Analysis of Energy-Delay-Product of a 3D Vertical Nanowire FET Technology

Ian O'Connor, Arnaud Poittevin, Sébastien Le Beux, Alberto Bosio, Zlatan Stanojevic, Oskar Baumgartner, Jens Trommer, Thomas Mikolajick, Guilhem Larrieu, Mukherjee Chhandak, et al.

Analysis of Energy-Delay-Product of a 3D Vertical Nanowire FET Technology

I. O'Connor, A. Poittevin, S. Le Beux*, A. Bosio
Lyon Institute of Nanotechnology, CNRS UMR 5270, Ecole Centrale de Lyon Ecully, France
*Dept. of Electrical and Computer Engineering Concordia University, Canada
ian.oconnor@ec-lyon.fr

J. Trommer, T. Mikolajick**
NaMLab gGmbH, 01187 Dresden, Germany,
**Institute for Semiconductors and Microsystems and cfaed, TU Dresden, 01187 Dresden, Germany.
Jens.Trommer@namlab.com

Z. Stanojevic, O. Baumgartner
Global TCAD Solutions, Vienna, Austria
z.stanojevic@globaltcad.com

G. Larrieu
LAAS-CNRS, Université de Toulouse, Toulouse, France
guilhem.larrieu@laas.fr

C. Mukherjee, C. Maneux
IMS Laboratory, University of Bordeaux CNRS UMS 5218,
351, Cours de la Libération - 33405 Talence, France
chhandak.mukherjee@ims-bordeaux.fr

***Abstract*— To sustain transistor scaling beyond lateral 7nm devices, gate-all-around (GAA) junction-less vertical nanowire field effect transistors (VNWFET) are a promising alternative. This work analyses the energy-delay-product (EDP) for a junction-less 3D vertical gate-all-around nanowire FET technology, with a physical channel length of 14nm. Comparisons with the EDP of a baseline 7nm FinFET technology are carried out. The analysis motivates a new 3D neural network compute cube (N^2C^2) concept. Our results show that a 10x gain in EDP can be achieved for a physical VNWFET gate length of 14nm.**

Keywords—Vertical junctionless NWFET, logic circuit simulation.

I. Introduction

Emerging computing paradigms for the Internet of Things (IoT), in particular edge computing and edge artificial intelligence (AI), target real-time operations including data creation, decision, and action where milliseconds matter. Existing solutions are computation-intensive and energy-hungry requiring server-based implementations, which introduce latency and jitter, as well as raising data protection and privacy concerns. Deterministic, secure, and real-time operation is important for self-driving cars, robotics, industry 4.0, augmented reality, and many other areas. However, despite recent improvements in algorithm efficiency, energy-efficiency of the hardware has become a challenge. Embedded lightweight energy-efficient hardware remains elusive.

Today's 2D electronic architectures suffer from "unscalable" interconnects and are thus still far from being able to compete with biological neural systems in terms of real-time information-processing capabilities with comparable energy consumption. Recent advances in materials science, device technology, and synaptic architectures have the potential to fill this gap with novel disruptive technologies that go beyond conventional technology. A promising solution comes from vertical nanowire field-effect transistors (VNWFETs) [1-2] that unlock the full potential of truly unconventional 3D networks through a unique integration approach [1] termed Logic Element Gate Overstacking (LEGO). This technology has motivated the concept of a flexible 3D neural network compute cube (N^2C^2) with high integration density and performance. While significant gains in silicon area and energy-efficiency are expected through the combination of extremely small elementary device footprint and intrinsically 3D device and circuit functionality, this work is the first attempt to quantify the gains in terms of energy-efficiency of 3D logic blocks based on VNWFET devices.

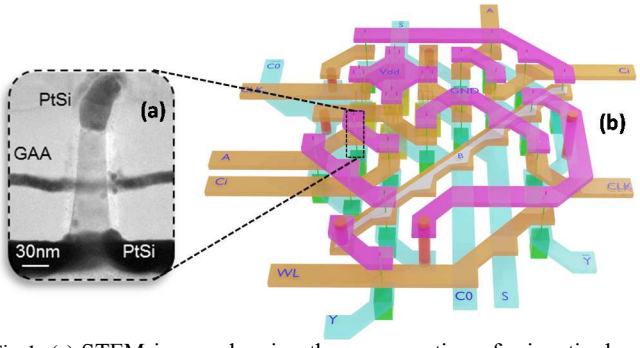


Fig 1: (a) STEM image showing the cross section of a junctionless vertical nanowire transistor, reproduced from [1], (b) 3D logic cell schematic of a 1-bit full adder using vertically stacked VNWFETs.

The rest of this paper is organized as follows: section II recalls the energy-delay product (EDP) metric and develops analytical equations for on-current and load capacitance based on geometrical and material parameters to compare EDP gains between VNWFETs and FinFETs. Section III applies this equation to explore EDP gains for both technologies and finally, Section IV concludes the work.

II. Analysis of EDP Improvement

The VNWFET technology under study [3, 4] is composed of a homogenous highly doped nanowire channel, patterned into boron doped ($2\times10^{19}cm^{-3}$) Si-substrate, working as a junction-less device [5]. The current between the silicided source/drain contacts is controlled by a gate-all-around structure having a physical channel length of 14nm (Fig. 1 (a)). Leveraging vertical integration of stacked VNWFETs, much higher compactness and flexibility of vertical device dimensions can be achieved (Fig. 1 (b)) and a significant area gain compared to a 7nm FinFET technology can be expected.

Energy-Delay Product (EDP) is a useful metric to compare the speed of energy-efficient circuits. While the Power-Delay Product (PDP), or the switching energy, measures the energy per function, it does not satisfactorily capture the speed. In a classical CMOS circuit, considering the PDP for a 0-to-1-to-0 computation cycle, the EDP can be written as

$$EDP = PDP * t_P = C_L V_{DD}^2 * t_P = \frac{C_L^2 V_{DD}^3}{I_{sat}} \quad (1)$$

where t_p represents the propagation time, C_L represents the load capacitance on the gate output, V_{DD} represents the supply voltage and I_{sat} represents the saturation current of the transistor through which the current is drawn from the voltage supply or sunk to the ground to change the output voltage state (we assume that the transistor through which the current is flowing is primarily in the saturation region).

The ratio between the EDP of the FinFET technology and the VNWFET technology can therefore be expressed as:

$$G_{Evf} = \frac{EDP_f}{EDP_v} = \frac{C_{Lf}^2 V_{DDf}^3}{I_{satf}} \frac{I_{satv}}{C_{Lv}^2 V_{DDv}^3} \approx \frac{C_{Lf}^2}{C_{Lv}^2} \frac{I_{satv}}{I_{satf}} \quad (2)$$

where the subscripts f and v stand for the FinFET and VNWFET, respectively. We assume that supply voltage values for both technologies are identical (i.e. $V_{DDf} = V_{DDv}$). Hence, for a given value n of G_Evf, the EDP of the VNWFET technology is n times lower than that of the FinFET technology.

A. *On-current (I_{sat}) considerations*

We also assume that material current density and carrier mobility are close enough for both technologies such that their influence can be ignored in this comparison. This then implies that the transistor saturation current can be approximated as

$$I_{sat} = \kappa \frac{W_g}{L_g} \quad (3)$$

where κ is a constant incorporating technology parameters and operating conditions, W_g represents the effective width of the transistor channel under the gate orthogonal to current flow, and L_g represents the length of the channel under the gate, i.e. the distance between source and drain. This also assumes that the transistor channel occupies all the space available in the given geometry – this implies that the channel is fully depleted in the FinFET and the radius of the nanowire is sufficiently small in the VNWFET to allow junctionless transport.

We can calculate W_g according to the geometry of both FinFET and VNWFET devices:

$$\left.\begin{aligned} W_{gf} &= 2(h_f + w_f) \\ W_{gv} &= 2\pi r_v \end{aligned}\right\} \quad (4)$$

where W_{gf} and W_{gv} represent the effective widths of the FinFET and VNWFET channels, respectively; h_f and w_f represent the height and width of a fin in the FinFET device; and r_v represents the radius of the VNWFET nanowire channel. The geometries and parameters of both devices are shown in Figs. 2 (FinFET) and 3 (VNWFET). Thus, the ratio between the saturation currents of both devices can be written as:

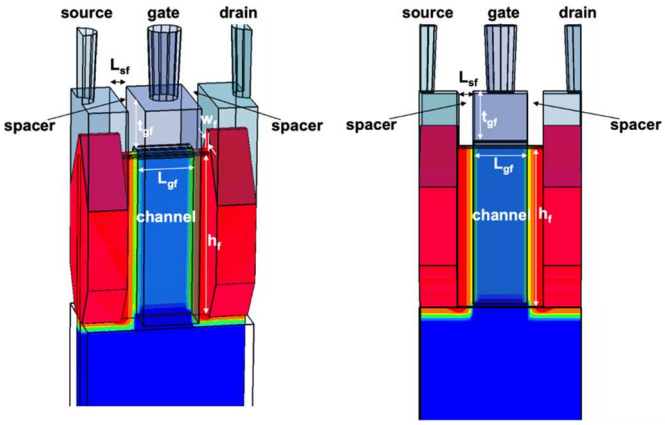
$$\frac{I_{satv}}{I_{satf}} = \frac{W_{gv}}{L_{gv}} \frac{L_{gf}}{W_{gf}} = \frac{\pi r_v}{(h_f + w_f)} \frac{L_{gf}}{L_{gv}} \quad (5)$$

In nanowire-based devices, the ratio of channel length L_{gv} to channel diameter $2r_v$ should be kept constant at 2:1 to preserve desirable behavior in the off state [6]. It is also important to avoid degradation of both ballistic and dissipative currents, which occurs with decreasing device size. In fact, ballisticity (i.e. the ratio of dissipative to ballistic current) also degrades for very small devices, with channel lengths below around 10nm. Both constraints can be combined resulting in:

$$L_{gv} \geq 10nm, r_v = \frac{L_{gv}}{4} \text{ subject to } r_v \geq 2.5nm \quad (6)$$

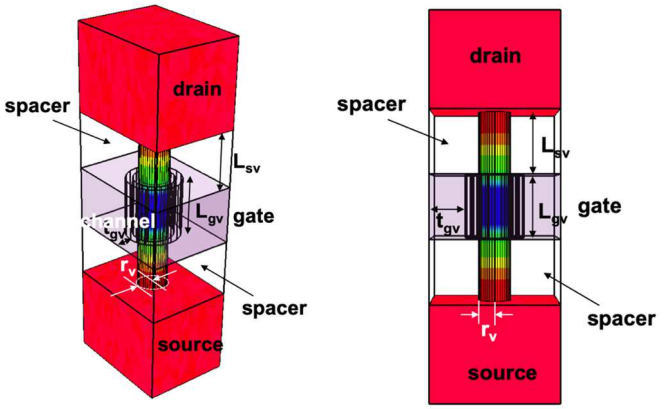
B. *Load capacitance (C_L) considerations*

In terms of capacitance, we will, in a first approach, consider that load capacitance is composed of the input gate capacitance C_{in} of F (fanout) logic gates on the output node, as well as interconnect capacitance C_w. Hence,



(a) (b)

Fig. 2: FinFET geometry (a) 3D view (b) lateral view perpendicular to the gate electrode; color scheme: red (resp. blue) indicates a high (resp. low) electron density for an n-channel device (vice versa for p-channel devices).



(a) (b)

Fig 3: VNWFET geometry (a) 3D view (b) lateral view; color scheme: red (resp. blue) indicates a high (resp. low) electron density for an n-FET (vice versa for p-FETs).

$$C_L = F(C_{in} + C_w) = F(C_{gc} + C_{gs} + C_w) \quad (7)$$

where the input gate capacitance is composed of the direct gate-channel capacitance C_{gc} linked to the gate-channel area, and the gate-source capacitance C_{gs} linked to the spacer geometry.

1) *Gate-channel capacitance*

The gate-channel capacitance can be expressed in terms of channel geometry for both devices as:

$$\left.\begin{aligned} C_{gcf} &= \frac{\varepsilon_{ox} L_{gf} 2(h_f + w_f)}{EOT} \\ C_{gcv} &= \frac{\varepsilon_{ox} L_{gv} 2\pi r_v}{EOT} \end{aligned}\right\} \quad (8)$$

Here, C_{gcf} and C_{gcv} are the gate-channel capacitance for FinFET and VNWFET devices, respectively; ε_{ox} and EOT are the dielectric permittivity for standard SiO_2 and the equivalent oxide thickness of the gate dielectric material (i.e. converted to the thickness it would have using SiO_2 as the gate dielectric), respectively. Assuming that the dielectric material is identical for both technologies, EOT writes:

$$EOT = t_{high-K}\frac{\varepsilon_{ox}}{\varepsilon_{high-k}} \quad (9)$$

where $\varepsilon_{high\text{-}k}$ is the permittivity of the hi-k gate dielectric. A typical value for EOT is around 0.89nm in current technologies where HfO_2 gate dielectrics are used.

2) Gate-source capacitance

The gate-source capacitance can be expressed in terms of spacer geometry for both devices as:

$$\left.\begin{aligned} C_{gsf} &= \frac{\varepsilon_s\, 2t_{gf}\left(2t_{gf}+h_f+w_f\right)}{L_{sf}} \\ C_{gsv} &= \frac{\varepsilon_s\left(4\left(t_{gv}+r_v\right)^2-\pi r_V^2\right)}{L_{sv}} \end{aligned}\right\} \quad (10)$$

where C_{gsf} and C_{gsv} represent the gate-source capacitance for FinFET and VNWFET devices, respectively; t_{gf} and t_{gv} represent the gate material thickness for FinFET and VNWFET devices, respectively and can be typically assumed to be between 10nm-20nm; L_{sf} and L_{sv} represent the spacer length (between gate and source) for FinFET and VNWFET devices, respectively, and can be typically assumed to be around 8nm; ε_s represents the spacer material dielectric permittivity. We assume that the spacer material (typically Si_3N_4) is identical for both technologies and that there are no fabrication issues for different dielectric materials for the gate (high-k) and for the spacer. We also assume that the gate material surrounds the channel with uniform thickness (overlap) equal to t_{gf} for the FinFET (although the lateral thickness is defined by the FinFET pitch), and is a square centered around the nanowire with minimum overlap equal to t_{gv} for the VNWFET. Note that this expression does not consider fringing capacitances, which is a reasonable assumption since this is not the dominant component. Hence one can write:

$$\left.\begin{aligned} C_{inf} &= \frac{\varepsilon_{ox}L_{gf}2\left(h_f+w_f\right)}{EOT} + \frac{\varepsilon_s\, 2t_{gf}\left(2t_{gf}+h_f+w_f\right)}{L_{sf}} \\ C_{inv} &= \frac{\varepsilon_{ox}L_{gv}2\pi r_v}{EOT} + \frac{\varepsilon_s\left(4\left(t_{gv}+r_v\right)^2-\pi r_V^2\right)}{L_{sv}} \end{aligned}\right\} \quad (11)$$

3) Wire capacitance

Wire capacitance is considered for local (gate-to-gate) interconnect, and is expressed as:

$$C_w = \frac{\varepsilon_{ox}w_m L_{gg}}{t_{ox}} \quad (12)$$

where w_m and L_{gg} represent the width and length of local (gate-to-gate) interconnect respectively; and t_{ox} represents the metal-substrate oxide thickness for interlayer dielectric SiO_2. L_{gg} can be directly linked to circuit compactness since it represents the lateral distance (pitch) between two gates. For a given improvement in compactness A_c between VNWFET and FinFET (i.e. $A_c = A_f / A_v$),

$$\frac{L_{ggf}}{L_{ggv}} = \sqrt{A_c} \quad (13)$$

Assuming identical values for ε_{ox}, t_{ox} and w_m between the FinFET and VNWFET technologies, we can also write:

$$\frac{C_{wf}}{C_{wv}} = \sqrt{A_c} \quad (14)$$

4) Overall expressions for the load capacitances

Leveraging (14) one can write the expression for the load capacitances in both cases as

$$\left.\begin{aligned} C_{Lf} &= F\left[C_{gcf} + C_{gsf} + C_{wf}\right] \\ C_{Lv} &= F\left[C_{gcv} + C_{gsv} + \frac{C_{wf}}{\sqrt{A_c}}\right] \end{aligned}\right\} \quad (15)$$

where C_{Lf} and C_{Lv} represent the load capacitances on FinFET and VNWFET logic gate outputs, respectively.

Thus, the ratio between the load capacitances of both devices can be written as:

$$\frac{C_{Lf}^2}{C_{Lv}^2} = \frac{\left[C_{gcf}+C_{gsf}+C_{wf}\right]^2}{\left[C_{gcv}+C_{gsv}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2} \quad (16)$$

And finally, the ratio between the EDPs can be expressed as,

$$G_{Evf} = \frac{C_{Lf}^2}{C_{Lv}^2}\frac{I_{Satv}}{I_{satf}} = \frac{\left[C_{gcf}+C_{gsf}+C_{wf}\right]^2}{\left[C_{gcv}+C_{gsv}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2}\frac{I_{Satv}}{I_{satf}} \quad (17)$$

Which can be further re-written in terms of geometric parameters as,

$$G_{Evf} = \frac{\left[\frac{\varepsilon_{ox}L_{gf}2\left(h_f+w_f\right)}{EOT}+\frac{\varepsilon_s\, 2t_{gf}\left(2t_{gf}+h_f+w_f\right)}{L_{sf}}+C_{wf}\right]^2}{\left[\frac{\varepsilon_{ox}L_{gv}2\pi r_v}{EOT}+\frac{\varepsilon_s\left(4\left(t_{gv}+r_v\right)^2-\pi r_V^2\right)}{L_{sv}}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2}\frac{\pi r_v}{\left(h_f+w_f\right)}\frac{L_{gf}}{L_{gv}} \quad (18)$$

5) Impact on 3D design

For this analysis, we assume that A_c is a constant and of the order of 5 according to [7], although it is anticipated that it will vary (negatively) with an increasing number of fins per FinFET / nanowires per VNWFET. This hypothesis is explored quantitatively in the next section and will require further analysis once an automated 3D place and route tool is available. This tends also to support the view that

a) VNWFET-based design performance improves for lower numbers of nanowires per VNWFET – not only for the pitch overhead but also because the on-current varies sublinearly with the nanowire number per VNWFET.

b) Computing should be kept local to enable short interconnects and limit energy consumption in parasitic elements (particularly relevant for mobile low power applications).

The proposed N^2C^2 concept (fig. 4) enabled by the VNWFET technology is exactly this – a regular 3D matrix of individual computing functions where intra-cube interconnect is short due to both the limited complexity of the N^2C^2 circuit and the limited number of targeted nanowires per VNWFET. While the benefit of local computing is known to limit the impact of interconnect delays, N^2C^2 goes beyond the current planar state-of-the-art by extending this principle to 3 dimensions. Fig. 4 illustrates the N^2C^2 concept as a scalable and flexible assembly of high-expressivity logic cells based on VNWFETs. By focusing on commonly used functions in deep learning and convolutional neural network architectures such as Multiply-Accumulate, combined with logic design styles suited to multiple transistors in series such as Pass-Transistor logic, we target compact and low-EDP neural network compute cube hardware capable of interfacing seamlessly with other, identical cubes through a versatile 3D interconnect framework. This will facilitate the development of physically regular and logically flexible 3D matrices for AI accelerators, including levers to explore hardware/software

co-design and approximation techniques to further improve energy-efficiency both at design time and at run-time.

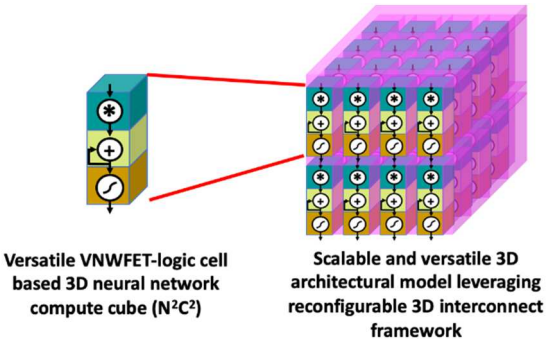


Fig 4: N^2C^2 concept

III. EDP Gain Analysis

For a FinFET aspect ratio (h_f/w_f) value of 60nm/7nm with 20nm physical FinFET gate length L_{gf}, and by varying the value of physical VNWFET gate length L_{gv} between 10nm-20nm for varying nanowire radius r_v, gate material thickness t_{gv}, spacer length L_{sv} and compactness A_c, EDP- and PDP-gain values have been calculated, as shown in Fig. 5.

Among the chosen radii, the lowest permissible value is 2.5nm (according to (6)) and the highest is 25nm which corresponds to the largest diameter available for the VNWFET technology. For low r_v, the on-current ratio I_{satv}/I_{satf} is lower than 1, but is offset significantly by the $(C_{lf}/C_{lv})^2$ ratio; while for high r_v, the on-current ratio approaches 2,5 while the $(C_{lf}/C_{lv})^2$ ratio tends towards unity. The results show that 10x gain in EDP between VNWFET and 7nm FinFET technology can be achieved for L_{gv}=14nm and a nanowire radius of 3.5nm. If the smallest fabricated device dimensions are considered, i.e. L_{gv}=14nm and r_v=11nm, an EDP gain of 4.3 has been predicted. Considering the worst-case scenario, for the largest already-fabricated devices (r_v=25nm and L_{gv}=14 nm [4]), a 2x EDP gain can still be observed. This analysis demonstrates that the VNWFET can be designed to be more energy-efficient compared to a 7nm FinFET.

IV. Conclusions

It is well known that transistor energy efficiency improves as gate length decreases. This is especially true for GAA NWFETs, among which the disruptive vertical device implementations allow the transition from 2D to truly 3D architectures. In this work, we developed a comparative analysis of logic energy-efficiency through first-order equations for EDP of VNWFET and FinFET technologies. Leveraging a set of realistic geometric and material parameter values, our work shows that a 10x gain in EDP over a baseline 7nm FinFET technology can be achieved for a physical VNWFET gate length of 14nm and even for actual fabricated devices, 4.3x gain in EDP has been predicted. These results pave the way towards 3D neural network compute cube required for dense and energy efficient non Von Neumann computing.

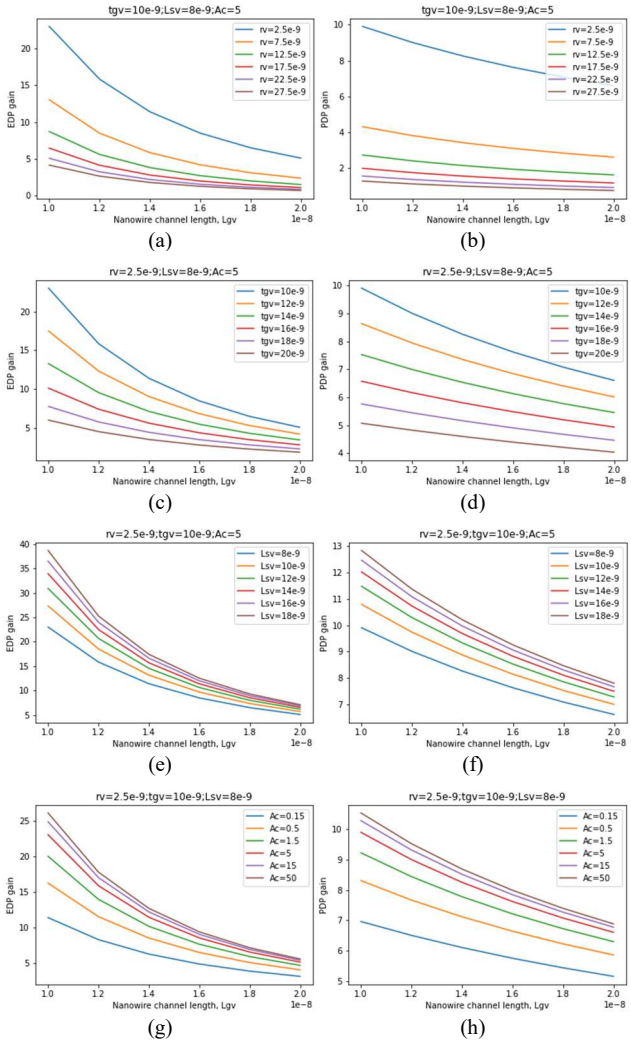


Fig. 5: EDP- and PDP-gain versus VNWFET gate length: a),b) for varying nanowire radius r_v; c),d) for varying gate material thickness t_{gv}; e),f) for varying spacer length L_{sv}; g),h) for varying compactness A_c.

Acknowledgment

This work was supported by the LEGO project (Grant ANR-18-CE24-0005-01) and by the project FVLLMONTI funded by European Union's Horizon 2020 research and innovation program under grant agreement N°101016776.

References

[1] G. Larrieu and X. L. Han " Vertical nanowire array-based field effect transistors for ultimate scaling", *Nanoscale*, vol. 5, pp. 2437-2441, 2013. DOI: 10.1039/C3NR33738C

[2] A. Veloso et al., "Vertical nanowire FET integration and device aspects", ECS Transactions, vol. 72 (4), pp. 31-42, 2016. DOI: 10.1149/07204.0031ecst

[3] Y. Guerfi and G. Larrieu "Vertical Silicon Nanowire Field Effect Transistors with Nanoscale Gate-All-Around", *Nanoscale Research Letters*, vol. 11, pp. 210, 2016. DOI: 10.1186/s11671-016-1396-7

[4] G. Larrieu, Y. Guerfi, X. L. Han and N. Clément "Sub-15 nm gate-all-around field effect transistors on vertical silicon nanowires", *Solid-State Electronics*, vol. 130, pp. 9-14, 2017. DOI: 10.1016/j.sse.2016.12.008

[5] J. P. Colinge *et al.* "Nanowire transistors without junctions", *Nat. Nanotechnol.* vol. 5, pp. 225-229, 2010. DOI: 10.1038/nnano.2010.15

[6] Z. Stanojevic, O. Baumgartner, M. Karner, F. Mitterbauer, H. Demel, and C. Kernstock, "Simulation Study on the Feasibility of Si as Material for Ultra-Scaled Nanowire Field-Effect Transistors," *Proc. EUROSOI-ULIS,* Vienna, Austria, pp. 25-27, 2016, DOI: 10.1109/ULIS.2016.7440074.

[7] C. Mukherjee, M. Deng, F. Marc, C. Maneux, A. Poittevin, I. O'Connor, S. Le Beux, A. Kumar, A. Lecestre, G. Larrieu, "3D logic cells design and results based on Vertical NWFET technology including tied compact model", *IFIP/IEEE International Conference on Very Large Scale Integration* (*VLSI-SoC*), 5-7 October 2020, Salt Lake City (UT), USA. DOI: arXiv:2005.14039v1.