

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

Ian O'Connor, Zlatan Stanojevic, Oskar Baumgartner, Chhandak Mukherjee, Guilhem Larrieu, Jens Trommer, Cristell Maneux

▶ **To cite this version:**

Ian O'Connor, Zlatan Stanojevic, Oskar Baumgartner, Chhandak Mukherjee, Guilhem Larrieu, et al.. Analysis of Energy-Delay-Product for a 3D Vertical Nanowire FET Technology for Logic Applications. 2020. hal-02732902

HAL Id: hal-02732902
https://hal.science/hal-02732902

Preprint submitted on 2 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Ian O'Connor[1], Zlatan Stanojevic[2], Oskar Baumgartner[2], Chhandak Mukherjee[3], Guilhem Larrieu[4], Jens Trommer[5] and Cristell Maneux[3]

[1] Lyon Institute of Nanotechnology, University of Lyon, CNRS UMR 5270, Ecole Centrale de Lyon Ecully, France
[2]Global TCAD Solutions GmbH, Vienna, Austria
[3]IMS Laboratory, CNRS UMR 5218, University of Bordeaux, France
[4]LAAS-CNRS, Université de Toulouse, France
[5]NaMLab GmbH, Dresden, Germany

***Abstract*— This document proposes a simplified analysis of the energy-delay-product (EDP) for a junction-less 3D vertical gate-all-around nanowire FET technology, with a physical channel length of 14nm, in comparison with the EDP of a baseline 7nm FinFET technology.**

I. Energy Delay Product

Energy-delay product (EDP) is a useful metric to compare the speed of energy-efficient circuits. While the power Delay Product (PDP) only measures the energy per function, it does not satisfactorily capture the speed. The EDP is written as,

$$EDP = PDP * t_P \quad (1),$$

where t_p represents the propagation time.

In a classical CMOS circuit, the PDP (or the switching energy) for a 0-to-1-to-0 computation cycle is defined by:

$$PDP = C_L V_{DD}^2 \quad (2),$$

where C_L represents the load capacitance on the gate output, and V_{DD} represents the supply voltage.

The switching or propagation time can also be computed by

$$t_p = \frac{C_L V_{DD}}{I} \quad (3),$$

Here, I represents the current drawn from the voltage supply or sunk to ground to change the output voltage state (on-current). For the purposes of this comparison, we will assume that the transistor through which the current is flowing is primarily in the saturation region and that this on-current equates to the saturation current I_{sat} of the transistor.

One can now write an expression for the EDP using (1)-(3):

$$EDP = PDP * t_p = \frac{C_L^2 V_{DD}^3}{I_{sat}} \quad (4).$$

II. EDP Improvement Equations

The EDP improvement of the VNWFET technology over the FinFET technology can therefore be expressed as:

$$G_{Evf} = \frac{EDP_f}{EDP_v} = \frac{C_{Lf}^2 V_{DDf}^3}{I_{satf}} \frac{I_{satv}}{C_{Lv}^2 V_{DDv}^3} \quad (5)$$

where the subscripts f and v stand for the FinFET and VNWFET, respectively.

We will assume that supply voltage values for both technologies are identical (i.e. $V_{DDf} = V_{DDv}$), which simplifies (5) further:

$$G_{Evf} = \frac{C_{Lf}^2}{C_{Lv}^2} \frac{I_{satv}}{I_{satf}} \quad (6)$$

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

We also assume that material current density and doping levels are identical for both technologies. This then implies that the transistor saturation current can be approximated as

$$I_{sat} = \kappa \frac{W_g}{L_g} \quad (7)$$

where κ is a constant that takes into account 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 material is fully depleted (i.e. the channel occupies all the space available in the given geometry – this implies that the radius of the nanowire is sufficiently small, which may be a limiting technological factor).

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

$$W_{gf} = 2(h_f = w_f), W_{gv} = 2\pi r_v \quad (8)$$

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. 1 (FinFET) and 2 (VNWFET).

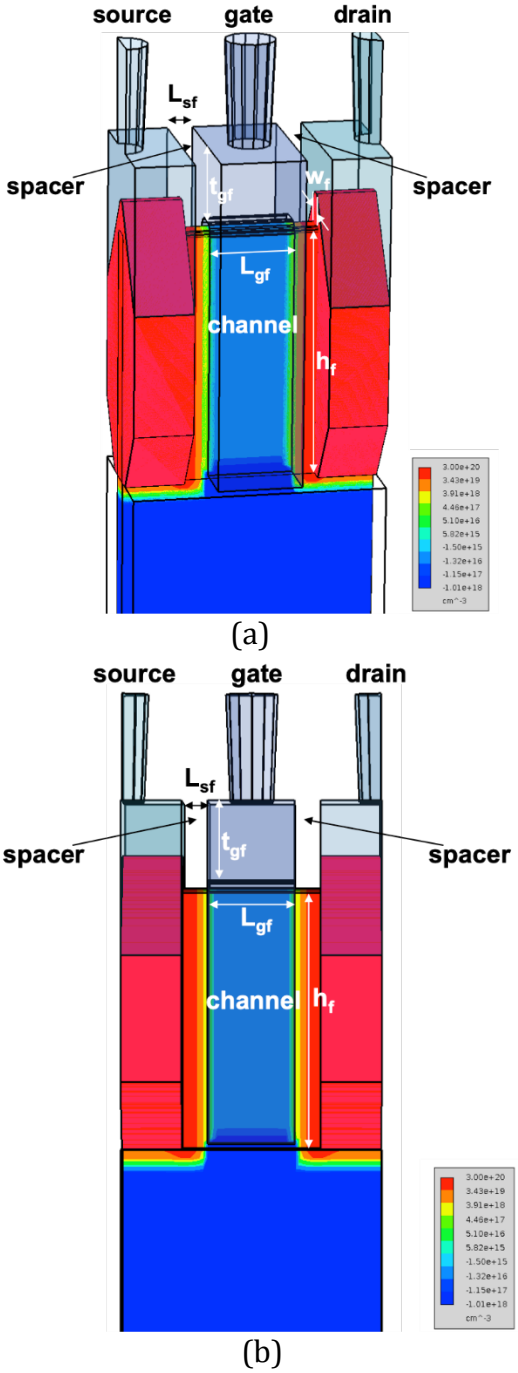


Fig. 1: FinFET geometry (a) 3D view (b) lateral view.

Thus, the ratio between the saturation currents of both devices can be written as:

$$\frac{I_{satv}}{I_{satf}} = \frac{W_{gv}}{L_{gv}}\frac{L_{gf}}{W_{gf}} = \frac{\pi r_v}{(h_f+w_f)}\frac{L_{gf}}{L_{gv}} \qquad (9)$$

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 [STA16]. 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 combine to give:

$$L_{gv} \geq 10nm, \quad r_v = L_{gv}/4 \geq 2.5nm \qquad (10)$$

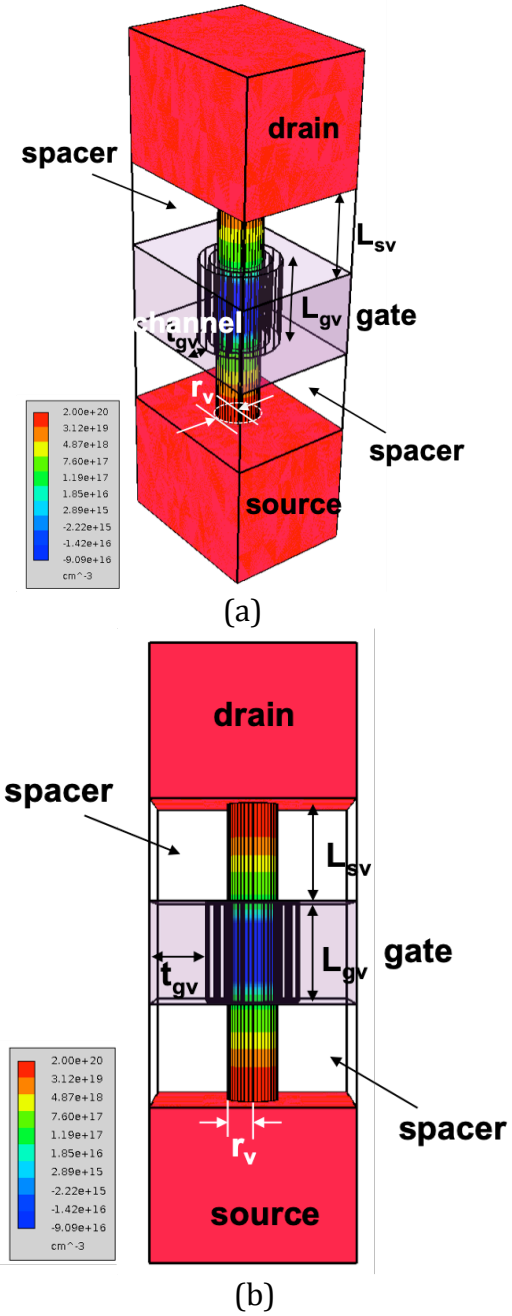


Fig. 2: VNWFET geometry (a) 3D view (b) lateral view

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,

$$C_L = F(C_{in} + C_w) \quad (11)$$

Input gate capacitance is in turn composed of direct gate-channel capacitance C_{gc} linked to the gate-channel area, and gate-source capacitance C_{gs} linked to the spacer geometry, as written by,

$$C_{in} = C_{gc} + C_{gs} \quad (12)$$

1. Gate-channel capacitance

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

$$\left.\begin{aligned} C_{gcf} &= \frac{\varepsilon_{ox} L_{gf} 2(h_f + w_f)}{EOT} \\ C_{gcv} &= \frac{\varepsilon_{ox} L_{gv} 2\pi r_v}{EOT} \end{aligned}\right\} \quad (13)$$

Here, C_{gcf} and C_{gcv} represent the gate-channel capacitance for FinFET and VNWFET devices, respectively; ε_{ox} and EOT represent the dielectric permittivity for standard SiO_2 and the equivalent oxide thickness for the gate dielectric material, respectively. We assume that the gate dielectric material is identical for both technologies. EOT, defined as:

$$EOT = t_{high-K} \frac{\varepsilon_{ox}}{\varepsilon_{high-k}} \quad (14),$$

is around 0.89nm in current technologies.

2. Gate-source capacitance

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

$$\left.\begin{aligned} C_{gsf} &= \frac{\varepsilon_s \, 2t_{gf}(2t_{gf} + h_f + w_f)}{L_{sf}} \\ C_{gsv} &= \frac{\varepsilon_s \left(4(t_{gv} + r_v)^2 - \pi r_V^2\right)}{L_{sv}} \end{aligned}\right\} \quad (15)$$

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; $\varepsilon_s$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 actually defined by 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 take into account fringing capacitances (a reasonable assumption since this is not the dominant component). Hence one can write,

$$\left.\begin{aligned} C_{inf} &= \frac{\varepsilon_{ox} L_{gf} 2(h_f + w_f)}{EOT} + \frac{\varepsilon_s \, 2t_{gf}(2t_{gf} + h_f + w_f)}{L_{sf}} \\ C_{inv} &= \frac{\varepsilon_{ox} L_{gv} 2\pi r_v}{EOT} + \frac{\varepsilon_s \left(4(t_{gv} + r_v)^2 - \pi r_V^2\right)}{L_{sv}} \end{aligned}\right\} \quad (16)$$

3. Wire capacitance

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

$$C_w = \frac{\varepsilon_{ox} w_m L_{gg}}{t_{ox}} \quad (17)$$

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. $A_c = A_f/A_v$),

$$\frac{L_{ggf}}{L_{ggv}} = \sqrt{A_c} \quad (18)$$

For this analysis, we assume that A_c is a constant and of the order of 5 according to [MUK20], although it is anticipated that it will vary (negatively) with increasing number of fins per FinFET / nanowires per VNWFET. This 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 number of nanowires per VNWFET.
b) computing should be kept local to enable short interconnect and limit energy consumption in parasitic elements (particularly important for mobile low power applications). The N^2C^2 (neural network compute cube) concept is exactly this – a regular 3D matrix of individual compute functions where intra-cube interconnect is short due to both the limited complexity of the N^2C^2 circuit and the targeted limited number of nanowires per VNWFET. While the locality of computing is a known technique to limit the impact of interconnect delay on

computing, the N^2C^2 concept goes beyond current planar state of the art by extending this principle to 3 dimensions.

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

$$\frac{C_{wf}}{C_{wv}} = \sqrt{A_c} \tag{19}$$

4. Overall expressions for the load capacitances

Leveraging (19) one can write the expression for the load capacitances in the two cases as,

$$\left.\begin{aligned} C_{Lf} &= F\left[C_{gcf} + C_{gsf} + C_{wf}\right] \\ C_{Lv} &= F\left[C_{gcv} + C_{gsv} + \frac{C_{wf}}{\sqrt{A_c}}\right] \end{aligned}\right\} \tag{20}$$

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:

$$\frac{C_{Lf}^2}{C_{Lv}^2} = \frac{\left[C_{gcf}+C_{gsf}+C_{wf}\right]^2}{\left[C_{gcv}+C_{gsv}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2} \tag{21}$$

And finally, the EDP gain can be expressed as,

$$G_{Evf} = \frac{C_{Lf}^2}{C_{Lv}^2}\frac{I_{Satv}}{I_{satf}} = \frac{\left[C_{gcf}+C_{gsf}+C_{wf}\right]^2}{\left[C_{gcv}+C_{gsv}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2}\frac{I_{Satv}}{I_{satf}} \tag{22}$$

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

$$G_{Evf} = \frac{\left[\frac{\varepsilon_{ox}L_{gf}2\left(h_f+w_f\right)}{EOT}+\frac{\varepsilon_S\, 2t_{gf}\left(2t_{gf}+h_f+w_f\right)}{L_{sf}}+C_{wf}\right]^2}{\left[\frac{\varepsilon_{ox}L_{gv}2\pi r_v}{EOT}+\frac{\varepsilon_s\left(4(t_{gv}+r_v)^2-\pi r_V^2\right)}{L_{sv}}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2}\frac{\pi r_v}{\left(h_f+w_f\right)}\frac{L_{gf}}{L_{gv}} \tag{23}$$

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 a nanowire radius r_v of 4nm, the EDP gain value varies as shown in Fig. 3. The results show that (for example) 10x gain in EDP between VNWFET and 7nm FinFET technology can be achieved for L_{gv}=14nm.

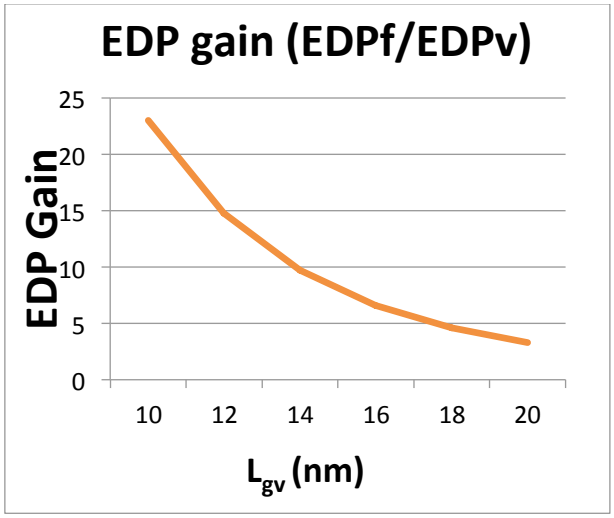


Fig. 3: EDP-gain versus VNWFET gate length

References

[1] Z. Stanojevic, O. Baumgartner, M. Karner, F. Mitterbauer, H. Demel, C. Kernstock, "Simulation Study on the Feasibility of Si as Material for Ultra-Scaled Nanowire Field-Effect Transistors," Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), Vienna (AT), 25-27 Jan. 2016, DOI: 10.1109/ULIS.2016.7440074

[2] 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 Int. Conf. Very Large Scale Integration (VLSI-SoC), 5-7 October 2020, Salt Lake City (UT), USA (submitted)