Performance and Design Considerations for Gate-All-Around Stacked-NanoWires FETs

S. Barraud, V. Lapras, B. Previtali, M. Samson, J. Lacord, S. Martinie, M.-A Jaud, S Athanasiou, F. Triozon, O. Rozeau, et al.

Performance and Design Considerations for Gate-All-Around Stacked-NanoWires FETs

S. Barraud[1], V. Lapras[1], B. Previtali[1], M.P. Samson[2], J. Lacord[1], S. Martinie[1], M.-A. Jaud[1], S. Athanasiou[1], F. Triozon[1], O. Rozeau[1], J.M. Hartmann[1], C. Vizioz[1], C. Comboroure[2], F. Andrieu[1], J.C. Barbé[1], M. Vinet[1], and T. Ernst[1]
[1]CEA, LETI, MINATEC campus and Univ. Grenoble Alpes, 38054 Grenoble, France, email: sylvain.barraud@cea.fr
[2] STMicroelectronics, 850 rue J. Monnet, 38920 Crolles, France

***Abstract*—**This paper presents recent progress on Gate-All-Around (GAA) stacked-NanoWire (NW) / NanoSheet (NS) MOSFETs. Key technological challenges will be discussed and recent research results presented. Width-dependent carrier mobility in Si NW/NS and FinFET will be analyzed, and intrinsic performance and design considerations of GAA structures will be discussed and compared to FinFET devices with a focus on electrostatics, parasitic capacitances and different layout options. The results show that more flexibility can be achieved with stacked-NS transistors in order to manage power-performance optimization.

I. INTRODUCTION

Gate-All-Around (GAA) FETs have long been recognized as offering the best solution to short-channel-effects (SCE) with a high current drivability per layout footprint due to 3D vertically stacked channels [1-4]. Moreover, horizontal GAA NW and NS also have the advantage of being fabricated with minimal deviation from FinFET (FF) devices in contrast to vertical NWs which require more disruptive technological changes. For these reasons, the GAA stacked-wire MOSFET architecture is today regarded as an attractive option to push CMOS scaling beyond 7/5nm nodes. Although the first 3D GAA stacked-wires transistors were demonstrated ten years ago [1-4], significant progress have been reported last year [5-6] with aggressive 44/48nm CPP (Contacted Poly Pitch) ground rules [7]. In this paper, we will first discuss recent progress on the fabrication of GAA stacked-wires FETs. Then, we shall be paying particular attention to intrinsic performances and design considerations of GAA structures (NW and NS) for an optimal performance and power efficiency *vs*. FF.

II. SCALING AND DEVICE ARCHITECTURE

For several decades, the Si CMOS technology has enabled manufacturers to produce integrated circuits with ever-increasing levels of performance and functionality. For example, **Fig. 1** presents the scaling rules of CPP and metal pitch (MP) from 90nm-*Bulk* to 7nm-*FinFET* architectures. If chipmakers seek to extend FinFET scaling at the 5nm node and beyond, they will likely need to move to the evolutionary alternative of GAA MOSFET architecture in order to keep low leakage current as the gate length gets smaller. Over the last year, GAA stacked-NW/NS MOSFETs were successfully demonstrated with a replacement metal gate (RMG) process [5-7]. GAA stacked-NWs FETs, which are close to the RMG FinFET technology, have nevertheless specific technical requirements. They are numbered "1 to 5" in **Fig. 2**. The fabrication started with the epitaxial growth of ($Si_{0.7}Ge_{0.3}$/*Si*) multilayers with ideally sharp interfaces. Then, dense arrays of fins were patterned to fabricate stacked-wires FETs. Multiple patterning techniques were used in order to meet the density targets of advanced nodes. After that, dummy gates and spacers were defined prior to the anisotropic etching of the (*SiGe/Si*) multilayers. Then, the SiGe layers were partially etched selectively to the Si ones to form inner spacers well-aligned and correctly dimensioned as shown in Fig. 2. Such spacers are essential in order to minimize parasitic capacitance. Finally, the Si wires were released during the RMG module prior to conformal HfO_2/TiN/W gate deposition. Nanowire widths up to 50nm [7] and even 75nm [6] were reported and the possibility to stack up to 13 NW levels without any strain relaxation was demonstrated [8]. In order to evaluate the benefits of GAA structures (NW and NS), the carrier mobility, electrostatics and parasitics should holistically be assessed from a power/performance optimization perspective. This is discussed in the next sections. **Fig. 3** shows the guideline proposed to benchmark FinFET and GAA architectures. The dimensions of devices are consistent with recent experimental results reported in ref. [7]. The layout footprint (LF), the fin and stack thickness as well as the fin-to-fin space ($S=\{LF-N\times W_{Fin}\}/\{N-1\}$) are assumed constant for the three devices.

III. CARRIER MOBILITY

Although electrostatic control is excellent in GAA NW FETs as the gate length becomes smaller, significant changes in transport properties associated to atomically small (<10nm) dimensions can alter device performance. The size-dependent carrier mobility in 3D multi-gate devices is primarily governed by facet-dominated transport [9-10] with high (resp. low) electron mobility in the (100) (resp. (110)) plane and high (resp. low) hole mobility in the (110) (resp. (100)) plane. Meanwhile, mobility in conventional NWs is often the worst due to additional quantum confinement effects resulting in rapid degradation in mobility at smaller size due to phonon and surface roughness scattering [11-14] (**Fig. 4**). Basically, as shown in **Figs. 5-6**, the computed carrier mobility interpolates between the Si NW and thin film limits [15]. For electrons, [110] n-FETs with width W>H show larger mobilities than those with H>W. This can be explained by band structure effects [16]. The Δ valleys split into light ($m^*=0.19m_0$) Δ_z valleys at Γ and heavier ($m^*=0.55m_0$) $\Delta_{x,y}$ valley off Γ. In the strong inversion regime, the electron gas is mostly confined in the light Δ_z valleys on the top and bottom (001) facets, and in the heavier $\Delta_{x,y}$ valleys on the lateral (110) facets. Therefore, wide [110] Si NS (W_{NS}>20nm) with dominant (001) facets

(W>H) perform better than tall [110] Si fin with dominant (110) facets (W<H). The trends are opposite for p-FETs, with a significant hole mobility improvement in [110] p-FET with H>W. Horizontal GAA NS for n-FETs and vertical GAA NS for p-FETs turn out to be the most effective solutions to promote electron and hole transport, respectively.

IV. PERFORMANCE AND DESIGN CONSIDERATIONS

In early days of CMOS, the FinFET technology alleviated several important challenges associated with the aggressive scaling of planar bulk CMOS devices. Such a strategy allowed to preserve robustness to SCE, but more importantly resulted in improved circuit delay owing to 3D integration and vertical channel orientation. This allowed the transistors to deliver more performance due to higher effective width (W_{eff}) per footprint. This idea must continue to be considered for GAA structures [17-18]. As shown in **Fig. 7**, the GAA NS structures could instead be used to maximize the effective width for a given layout footprint, which will improve the drive current without increasing power density. As compared to FinFET, the conventional square (or round) NW has a lower W_{eff} for a given layout footprint while the effective width can be significantly enhanced with wide and thin NS compared to conventional FinFETs, as shown in **Fig. 8**. A large footprint allows wider NS which yields higher W_{eff} and ultimately better performances while maintaining Drain-Induced-Barrier-Lowering (DIBL) lower than in short-channel FinFET devices, as shown in **Fig. 9**. Basically, the DIBL *vs.* W_{eff} for a given footprint are plotted in **Fig. 10** to clearly emphasize the trade-off between the channel electrical width which govern the intrinsic performance of devices (modulated by the carrier mobility discussed above) and the immunity to SCE. In considering three layout footprints and a 16nm gate length, Fig. 9 shows that GAA stacked-NS can be thought of as a practical compromise between speed and power dissipation. However, to make the best of GAA stacked-NS structures, effective current I_{eff} enhancement achieved by higher μ_{eff} and W_{eff} should not be canceled by an excessive increase in parasitic capacitances. This argument is often to the fore when arguing against the development of stacked-NW/NS transistors. Indeed, a significant increase of parasitic capacitance will impact circuit performance by increasing the load capacitance C_{eq} defined in **Fig. 11** and increasing the switching delay τ_p. **Fig. 12** summarizes the calculation of *gate-to-channel* capacitance (C_{inv}) in inversion regime and *gate-to-drain* capacitance (C_{gd0}) of FinFET and GAA structures used in this work. While there are inevitably an increase of C_{inv} for wider GAA NS structures due to increased W_{eff}, the C_{gd0} capacitance is less sensitive to a change in width of NS [18]. The rate of increase in C_{inv} and C_{gd0} capacitances for FinFET is shown in **Fig. 13**. +28% improvement in C_{inv} is achieved against only 6% in C_{gd0} at 40nm width (W_{NS}). Since the *gate-to-drain* capacitance depends slowly on W_{NS}, and assuming the same wiring load capacitance of back-end of line ($C_{back\text{-}end}$) for the three devices (FF, NW and NS), the metric $C_{eq}\times V_{DD}/I_{eff}$ must be improved at constant layout footprint. I_{eff} represent the effective current and is defined in Fig. 11. The effective capacitance C_{eq} defined as the sum of the three contributions (C_{inv}, C_{gd0} and $C_{back\text{-}end}$) is shown in **Fig. 14**. The larger the footprint will be (allowing wide width of NS), the larger C_{eq} will be due to the effective width enhancement. However, it should be noted that C_{eq} normalized by W_{eff} can be significantly reduced for wide NS widths. Likewise, the switching delay has been calculated for different layout footprints. The delay and C_{eq} reduction over FinFET is shown in **Fig. 15** for LF=57nm (resp. 82nm) related to a 3-fins library cell (resp. 4-fins library cell). C_{eq} is reduced for NWs (W=7nm) but no delay reduction is achieved, while performance can be significantly improved for nanosheet design having wider wires. A delay reduction of around 20% is expected for W_{NS}~30nm. As shown in **Fig. 16** and **Fig. 17**, nanosheets have a more effective width for a given footprint and therefore drive a capacitive load better [6,17,18]. The benefit offered by GAA stacked-NS enabling to relax the fin pitch by using double or triple stack to match or overcome the effective width at constant footprint may be reduced when considering ultra-scaled standard cell height with small fin pitch and few fins. The reduction of fin number and fin pitch (**Fig. 18**) will result in lower NS width: the W_{eff} enhancement and the τ_p reduction may then be minimized. Finally, a fine tuning of NS width with EUV lithography would tentatively result in improved power/performance management through the modulation of threshold voltage and subthreshold slope, as shown in **Fig. 19**. For a given cell height, **Fig. 20** give an example of 2-input NAND gate with two possible wire widths (7 and 30nm).

V. CONCLUSION

Obviously, the technological challenges facing the development of a new technology platform featuring GAA nanosheet transistors are still numerous (shape optimization with reduced roughness, inner spacer, access optimization and strain management). Nonetheless, some significant experimental advances have been made recently and show the high competitiveness of this technology for future technology nodes. Nanosheet transistors offers more freedom to designers for the power-performance optimization thanks to a fine tuning of the device width.

ACKNOWLEDGMENT

This work was partly funded by the French Public Authorities through the NANO 2017 program and EQUIPEX FDSOI11. It is also partially funded by the SUPERAID7 (grant N° 688101) project.

REFERENCES

[1] T. Ernst *et al.*, IEDM, 10.1109/IEDM.2006.346955, 2006, [2] L.K. Bera *et al.*, IEDM, 10.1109/IEDM.2006.346841, 2006, [3] C. Dupré *et al.*, IEDM, 10.1109/IEDM.2008.4796805, 2008, [4] E. Bernard *et al.*, VLSI, pp. 16-17, 2008, [5] H. Mertens *et al.*, VLSI, 10.1109/VLSIT.2016.7573416, 2016, [6] S. Barraud *et al.*, IEDM, 10.1109/IEDM.2016.7838441, 2016, [7] N. Loubet *et al.*, VLSI, 10.1109/VLSIT.2017, 2017, [8] T. Ernst *et al.*, Micro. Eng., vol. 88, pp. 1198-1202, 2011, [9] J. Chen *et al.*, VLSI, pp. 32-33, 2008, [10] J. Chen *et al.*, VLSI, pp.175-176, 2010, [11] Y.M. Niquet *et al.*, ULIS, pp.49-52, 2012, [12] N. Neophytou *et al.*, PRB, 84, pp. 085313-185313-15, 2011, [13] S. Jin *et al.*, JAP, 102, pp.083715—1 083715-14, 2007, [14] M.V. Fischetti *et al.*, JAP, 80, pp. 2234-2252, 1996, [15] Z. Zaiping *et al.*, IEEE TED, 64, pp. 2485-2491, 2017, [16] Y.M. Niquet *et al.*, NanoLetters, 12, pp. 3545-3550, 2012, [17] S.-D. Kim *et al.*, S3S, 10.1109/S3S.2015.73335212015, [18] L. Gaben *et al.*, SSDM, 2015, [19] J. Lacord *et al.*, IEEE TED, 63, pp. 781-786, 2016.

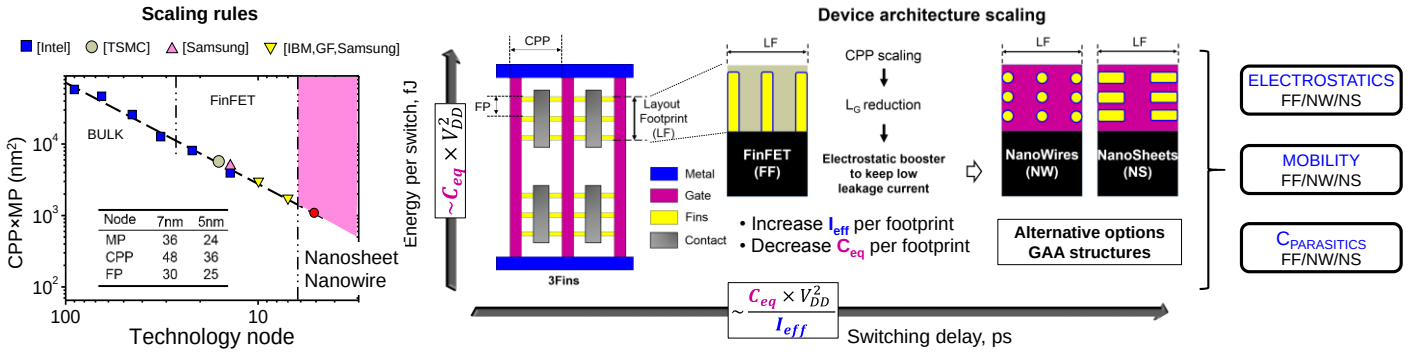

Fig.1: (Left) Standard cell area scaling *vs.* technology node. (Right) The CMOS scaling is governed by the Contacted Poly Pitch (CPP) in the *x*-direction and the Metal Pitch (MP) in the *y*-direction. At 5nm node, two GAA structures (NW and NS) can be proposed as an alternative to FinFET. Electrostatics confinement, carrier mobility and parasitics need to be investigated in order to minimize the effective load capacitance (C_{eq}) and/or maximize the effective drive current (I_{eff}).

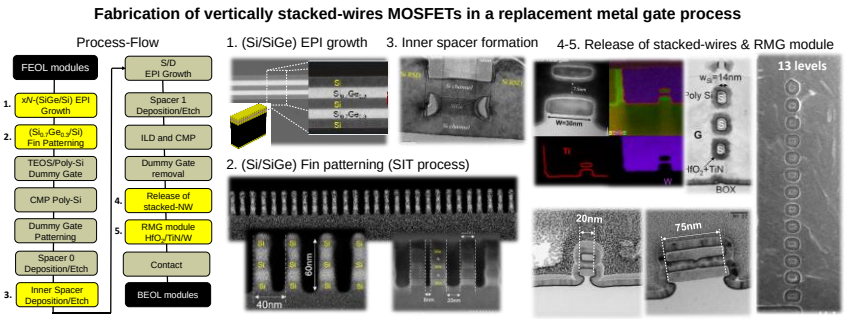

Fig.2: Process flow of stacked-NW/NS FETs. The steps numbered '1' to '5' are specific technical requirements for NW/NS FETs (as compared to FinFET devices). Cross-sectional Transmission Electron Microscopy (TEM) images are shown at various stages of the fabrication process. (1) Growth of (Si/SiGe) superlattices with 3 levels of Si layers stacked upon one another; (2) Etching of (Si/SiGe) fins with a SIT process; (3) Stacked-wires FETs after the integration of inner spacers and (4-5) stacked-NS FETs with a HfO_2/TiN/W gate stack with W_{NS} up to 75nm.

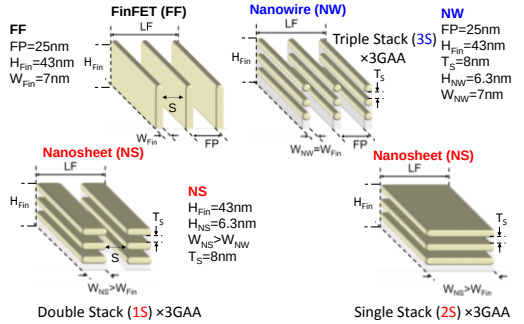

Fig.3: Guideline for benchmarking FinFET and GAA stacked-NW/NS architectures. For a given footprint (LF) and Fin thickness (H_{Fin}), several GAA Nanosheets structures (single-, double-, triple-stack, etc.) can be considered to overcome FinFET performance thanks to higher effective width. S is defined as $(LF-N\times W_{Fin})/(N-1)$.

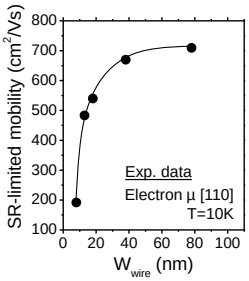

Fig.4: Experimental data of surface-roughness limited mobility ($N_{inv}=10^{13}$ cm^{-2}) *vs.* W_{wire} at 10K.

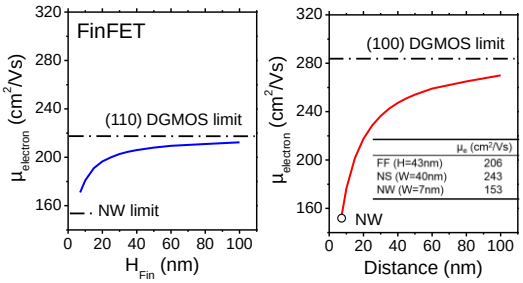

Fig.5: (Left) $\mu_{electron}$ ($N_{inv}=10^{13}$cm^{-2}) *vs.* H_{Fin} for [110] FinFET transistors (W=7nm). (Right) $\mu_{electron}$ *vs.* W_{NS} for [110] GAA NW and NS FETs (H=7nm). The horizontal doted lines are the reference in (110) and (100) double-gate FET.

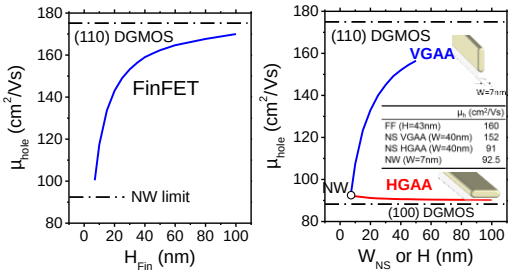

Fig.6: (Left) μ_{hole} ($N_{inv}=10^{13}$cm^{-2}) *vs.* H_{Fin} for [110] FinFET transistors (W=7nm). (Right) μ_{hole} *vs.* W_{NS} or H for [110] horizontal (HGAA, H=7nm) and vertical (VGAA, W=7nm) GAA FETs.

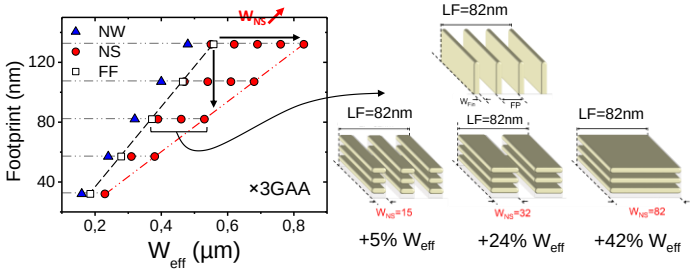

Fig.7: In contrast to NWs and FinFETs, a significant increase in W_{eff} can be achieved with GAA NS transistors for a given footprint (LF). The wider W_{NS} will be, the higher the effective width enhancement will be. Computations done for a stack of 3 GAA with the dimensions of devices given in Fig. 3.

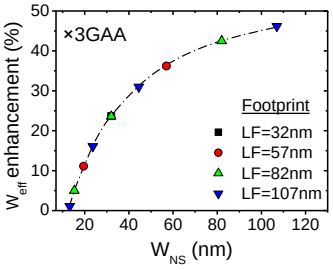

Fig.8: W_{eff} enhancement (compared to FF with FP=25 nm) *vs.* W_{NS} for different layout footprints (LF). Here, $H_{NS}=H_{NW}=6.3$nm (cf. Fig. 3).

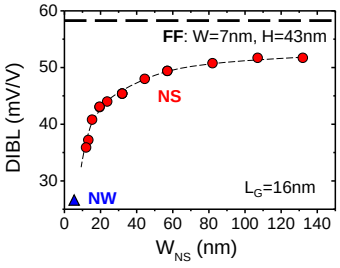

Fig.9: DIBL *vs.* W_{NS}. A fine tuning of W_{NS} allows a DIBL modulation between NW and FF. Here, $H_{NS}=H_{NW}=6.3$nm (cf. Fig. 3).

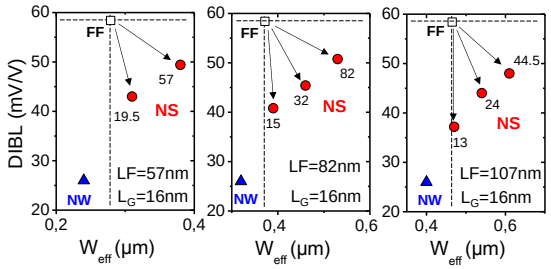

Fig.10: DIBL *vs.* W_{eff} for three layout footprints. A better compromise between high W_{eff} and low DIBL is achieved for GAA stacked-NS structures. If low DIBL is shown for NW, W_{eff} is strongly reduced *vs* NS. H_{NS}=H_{NW}=6.3nm (cf. Fig. 3).

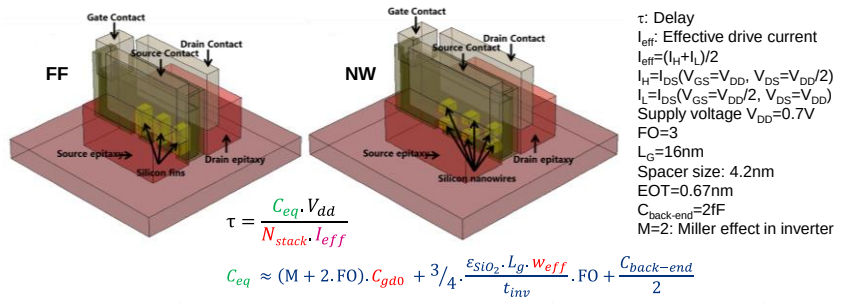

Fig.11: Bird view of FinFET and stacked-NW structures used for TCAD simulations of parasitic capacitances and delay computation. The effective load capacitance C_{eq} results from (i) the gate-drain capacitance C_{gd0}, (ii) the inversion capacitance and (iii) the back-end component [19]. A thin SiN spacer width is considered (W_{spacer}=4.2nm). Other device dimensions are given in Fig. 3.

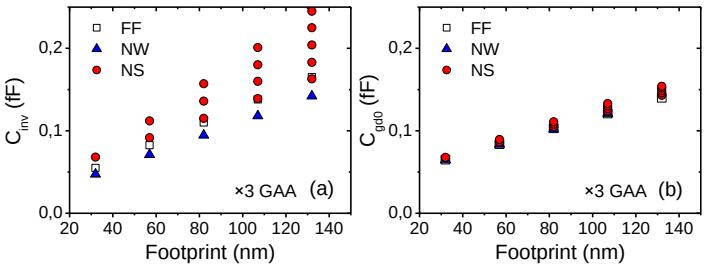

Fig.12: Inversion (a) and gate-to-drain (b) capacitances *vs.* layout footprint for FF and stacked-NW/NS structures. As compared to FF (with FP=25nm) devices, an increase of W_{NS} for a constant footprint results in a C_{inv} enhancement due to higher W_{eff}. However, a small improvement of C_{gd0} is observed as W_{NS} increases.

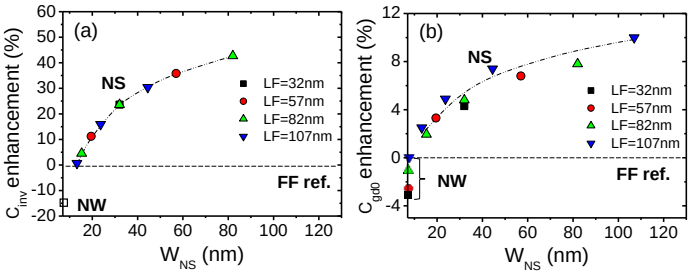

Fig.13: C_{inv} (a) and C_{gd0} (b) enhancement over FinFET (with FP=25nm) for stacked-NS/NW structures. Capacitances are calculated for different layout footprints (LF). Whatever the LF, C_{inv} and C_{gd0} increase with W_{NS} (+ 28% improvement for C_{inv} against only + 6% for C_{gd0} at W_{NS}=40nm).

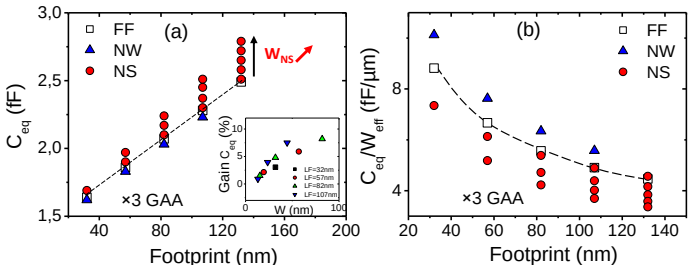

Fig.14: (a) Effective load capacitance C_{eq} as a function of LF for FF (with FP=25nm) and stacked-NW/NS structures. The larger the footprint is, the wider the NS will be, resulting in larger C_{inv} contribution at high W_{eff}. However, the normalized effective load capacitance is significantly reduced for NS (b).

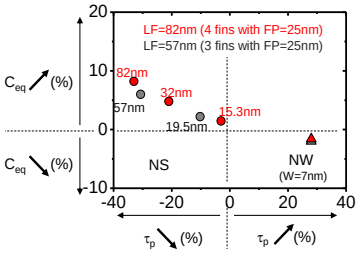

Fig.15: τ_p and C_{eq} reduction over FinEET (with FP=25nm) technology for different LF. A τ_p reduction of ~20 % is expected at W_{NS}=30nm.

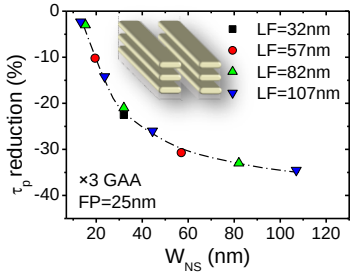

Fig.16: Delay reduction over FF technology (with FP=25nm) for different footprints (LF). The delay reduction is mainly dependent on W_{NS}.

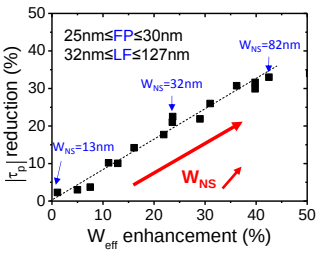

Fig.17: Whatever the FP and LF values, τ_p is reduced in NS as W_{eff} increases, with a clear improvement over FinFET devices. Wider NS yield reduced τ_p.

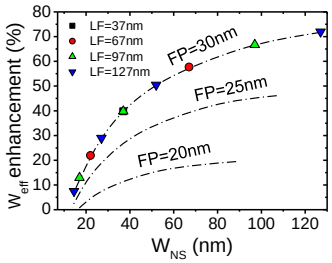

Fig.18: Effective width enhancement (compared to FF devices related to different fin pitch FP) *vs.* W_{NS} for different layout footprints (LF). The fin pitch reduction results in lower W_{NS} values and therefore lower effective width. The switching delay reduction will then be less.

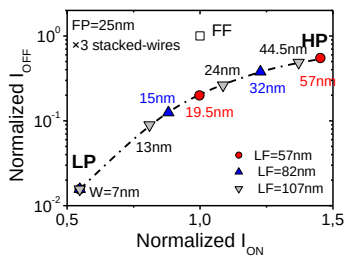

Fig.19: Normalized I_{OFF} *vs.* I_{ON} current (over FinFET with FP=25nm) for L_G=16nm. A tuning of the NS width allows one to modulate the threshold voltage and the substhreshold slope, enabling a better performance-power optimization.

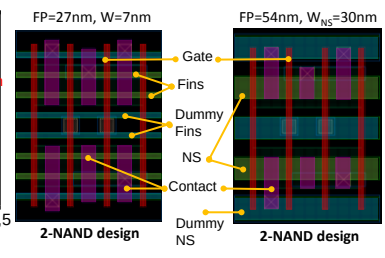

Fig.20: Layout example of 2-input NAND gates for NW (left) and NS (right) configurations. For the same cell heigth, the Fin/stack patterning can be relaxed for NS (FP=54nm against 27nm) in order to have wider NW (30nm), and therefore several design options to manage power consumption and performance as shown in Fig. 19.