Modelling of vertical and ferroelectric junctionless technology for efficient 3D neural network compute cube dedicated to embedded artificial intelligence (Invited)

Cristell Maneux, Mukherjee Chhandak, Marina Deng, Maeva Dubourg, Lucas Réveil, Georgeta Bordea, Aurélie Lecestre, Guilhem Larrieu, Jens Trommer, Evelyn T Breyer, et al.

Modelling of vertical and ferroelectric junctionless technology for efficient 3D neural network compute cube dedicated to embedded artificial intelligence

C. Maneux[1], C. Mukherjee[1], M. Deng[1], M. Dubourg[1], G. Bordea[1], A. Lecestre[2], G. Larrieu[2], J. Trommer[3], E.T. Breyer[3], S. Slesazeck[3], T. Mikolajick[3], O. Baumgartner[4], M. Karner[4], D. Pirker[4], Z. Stanojevic[4], David A. Atienza[5], A. Levisse[5], G. Ansaloni[5], A. Poittevin[6], A. Bosio[6], D. Deleruyelle[6], C. Marchand[6], I. O'Connor[6]

[1]IMS, University of Bordeaux, UMR CNRS 5218, Bordeaux INP, Talence, France, cristell.maneux@u-bordeaux.fr, [2]LAAS-CNRS, UPR 8001, CNRS, Université de Toulouse, Toulouse, France, [3]NaMLab gGmbH, Dresden, Germany, [4]Global TCAD Solutions GmbH, Vienna, Austria, [5]EPFL, Lausanne, Switzerland, [6]INL, University of Lyon, CNRS UMR 5270, Ecole Centrale de Lyon, Ecully, France.

Abstract— This paper presents the set of simulation means used to develop the concept of N^2C^2 (neural network compute cube) based on a vertical transistor technology platform. On the basis of state-of-the-art junctionless nanowire transistors (JLNT), TCAD simulation, compact modeling and EM simulation are leveraged through a Design-Technology Co-Optimization (DTCO) to achieve innovative 3D circuit architectures. Further, System-Technology Co-Optimization (STCO) implications on 3D NN system architecture are explored.ss

I. Introduction

In the context of the fourth industrial revolution along with unprecedented growing demand for neural networks (NNs), technology solutions still rely on transistors inherited from Moore's Law, continuously optimized to meet the requirements of von Neumann machines. The energy efficiency of von Neumann based processors is limited by the data transfer between the memory and the computing cores implemented in 2D integration schemes. This results in a crippling limitation for NNs efficiency. Since natural NNs feature a 3D framework, ideal hardware architectures should involve planes composed of k levels of (m*n) cubes. Each cube is a configurable 3D NN compute cell (N^2C^2) aimed at element-wise matrix multiplication including flexible means to call coefficients from non-volatile memory and for data routing to subsequent layers. To aim at this, several innovations are targeted:

- Leveraging 3D integration for gate length and contact area scaling, JLNT [1] can circumvent process challenges such as obtaining an abrupt shallow doping gradient at the source/drain junction, ensuring high and uniform body doping and minimal S/D contact resistance. In particular, vertical nanowire FETs (VNWFETs) represent a promising approach for monolithic 3D stacks to simultaneously address challenges of compactness, heat dissipation, and interconnect length.
- Embedded non-volatile memory (e-NVM) cells using 3D ferroelectric (FE) gated vertical transistors, utilizing an inherent remanent polarization, could also be used. These devices intimately incorporate multi-bit memory capability within the computing elements and path the way for new concepts computing-in-memory. Hence, they reduce the dynamic power consumption. Additionally, the 3D N^2C^2 will exhibit multiple means of configuration ensuring plasticity at the 3D NNs level, namely:
  - The number of inputs: possibility to configure the vertical routing of data between layers in both directions.
  - The synaptic coefficients: possibility to program them in NVM elements and connect them to the multipliers.
  - Activation function can be efficiently programmed in NVM elements in a coarse-grain Logic in Memory (LiM) approach.
- To enable a regular 3D matrix of configurable logic functions, a versatile and scalable inter-cube interconnect framework should be capable of housing multiple (in the order 10^6) non-volatile N^2C^2 structures in the x,y,z planes and routing all inter-cell data, control signals and power lines in an efficient, regular and organized way.

The paper presents the N^2C^2 modelling framework to ensure efficient DTCO as well as models for the higher 3D NN system architecture level enabling STCO.

II. Vertical GAA JLNTs

Vertical JLNTs with nanoscale metallic GAA (~14nm) and symmetrical silicided source and drain contacts (Fig. 1) have been demonstrated [1].

A. TCAD simulations

Advanced 3D technologies show an especially strong interaction between individual components, making accurate predictive simulations important for DTCO. The DTCO simulation flow applied here relies on an accurate extraction of the parasitic network (PEX) from a detailed 3D TCAD model of the logic cells built from layout and technology information [2]. Measurement data combined with insight from TCAD are key enablers of the compact model described in the next section. The automatically extracted PEX netlists are combined with the compact models of the vertical GAA JLNTs to run transient simulations of 3D logic cells enabling power, performance, and area analysis [3].

B. Compact models

The GAA JLNT SPICE compact model is based on a unified charge-based control model [4], accounting for the current calculation with short channel effects (for 14 nm NW channel length), velocity saturation, drain-induced barrier lowering (DIBL) as well as band-to-band tunneling (BTBT) and gate-induced drain leakage (GIDL). Schottky contact formation at the source and drain access regions is also included through thermionic and leakage current branches [5]. The compact model has been calibrated against both experimental and TCAD simulation data depicting a good model accuracy (Fig. 2). This VerilogA compact model has been coupled to the PEX netlist of the NAND (Fig.3) 3D logic cell, thus resulting in delay increase of 0.73 ns compared to the ideal scenario due to parasitic elements (Fig. 4). This confirms the need for accurate extraction of the PEX in order to capture the parasitic coupling effects in the designed 3D structures.

A Preisach-based [6] model (Fig. 5) of hysteresis is used for the non-volatile behaviour of the FE gate stack featuring a remanent polarization P_r and a coercive field E_c. The overall polarization P is described by integrating over the individual dipoles (superposition), approximating a hyperbolic tangent function for the saturated polarization curve: $P(E,k,P_{off}) = k \cdot P_{sat} \cdot tanh\left(\frac{E_{eff} \pm E_c}{2\delta}\right) + P_{off}$ where $\delta = E_c \left\{ln\left(\frac{1+P_r/P_{sat}}{1-P_r/P_{sat}}\right)\right\}^{-1}$. P_{sat} is the saturation polarization.~~,~~ The model fitted to experimental data of a capacitor (Fig. 6) ~~and~~ was area matched to the JLNT compact model to yield the predictive-NVM behavior (Fig. 7).

C. EM test structure simulations

To validate device and circuit parasitic element values (series resistive, inductive effects and capacitive couplings), test structure EM simulations are validated beforehand against on-wafer S-parameter measurements of dedicated Open and Short structure designs at different reference planes. The methodology was applied to an Open structure dedicated to JLNT de-embedding. The resulting equivalent electrical model of the interconnections (Fig. 8) has been included in the compact model. Agreement between measurements, EM simulation and electrical model simulation is observed in Fig. 9 up to 40 GHz.

III. JLNTs 3D Logic Cells and RFETs for Routing

In this section, we leverage the developed compact model to assess the performance metrics of a small library of conventional 1-bit Boolean logic cells implemented in the VNWFET technology. In Figs. 10-1~~3~~4 3D layout views of INV, NAND, NOR, and XOR gates are shown, respectively. Equivalent drive strength was adjusted to a single inverter via number of nanowires per VNWFET (1 n-type nanowire, 3 p-type nanowires) and assuming 45° metal routing. Fig. 14 illustrates transient simulation results for a NAND gate loaded with a capacitance varying from 10fF-80fF under a V_{dd}=1V supply voltage, completed in 120 seconds on an Intel® Xeon® Silver 4114 CPU running at 2.2GHz and demonstrating successful use of the compact model. We extracted typical performance characteristics of each gate necessary to enable logic synthesis – results are summarized in Fig. 15. Delay is measured as the time difference $t_r = t|_{V_{out}=0.95V_{dd}} - t|_{V_{in}=0.05V_{dd}}$ (resp. $t_f = t|_{V_{out}=0.05V_{dd}} - t|_{V_{in}=0.95V_{dd}}$). Transition energy E_{trans} is calculated as $E_t = V_{dd} \int_{t|_{V_{out}=0.05V_{dd}}}^{t|_{V_{out}=0.95V_{dd}}} I_{dd} dt$ for 0↗1 transitions at the output, and $E_t = V_{dd} \int_{t|_{V_{out}=0.95V_{dd}}}^{t|_{V_{out}=0.05V_{dd}}} I_{dd} dt$, for 1↘0 transitions at the output.

Adding ferroelectric material to the VNWFET gate stack as described in section II.B enables non-volatile logic as well as non-volatile reconfigurability. For example, a non-volatile full adder (Fig. 16, based on [7]) is able to store one of the summands in a non-volatile manner, which is of particular interest in multiplication operations used in digital filters or convolutional neural networks, and where one summand constantly varies (data), while the other one rarely varies (coefficients). Reconfigurable in-memory computing is enabled by the conjunction of ferroelectric VNWFETs with classical LookUp Table (LUT) circuit structures such as a LUT2 illustrated in Fig. 17 [~~ss~~8] and where the output depends both on the inputs S_0 and S_1 as well as on the stored states (here 4 states stored in a non-volatile manner to reflect a 2-input truth table), i.e. $Y = A.\overline{S_1}.\overline{S_0} + B.\overline{S_1}.S_0 + C.S_1.\overline{S_0} + D.S_1.S_0$.

Reconfigurable transistors (RFETs) leverage a multitude of gates on a single ambipolar channel, to dynamically select both conduction state, as well as carrier polarity (Fig. 18). RFETs connected to a common body [9] enable inherent X-to-1 multiplexing in a single transistor with X independent drain contacts. This way, a flexible reconfigurable input routing between N^2C^2 can be realized. Non-volatile RFETs in a 3D tile (Fig. 19) intimately incorporate multi-bit memory capability within computing elements, thus opening the way for a new concept for computing-in-memory.

IV. NNs Conceptual Perspectives

The N^2C^2 (Fig. 20) 3D building block concepts described above can be connected via a reconfigurable 3D interconnect network to construct powerful scalable and versatile 3D computing accelerator architectures (Fig. 21). Indeed, the computational patterns of NNs applications exhibit a high level of regularity, showcasing compute-intensive hotspots, for the most part embodied in Matrix-Matrix (MM) or Matrix-Vector multiplications. MM multiplications consume between 66% and 90% of the run-time for the BERT transformer NN (Fig. 22), thus requiring an efficient and dedicated support, especially in the light of the ever-increasing complexity of NN models.

Acknowledgment

This work is supported by ANR LEGO (Grant 18-CE24-0005-01), Horizon 2020 3εFerro (Grant N° 780302), and FVLLMONTI (Grant N°101016776).

References

[1] G. Larrieu, XL Han, (2013) Nanoscale 5 (6), 2437-2441.
[2] Z. Stanojević et al., in Proc. of ESSDERC 2018, pp. 202-205.
[3] G. Rzepa et al., in Proc. of IEEE IRPS, 2021, pp. 1-6.
[4] A. Hamzah et al., Physica Scripta, vol. 94, pp. 105813, 2019.
[5] C. Mukherjee et al., in Proc of IEEE VLSI SOC 2020, pp. 76-81.
[6] F. Preisach, in Zeitschrift für Physik, vol. 94, pp. 277-302, 1935.
[7] X. Yin et al., in Proc. of ICCAD, 2016, pp. 121:1 – 121:8.
[8] E.T. Breyer et al., in Proc. of ESSDERC 2019, pp. 118-121.
[9] T. Baldauf et al. IEEE EDL 39(8), 1242-1245 (2018).

[10] D. Dice et al., "Optimizing Inference Performance of Transformers on CPUs", 2021, arXiv:2102.06621.

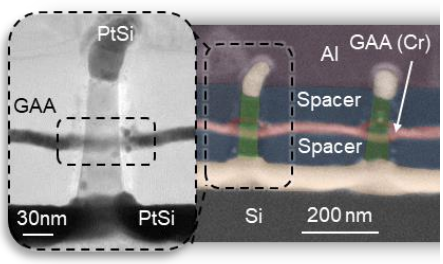

Fig. 1. Vertical JLNTs with 14nm GAA and S/D silicided contacts (adapted from [1]).

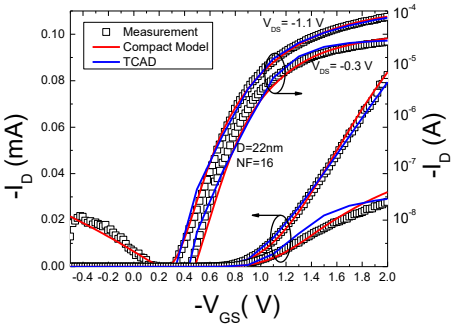

Fig. 2. Model calibration for JL VNWFETs.

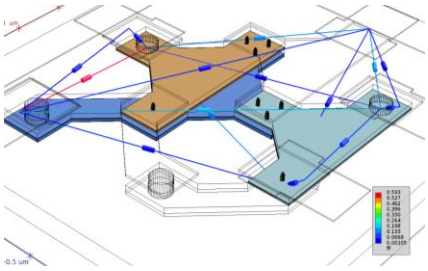

Fig. 3. NAND cell layout illustrating the main capacitive parasitic elements calculated by PEX.

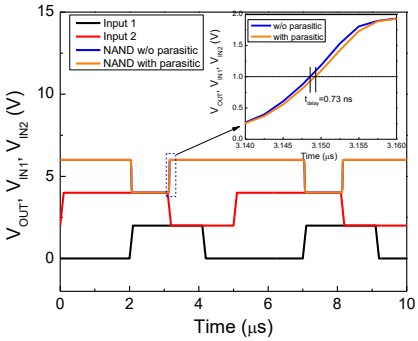

Fig. 4. Transient simulation of the NAND gate with and without parasitic elements.

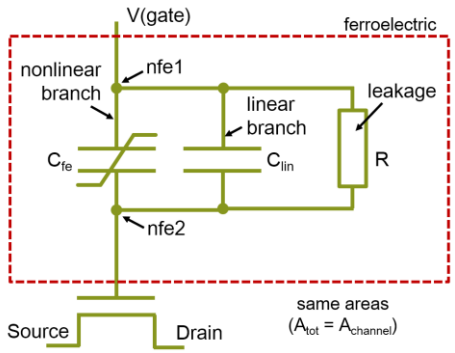

Fig. 5. Schematic of the Preisach-based Verilog-A FeCap model as gate extension for the VJLNW-FETs.

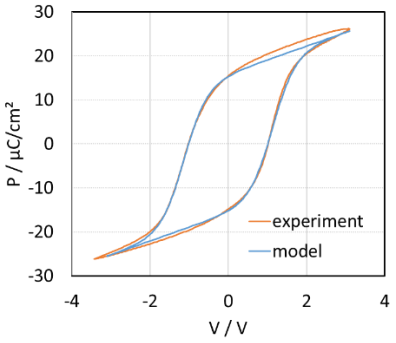

Fig. 6. E_c, P_r and P_{sat} have been calibrated to experimental data of a ferroelectric Hafnium-Zirconium-Oxide capacitor with 8.5 nm thickness.

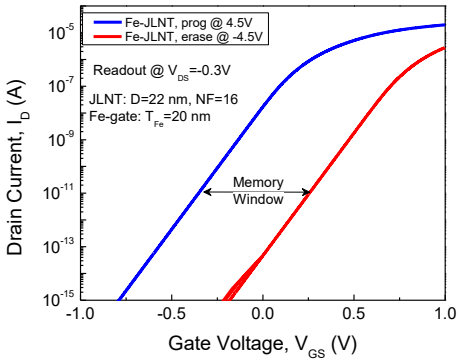

Fig. 7. Simulated predictive FE-JLNT program and erase characteristics with FE e-capacitor calibration data. The effect of GIDL is neglected in the compact model simulation.

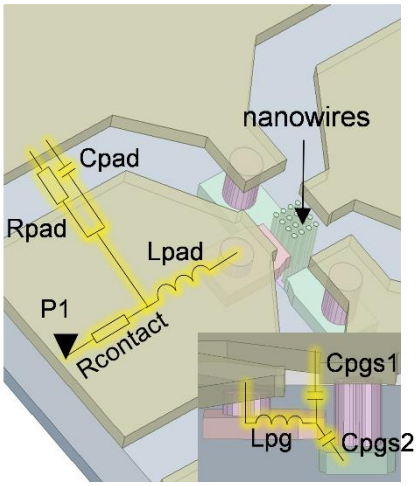

Fig. 8. EM simulation of JLNT test structure and interconnect modelling. The parameters of the distributed equivalent electrical circuit representing the parasitic effects were extracted using EM simulation of the structures' layout.

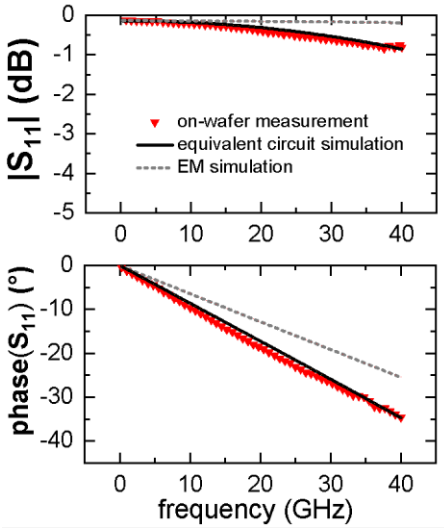

Fig. 9. Comparison between S-parameter measurement and simulation of Open test structure up to 40 GHz. A good agreement is obtained by adding additional capacitive coupling at close proximity to the nanowire location.

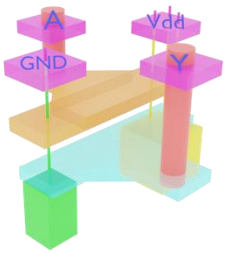

Fig. 10. 3D representation of inverter gate.

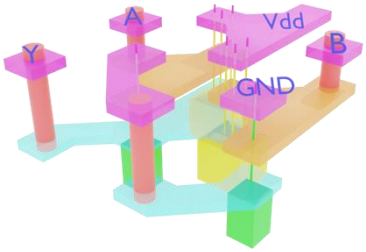

Fig. 11. 3D representation of NAND gate.

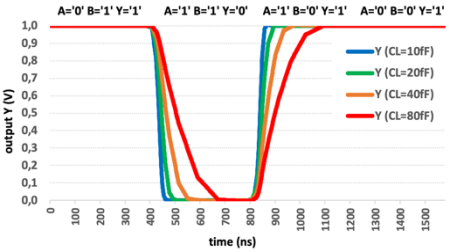

Fig. 14. Transient simulation of the NAND gate.

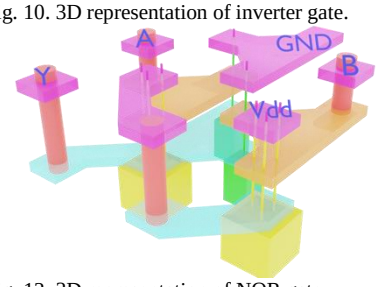

Fig. 12. 3D representation of NOR gate.

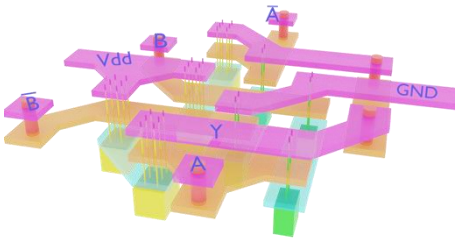

Fig. 13. 3D representation of XOR gate.

cell	transition	t_d (ns)	t_{trans} (ns)	E_{trans} (fJ)
INV	A↓ Y↑	9.25	t_r=11.47	1.010
	A↑ Y↓	4.75	t_f=5.64	0.022
NAND	A↓B Y↑	25	t_r=37.5	20.89
	A↑B Y↓	25	t_f=37.5	0.004
	A B↓Y↑	25	t_r=37.5	21.45
	A B↑Y↓	25	t_f=37.5	0.266
NOR	A↓/B Y↑	27.5	t_r=50	6.170
	A↑/B Y↓	20	t_f=33.3	333.3
	/A B↓Y↑	25	t_r=40	7.332
	/A B↑Y↓	13	t_f=16.6	0.008
XOR	A↑/B Y↑	45	t_r=66	3.518
	A↓/B Y↓	22.5	t_f=36	0.005
	/A B↑Y↑	40	t_r=60	4.267
	/A B↓Y↓	25	t_f=35	0.004

Fig. 15. Energy and delay characteristics of logic cells.

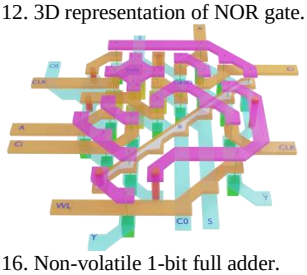

Fig. 16. Non-volatile 1-bit full adder.

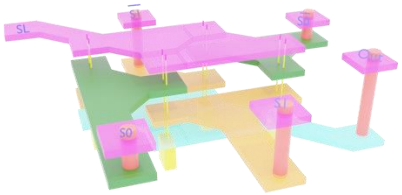

Fig. 17. 2-input non-volatile LUT.

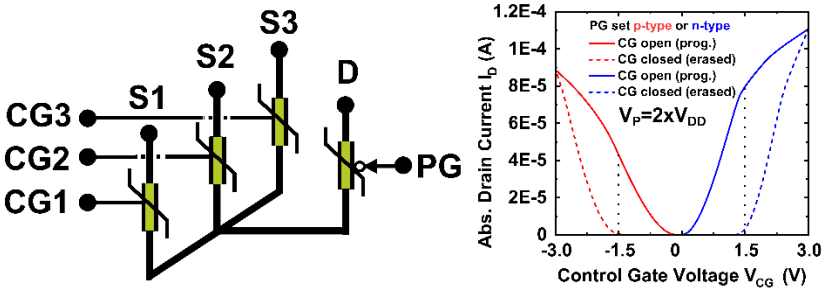

Fig. 18. Schematic of a vertical- reconfigurable transistor with FE control gate (CG) electrodes on a common body. The output current gives the sum of the number of source (S) inputs that are set ON by programming the FE control gates (CG). The polarity gate (PG) at drain (D) sets the carrier type. While volatile RFETs require a constant input to polarity gates, FE gate oxides are written by pulses, reducing the dynamic power consumption.

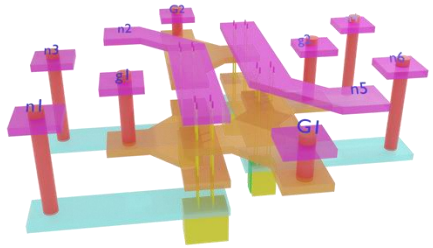

Fig. 19. Non-volatile circuit tile based on ambipolar reconfigurable FETs.

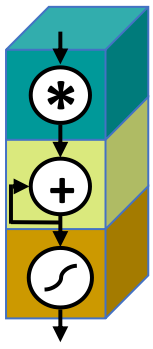

Fig. 20. Versatile VNWFET-logic based 3D neural network compute cube (N^2C^2).

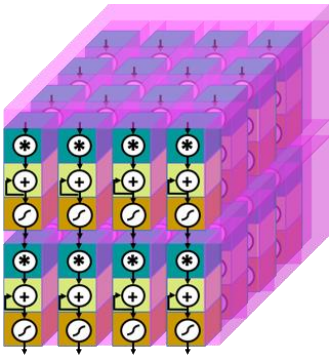

Fig. 21. NNs architecture based on the N^2C^2 concept illustrating the twin properties of physical regularity and functional versatility for in-memory processing.

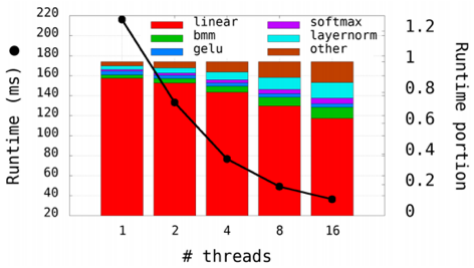

Fig. 22. Breakdown of BERT run-time, from [10]. Linear and bmm operations are realized as MM multiplications.