Vertical Stacking Possibilities of GAA-NSs: Role of MoL/Contact Module for Continued Tau (τ)-Scaling

Himanshu Sharma
Huawei Technologies R&D
Leuven, Belgium
himanshu.sharma1@huawei.com

Krishna K. Bhuwalka
Huawei Technologies R&D
Leuven, Belgium
krishna.bhuwalka@huawei.com

Dmitry Yakimets
Huawei Technologies R&D
Leuven, Belgium
dmitry.yakimets@huawei.com

Pieter Schuddinck
Huawei Technologies R&D
Leuven, Belgium
pieter.schuddinck@huawei.com

Gerhard Rzepa
Global TCAD Solutions
Vienna, Austria
g.rzepa@globaltcad.com

Lei Hou
Huawei Technologies R&D
Leuven, Belgium
houlei17@hisilicon.com

Ming Jin
Huawei Technologies R&D
Leuven, Belgium
jinming17@hisilicon.com

Markus Kraner
Global TCAD Solutions
Vienna, Austria
m.karner@globaltcad.com

Zheng Yu
Huawei Technologies R&D
Leuven, Belgium
yuzheng1@hisilicon.com

***Abstract*— We investigate the underlying factors limiting optimal performance of state-of-art gate-all-around nanosheets (GAA NS) to 3 or 4 stacked sheets (N_{NS}) and how such factors can be overcome to enhance τ-(RC) scaling using taller stacks. Thus, device external resistance dominated by current crowding in source-drain regions (R_{SD}), rather than capacitance is identified as most detrimental factor. Source/drain resistance can be improved by full wrap-around contact (WAC) (with capacitance penalty) or by replacement of S/D epi with metal (without capacitance penalty). We thus, show that improving S/D resistance can allow taller stacks with near ideal vertical scaling (up to 8 stacked sheets in this study). Impact of pFET stress loss for metal S/D is further investigated along with benefits for thermal heat dissipation.**

***Keywords*— GAA-NS, Nanosheet, τ-Scaling, Tau-Scaling, Stack**

I. Introduction

Gate-all-around nanosheet devices have been introduced at 3nm/2nm logic technology nodes as a key enabler for continued transistor delay (τ) scaling [1-3]. A key question arises in terms of scalability and multiple node extendibility. Multiple prior works have focused on metal-pitch (M_x) and standard-cell height (CH) scaling, which has resulted in innovations like the fork-sheet (FS) devices [4] and (complementary) C-FETs [5]. Although these devices allow cell height scaling, their vertical stacking seems limited to 3 or 4 nanosheets per device, beyond which severe degradation in speed and power is observed [6]. Sub-threshold swing (S_{SAT}) improvement has been studied by reducing the Si-sheet thickness down to 3 nm for a gate length of 6 nm [7, 8]. While cell height and intrinsic device scaling are important for performance boost, feasibility of vertical stacking of nanosheets have not been investigated in detail, though there has been experimental demonstration of working devices with up to 7 sheets [9, 10]. Furthermore, GAA NS devices suffer from degraded self-heating (SH) effects due to isolated gates [11]. In taller nanosheet stacks, these impacts are expected to get worse resulting in significant challenges for thermal management and reliable device operation. Thus, there is need to understand SH in taller devices and ways to mitigate the impact.

In this paper, using a well calibrated TCAD based DTCO flow [6], we do an in-depth analysis of nanosheet stack scaling. Intrinsic and extrinsic resistance-capacitance (RC) components are decoupled to identify performance limiting factors with vertical stack scaling. Thus, we show that while intrinsic channel resistance ($R_{CHANNEL}$) and effective circuit capacitance (C_{EFF}) scale nearly linearly with stacking, the increase in source/drain resistance (R_{SD}) (due to non-uniform current distribution in the epitaxial region) degrades performance as the number of sheets (N_{NS}) increases. This resistance issue can be mitigated by (i) wrapped around contact (partially mitigated); or by (ii) fully replacing Si/SiGe epitaxy with metal source-drain (fully mitigated) (schematic in Fig. 1). Analysis of the S/D metal fill is further extended to include the impact of stress loss (due to epi-replacement) as well as self-heating benefits (due to low-resistive metal pillar).

II. TCAD/DTCO Methodology and Device Parameters

The semi-classical TCAD/DTCO flow used for device parameters extraction, calibration and experimental validity from physics-based to drift-diffusion and SPICE simulations have been previously discussed and reported in [6,8] for NS channel thickness T_{NS} ranging from 3 ~ 6 nm and gate-length L_G of 6 ~ 15 nm.

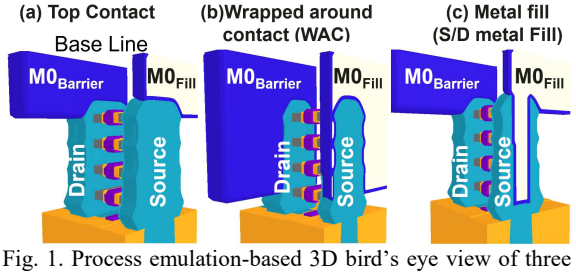

Fig. 1. Process emulation-based 3D bird's eye view of three different types of contact schemes for a 4-sheet GAA-NS device. (a) Top-contact over source/drain epitaxy is conventional contact scheme; (b) Fully wrapped around contact (WAC) is an advanced contact scheme to lower contact resistivity; (c) full source/drain metal fill scheme is new proposal in this manuscript. Here source cross-sections are shown, and M0 barrier is removed on the source side purely for illustrative purposes.

Baseline NS device and standard cell dimensions are discussed in detail in [6]. For this work, intrinsic device with

T_{NS} 5 nm, L_G 12 nm with junction ~2.25 nm/dec and spacer thickness T_{SP} of 6 nm are used (fixed CPP 45 nm). Baseline NS width (W_{NS}) is 30nm and N_{NS} is 3 with a constant vertical pitch of 15 nm. In this work, we focus on vertical stack-scaling and its dependency on middle-of-line (MoL) and source/drain contact module. In our experiments, W_{NS} is varied from 15 nm to 35 nm and nanosheet stack N_{NS} is varied from 2 to 8. For each case, source/drain recess for epi-growth is adjusted to reach the bottom of fin. The same criterion is applied for WAC and full S/D metal contact schemes. All the devices are renormalized (Vt-adjusted) for a fixed I_{OFF} of 1 nA/device at V_{DD}=0.75V. For the sake of completeness, it should be emphasized that current crowding effects in the S/D epitaxy (high doping of $1x10^{21}cm^{-3}$), metal-semiconductor Schottky barrier (with interface resistivity of $1x10^{-9}$ Ω-cm^2) and MoL and S/D metal resistivity are all taken into account, which are critical for this work and its accuracy.

III. Results and Discussion

Fig. 2 shows the speed-at-iso-leakage data for stack number scaling from 2 to 8 stacks per NS when baseline top-contact is used. As can be seen, an optimal frequency at 4 sheets is achieved, consistent with previous reported work [6], beyond which performance drops severely for all W_{NS}.

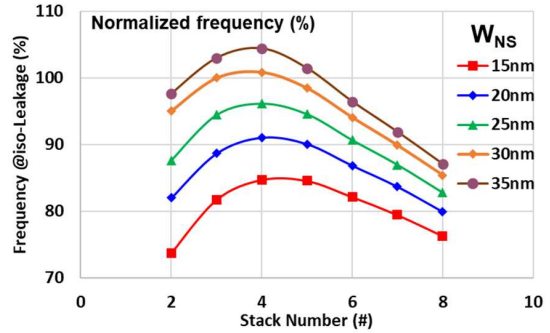

Fig. 2. Normalized frequency-at-iso-leakage (V_{DD} 0.75V) for multiple W_{NS} and N_{NS} scaling from 2 to 8. Data is an extension of results to 8-sheet stacks from previous study [6] (5-stage RO with FO3).

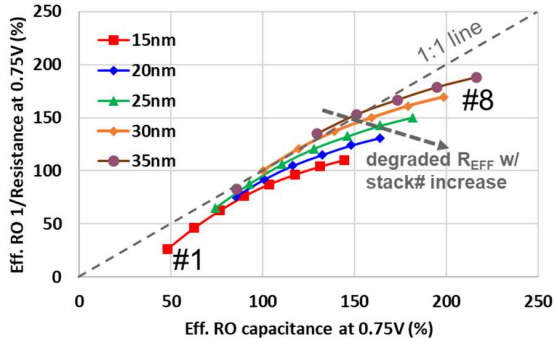

Fig. 3. Normalized ring-oscillator (RO) $1/R_{EFF}$ vs. C_{EFF} for multiple number of stacked sheets (N_{NS} = 1-8) and W_{NS} (15-35 nm).

We further investigate the impact with R-C component breakdown. Fig. 3 shows the SPICE simulated ring-oscillator's (RO) effective circuit current I_{EFF} (~V_{DD}/R_{EFF}) vs. effective capacitance C_{EFF} at 0.75V. While C_{EFF} increases near linearly with stack number scaling, I_{EFF} does not increase as expected. This leads to a severely degraded performance at taller stacks.

A breakdown of device (external resistance R_{EXT} and channel resistance $R_{CHANNEL}$) and circuit (R_{EFF} and C_{EFF}) RC as a function of effective width W_{EFF} is shown in Fig. 4. R_{EXT} is gate-voltage independent device parasitic resistance (which include S/D, overlap resistance and contact interface resistance) while $R_{CHANNEL}$ is gate-voltage dependent resistance which is extracted from $R_{TOTAL}(V_{GS}) = V_{DD}/I_{ON} = R_{EXT}+R_{CHANNEL}(V_{GS})$ [8]. W_{EFF} is defined as total perimeter width of NS such that $W_{EFF} = 2N_{NS}(W_{NS}+T_{NS})$. While R_{EFF} and C_{EFF} are extracted from SPICE simulations of RO, R_{EXT} and $R_{CHANNEL}$ are extracted from device TCAD for same conditions. As can be seen, while $1/R_{CHANNEL}$, $1/R_{EFF}$ and C_{EFF} scale near linearly, R_{EXT} increases as W_{EFF} increases (effectively as N_{NS} increases). The data is shown for nFETs, but pFETs follow similar trends. This is more clearly visible in Fig. 5 where the current density for nFET is shown for 3 sheets and 6 sheets cases. Taller stacks result in potential drop in the S/D epitaxy region, leading to lower current density in the bottom sheets.

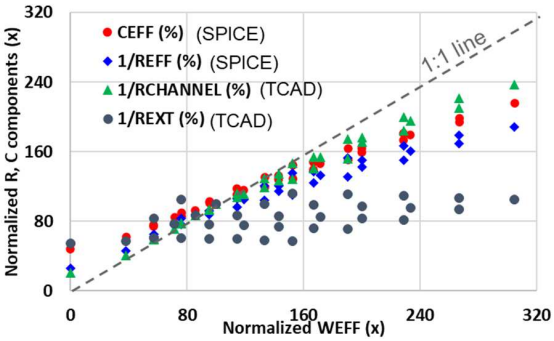

Fig. 4. Normalized RC components for NS with 2~8 stacks and WNS from 15nm to 35nm. As can be seen, device R_{EXT} scales worst amongst all components.

Two possible ways are proposed to improve current density in bottom sheets, as discussed in Fig. 1. A WAC with silicide and M0 contact metal extending to the bottom of S/D can be done to mitigate some effects. This can be achieved by conformal deposition of titanium using a chemical vapor deposition (CVD) process or advanced atomic layer deposition (ALD) process. Alternatively, etching a hole in the epi after formation and RMG process, and subsequent deposition of filling Ti can also be used. We discuss pros and cons of each MoL process.

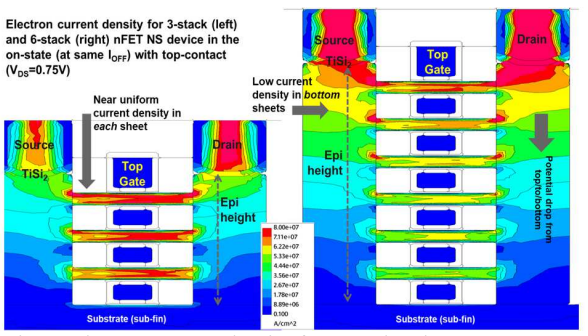

Fig. 5. Electron current density for 3- and 6-sheet nFETs NSs. The 3-sheet stack shows more uniform current distribution while in the 6-sheet stack, current density drops for lower sheets, resulting in overall performance drop.

As can be seen in Fig. 6, current density improves for WAC (centre) and S/D metal contact (right) over baseline top-contact case

(left) due to improved carrier injection and reduced contact resistance.

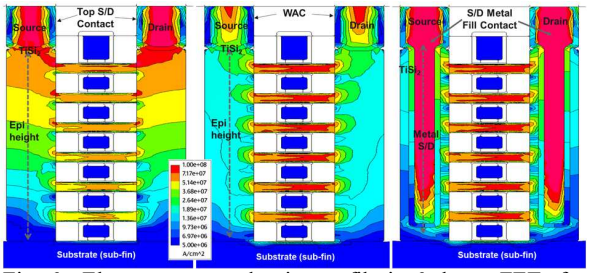

Fig. 6. Electron current density profile in 6-sheet nFETs for three different contact schemes as shown in Fig. 1. For each case, the data is extracted at constant off-current.

Fig. 7 shows the normalized RO frequency at iso-leakage for GAA NS devices with N_{NS} scaling from 2 to 8 for the three distinct contact schemes. As can be seen, while for conventional top-contact, optimal performance is achieved for a stack of 3 or 4 sheets, near ideal RC scaling is achieved for WAC and Metal S/D Fill contact schemes.

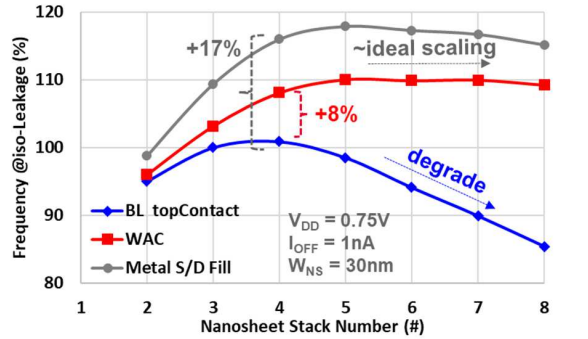

Fig. 7. Normalized frequency-at-iso-leakage for GAA NS devices with N_{NS} increasing from 2 to 8 for the three MoL contact schemes. For the baseline optimum of 4 sheets, the performance increases by 8% for WAC and by 17% for the Metal fill S/D contact scheme. Beyond 4 sheets, near ideal-scaling is achieved for both optimized contact schemes.

Thus, for a stack of 4 sheets, ~8% and 17% speed improvement is achieved for the proposed two MoL cases. Beyond 4 sheets, performance does not degrade, but remains constant, indicating near linear RC behavior with W_{EFF}. Thus, normalized ($1/R_{EXT}$) is shown in Fig. 8 for multiple N_{NS} and W_{NS} as a function of W_{EFF}.

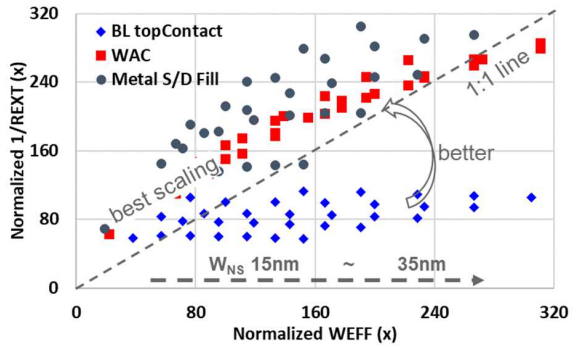

Fig. 8. $1/R_{EXT}$ (extracted from TCAD) for all three contact modules is shown, for multiple N_{NS} and W_{NS}. As can be seen, both WAC and Metal S/D Fill follow better than W_{EFF} scaling, thereby improving performance.

As can be seen, WAC and Metal S/D Fill cases show better than linear scaling with increasing W_{EFF}. Metal S/D Fill shows even better performance than WAC, due to the linear increase of metal fill area with increasing WNS, resulting in lower access resistance. Further, WAC comes with an added C_{EFF} penalty as is shown in Fig. 9. Wrapping contact around S/D epi increases the parallel plate area between S/D and gate lines, resulting in C_{EFF} penalty. In case of Metal S/D Fill case, as it only replaces some S/D epi, there is no change in C_{EFF}, leading to maximal performance gain.

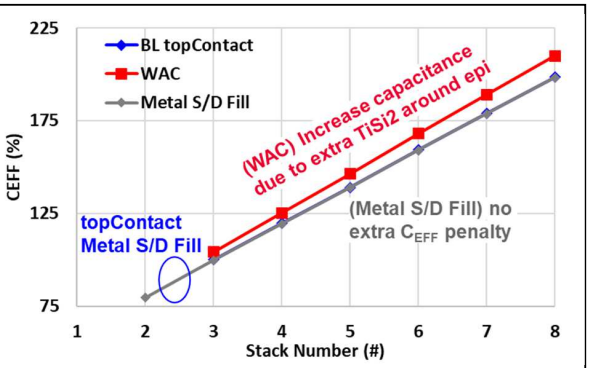

Fig. 9. RO C_{EFF} for three contact schemes for W_{NS} 30nm and N_{NS} scaling from 2 to 8. C_{EFF} increases linearly with N_{NS}, however, WAC shows larger C_{EFF} resulting in degraded performance.

Due to the presence of inner spacers in GAA NS devices, S/D epi growth happens from multiple disconnected seeding locations, which are more likely to result in S/D epitaxy with dislocations, which can lead to stress loss. We have discussed the implications of stress loss on GAA performance and trade-off scenarios in detail for a 3-stack NS device [12]. Replacing S/D epitaxy with metal S/D might lead to further loss of stress in the pFETs, in case a high quality epitaxy can be grown despite the inner-spacers. The resulting frequency-at-iso-leakage is shown in Fig. 10. As can be seen, stress loss (-1.5 GPa to 0 GPa) for pFETs results in roughly 16% performance loss at baseline stack of 3. On the other hand, at tall stacks, beyond 4 sheets, better scaling is achieved with metal S/D even with cases where full stress can be maintained in conventional GAA NS devices with top-contact. The trade-off between degraded $R_{CHANNEL}$ and improved R_{EXT} being compensated at 4 sheets, beyond which there is significant performance gain with metal S/D cases at taller stacks.

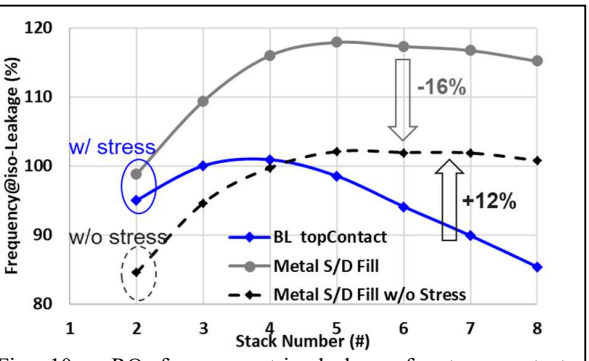

Fig. 10. RO frequency-at-iso-leakage for two contact schemes: Top Contact and metal S/D fill with stress (- 1.5 GPa) and Metal S/D Fill without stress (0 GPa) as the number of GAA stack sheets scales from 2 to 8.

Finally, we further show the advantages of metal S/D from self-heating perspective. Fig. 11 illustrates the heat flux density and temperature distribution for nFET devices with conventional top contact (baseline) and with the replacement metal S/D. The heat flux density is noticeably higher in the metal S/D case, indicating more efficient heat dissipation through the low-resistivity metal compared to Silicon or SiGe. Although the electron current density and thus electron-induced local heat is higher in the metal S/D case, its low electrical resistivity limits heat generation, and its high thermal conductivity efficiently removes heat, resulting in a lower peak temperature.

The temperature map further shows that the thermal spread is larger in the top S/D contact, while the metal S/D fill confines and directs heat more effectively, resulting in a narrowed thermal profile across the nanosheet stack.

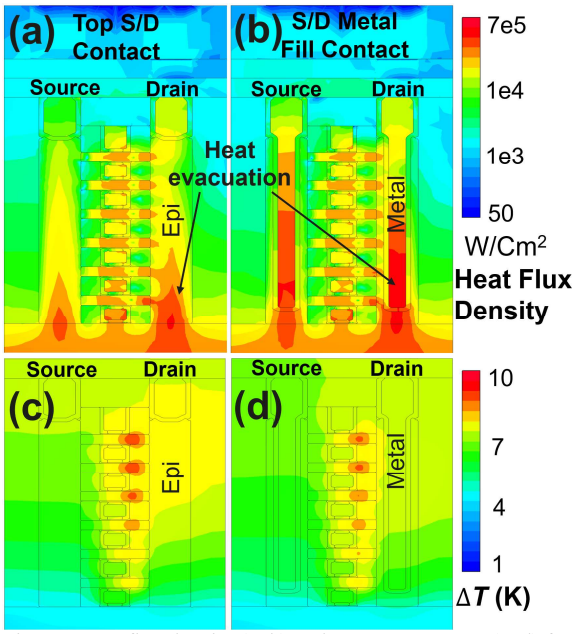

Fig. 11. Heat flux density (a–b) and temperature map (c–d) for nFET devices with top-contact and metal-filled S/D, highlighting heat evacuation through epitaxy versus metal and also the resulting thermal distribution across the channel and S/D regions. In these simulations, we assume that 100% of the generated heat is dissipated through the substrate, without introducing additional thermal resistances within the device.

Conclusions

In this paper, we have studied and identified limiting factors for vertical scaling of stacks in GAA nanosheet based devices. We showed that the performance is limited by increased source/drain resistance in taller stacks rather than capacitance. Thus, we further showed that using a wrapped-around contact (WAC) scheme can improve vertical scaling and performance (~8%), albeit at the cost of increased capacitance. Alternatively, replacing S/D epitaxy by metal results in increased performance by ~17% without C_{EFF} penalty. Both WAC and replacement S/D epitaxy schemes maintain near-linear RC scaling beyond 4 sheets. Impact of Stress loss in pFETs in case of replacement metal is further investigated. It is observed, that beyond 4 sheets, resistance trade-offs enable better overall performance with replacement metal compared to conventional top contact with stress. Additionally, from a thermal perspective, replacement metal S/D enhances heat dissipation, as confirmed by heat flux density and temperature maps. Overall, careful S/D engineering enables taller nanosheet stacks with higher speed, reduced self-heating, and balanced parasitic and strain trade-offs, providing a practical pathway for high-performance, thermally robust GAA nanosheet devices for beyond 2nm technology nodes.

References

[1] Tingbo He, "New Semiconductor Path in Practice", presented at the IEEE Int. Symp. Circuits and Systems (ISCAS), Shanghai, China, 2026.

[2] G. Bae, D.-I. Bae, M. Kang, S. M. Hwang, S. S. Kim, B. Seo, et al., "3nm GAA technology featuring Multi-Bridge-Channel FET for low power and high-performance applications," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2018, pp. 28.7.1-28.7.4.

[3] G. Yeap, S. S. Lin, H. L. Shang, H. C. Lin, Y. C. Peng, M. Wang, et al., "2nm Platform Technology Featuring Energy-Efficient Nanosheet Transistors and Interconnects Co-Optimized with 3DIC for AI, HPC and Mobile SoC Applications," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4.

[4] K. K. Bhuwalka, K. M. Choi, T. Okagaki, D. W. Kim, J. C. Kim, "Semiconductor Device", US Patent no. 11876097.

[5] J. Ryckaert, P. Schuddinck, P. Weckx, G. Bouche, B. Vincent, J. Smith, et al., "The Complementary FET (CFET) for CMOS scaling beyond N3," 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2018, pp. 141-142.

[6] K. K. Bhuwalka et al., "Optimization and Benchmarking FinFETs and GAA Nanosheet Architectures at 3-nm Technology Node: Impact of Unique Boosters," in IEEE Transactions on Electron Devices, vol. 69, no. 8, pp. 4088-4094, Aug. 2022.

[7] A. Agrawal, W. Chakraborty, W. Li, H. Ryu, B. Markman, S. H. Hoon, et al., "Silicon RibbonFET CMOS at 6nm Gate Length," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4.

[8] K. K. Bhuwalka et al., "Contacted-Poly-Pitch (CPP) Scaling Limitations of Si Gate-All-Around Nanosheet (GAA-NS) Devices," in IEEE Transactions on Electron Devices, vol. 72, no. 12, pp. 6437-6446, Dec. 2025, doi: 10.1109/TED.2025.3615917.

[9] S. Barraud et al., "7-Levels-Stacked Nanosheet GAA Transistors for High Performance Computing," 2020 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2020, pp. 1-2, doi: 10.1109/VLSITechnology18217.2020.9265025.

[10] Michelly de Souza, et. al., Analysis of electron mobility in 7-level stacked nanosheet GAA nMOSFETs, Solid-State Electronics, Volume 227, 2025, 109115, ISSN 0038-1101, https://doi.org/10.1016/j.sse.2025.109115.B.

[11] Vermeersch et al., "Self-Heating in iN8–iN2 CMOS Logic Cells: Thermal Impact of Architecture (FinFET, Nanosheet, Forksheet and CFET) and Scaling Boosters," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 371-372, doi: 10.1109/VLSITechnologyandCir46769.2022.9830228.

[12] A. Rawat et al., "Performance Trade-Off Scenarios for GAA Nanosheet FETs Considering Inner-spacers and Epi-induced Stress: Understanding & Mitigating Process Risks," ESSDERC 2021 - IEEE 51st European Solid-State Device Research Conference (ESSDERC), Grenoble, France, 2021, pp. 55-58, doi: 10.1109/ESSDERC53440.2021.9631815.