Contact Poly-Pitch Scaling of Si GAA-NSs down to A10 Node: Role of Nanosheet Thickness and Junction Profile

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

Lee-Chi Hung
Global TCAD Solutions
Vienna, Austria
r.hung@globaltcad.com

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

Franz Schanovsky
Global TCAD Solutions
Vienna, Austria
f.schanovsky@globaltcad.com

Zlatan Stanojevic
Global TCAD Solutions
Vienna, Austria
z.stanojevic@globaltcad.com

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

Wei Li
Huawei Technologies R&D
Leuven, Belgium
liwei@huawei.com

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

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

***Abstract*— Contact-poly-pitch (CPP) scalability of Gate-All-Around-nanosheet (GAA-NS) device is investigated in detail in combination with sheet thickness (T_{NS}) and junction abruptness (δ) both allowing for spacer-thickness (T_{SP}) and gate-length (L_G) scaling. Using an extensively calibrated advanced transport model combining with the DTCO flow we account for accurate device and circuit intrinsics and parasitics. We highlight trade-off between T_{SP} and L_G scaling (L_G scaling better than T_{SP}); T_{NS} can be scaled to ~3.5nm maintaining iso-performance allowing for a minimum CPP of ~38nm (L_G ~10nm). Aggressively improving δ can further scale CPP to ~34nm (L_G ~6nm), though T_{NS} scaling gain saturates due to severe mobility degradation.**

Keywords*— *Gate-all-around, Nanosheet, DTCO, scalability

I. Introduction

Gate-all-around nanosheet devices have been introduced at 3nm/2nm logic technology nodes [1, 2]. A key question arises in terms of scalability, particularly contact-poly-pitch (CPP) and challenges there-of. Multiple prior work has focused on metal-pitch (M_x) and standard-cell height (CH) scaling which has resulted in innovations like the fork-sheet (FS) devices [3] and (complementary) C-FETs [4]. Possibility of Si-sheet thickness scalability [5] for sub-threshold swing (S_{SAT}) improvement of up to 3nm has been investigated for a gate-length of 6nm [6]. Though the results are promising, a detailed performance-power analysis is missing to quantify how much gate-length (hence CPP) can be scaled. This is important, as, scaling thickness and gate-length not only results in mobility (transport) and sub-threshold swing (short-channel effects) but also capacitances changes. Further, n- and p-FETs behave differently with nanosheet thickness T_{NS} scaling [7, 8, 9]. Extrinsic parasitics like middle-of-line (MoL) capacitance C_{MOL} and external resistance R_{EXT} (described in Fig. 2) cause a dampening effect on intrinsic gains. Thus, at least a standard-cell level analysis is required for meaningful quantitative analysis.

In this paper, we discuss an in-depth analysis of thickness and junction gradient (δ) requirements for CPP scaling using a predictive physics-based DTCO flow based on quantum-mechanics and non-equilibrium transport [9]. For each n/p-FET device (each L_G, T_{SP}, T_{NS}) sub-band BTE simulations are done to calibrate semi-classical parameters. Semi-classical

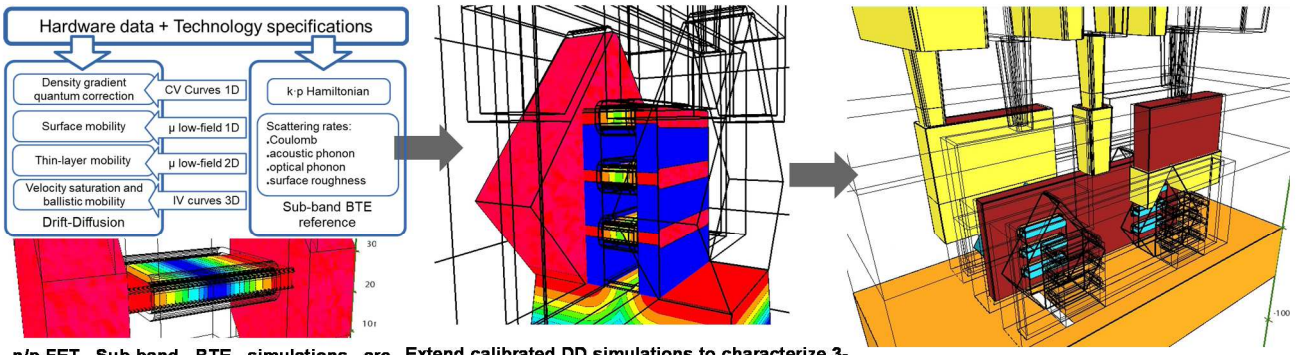

Fig. 1. The model parameters for the drift-diffusion (DD) semi-classical simulations have been obtained from calibration to the sub-band Boltzmann transport equation (SBTE) flow as developed in [9]. It uses a base calibration for both SBTE and DD from measured low-field mobility curves. The density gradient quantum correction, mobility models for surface, thin-layer, velocity saturation, and ballistic effects in the DD simulation are then refined to achieve a match with SBTE for a set of capacitance-voltage, low-field mobility, and current-voltage curves. This is repeated for *each* DoE for different geometry: T_{NS} (5nm to 3nm) and L_G (15nm to 6nm).

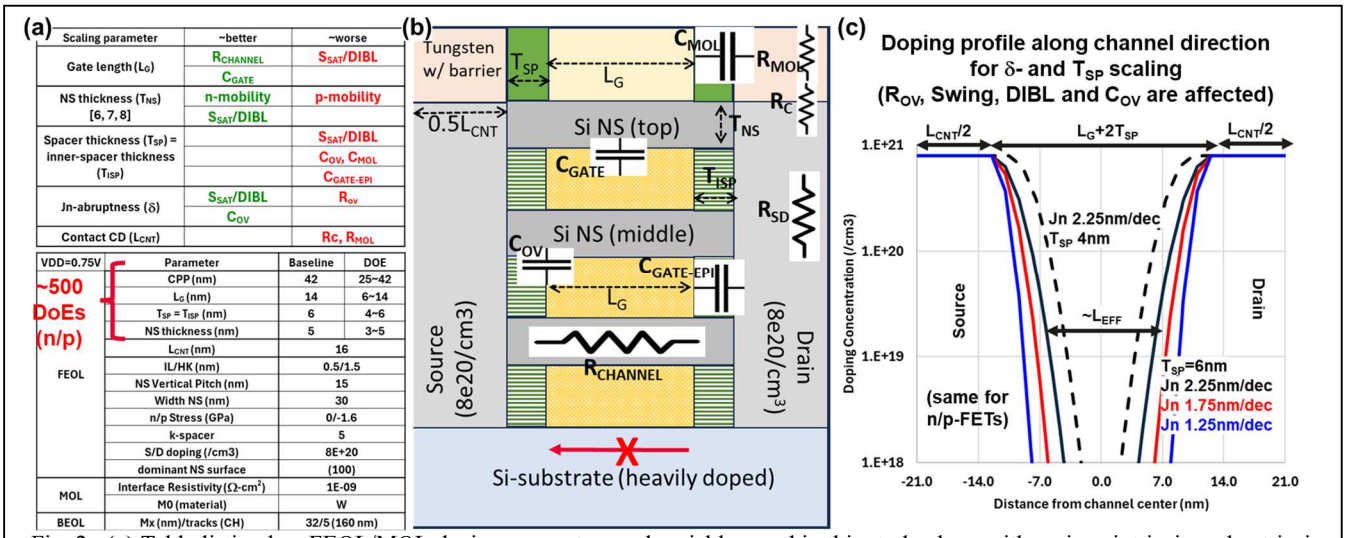

Fig. 2. **(a)** Table listing key FEOL/MOL device parameters and variables used in this study along with various intrinsic and extrinsic trade-off electrical parameters (S_{SAT}: sub-threshold swing, DIBL: drain-induced barrier lowering; R-: resistance; C-: capacitance; -ov: overlap; -MOL: middle-of-line, -GATE: gate-capacitance, -GATE-EPI: Gate-to-epitaxy (inner and outer)); **(b)** schematic cross-section of 3-stack NS indicating various parasitic components; **(c)** doping profile along the channel direction for δ and T_{SP} (also source-drain epitaxy proximity) scaling. Effective-gate-length, L_{EFF} (sensitive to T_{SP}, δ) is indicated; p-stress is assumed constant.

simulations are extended to a 3-stack n/p-nanosheet with a width W_{NS} of 30nm and a fixed vertical pitch of 15 nm. Thus, parasitics from device (source-drain resistance R_{SD}, interface resistance R_C, overlap capacitance, C_{OV}; gate-to-epi capacitance, $C_{GATE-EPI}$) and middle-of-line components (R_{MOL}, C_{MOL}) are all included.

II. TCAD/DTCO Methodology and Device Parameters

Fig. 1 describes the TCAD/DTCO flow used for device parameters calibration from physics-based to drift-diffusion and SPICE simulations [9,10]. It should be emphasized that for each n- and p-FETs design-of-experiment (DoE) parameter combination (T_{NS}, L_G, T_{SP}, δ) unique compact-model parameters are extracted, inverter standard-cells are defined for parasitic (PEX) extraction and SPICE simulations (5-stage inverter ring-oscillator with fan-out 3) are done. This is done to preserve the sanctity of the calculations down to T_{NS} of 3nm. n- and p-FETs have inner-spacers where thickness of inner-spacers is same as that of outer-spacers (T_{ISP}=T_{SP}) for all DoE. Its should be noted, that we have reported that relaxing this requirement can help improve C_{EFF} by aggressively scaling top-L_G in multi-stack GAA-NS [11], the same is not applied here for simplicity. Dominant NS surface is (100) for a standard wafer orientation of (100) and channel direction of <110>. Uniform and constant channel compressive stress is assumed in the p-FET (-1.6 GPa). Stress assumption is such that n- and p-FETs have nearly balanced performance at the baseline condition. CPP for each L_G, T_{SP} is adjusted such that contact-length, L_{CNT} is fixed at 16 nm corresponding to a N2 baseline CPP of 42 nm (L_G=14 nm, T_{SP}=6 nm, spacer-κ=5) at T_{NS}=5 nm (δ=2.25 nm.dec^{-1}). L_{CNT} is fixed as scaling it would always degrade performance due to R_C, R_{MOL} increase, unless MoL process/material optimization is considered, i.e. improving contact resistivity (ρ_{cnt}) as well as plug contact material allowing for barrier-less MoL. Also, its impact is nearly independent of T_{NS}, L_G, T_{SP}, δ. The semi-classical DTCO flow and baseline NS device and standard cell dimensions are discussed in detail in [12]. All the devices are renormalized (Vt-adjusted) for a fixed I_{OFF} of 1nA at V_{DD}=0.75V. Fig. 2 details the key parameter trade-off

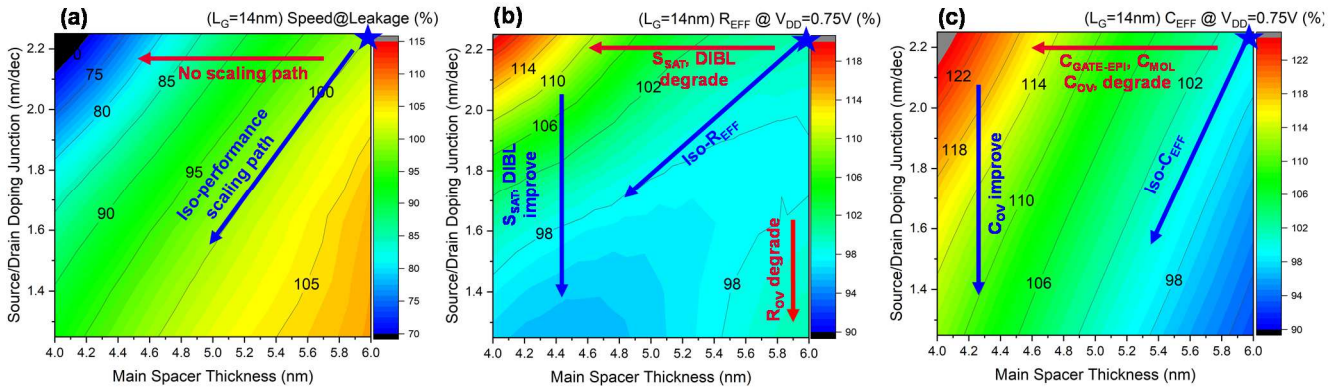

Fig. 3. (a) 5-stage Inverter (INV) with Fan-out-3 Speed-at-iso-leakage (each n/p I_{OFF}=1nA, V_{DD}=0.75V) [10] as a function of junction-gradient (δ) and spacer-thickness for 3-Stack NS devices (L_G=14nm, T_{NS}=5nm). Scaling T_{SP} (fixed δ) inherently results in both (b) R_{EFF} (S_{SAT}, DIBL) and (c) C_{EFF} ($C_{GATE-EPI}$, C_{OV}, C_{MOL}) degradation. Junction improvement by process engineering provides for iso-performance T_{SP} scaling (S_{SAT}, DIBL, C_{OV} improve), albeit by a few nanometers only. Each geometry has parasitic RC extracted from unique layout. R_{EFF}, C_{EFF} extracted from inverter delay, I_{DDQ}, I_{DDA} at V_{DD}=0.75V.

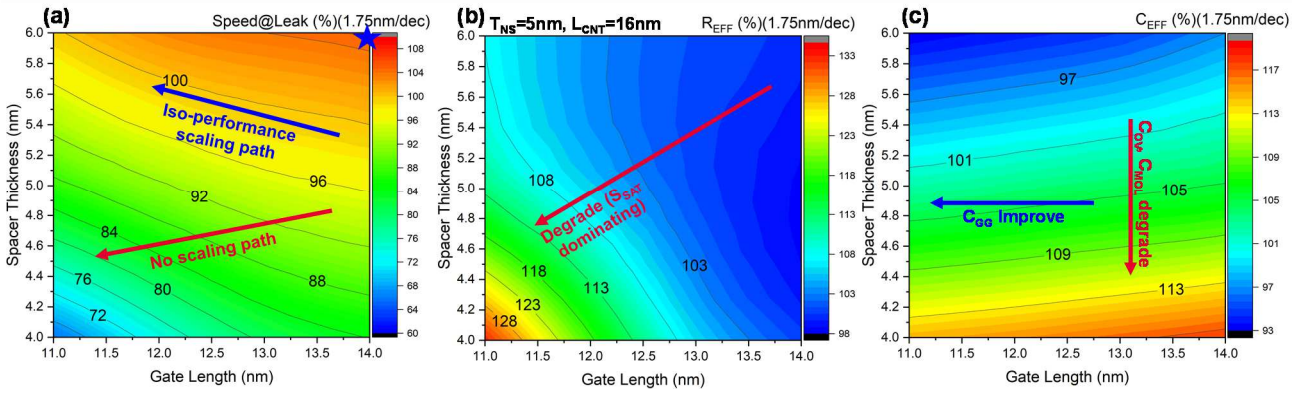

Fig. 4. Impact of gate-length and spacer-thickness scaling on performance for a T_{NS}=5nm, L_{CNT}=16nm (δ=1.75nm/dec). Performance degrades as L_G (S_{SAT}, DIBL) or T_{SP} (S_{SAT}, C_{MOL}, $C_{GATE-EPI}$, C_{OV}) scales. However, an iso-performance path exists with L_G scaling if T_{SP} can be increased to compensate for S_{SAT} degradation and C_{EFF} improvement **(a)** speed-at-iso-leakage; **(b)** R_{EFF}; **(c)** C_{EFF} (0.75V).

scenarios, DoEs considered as well as junction profile definitions.

III. Results and Discussion

Fig. 3a shows the detrimental impact of T_{SP} scaling on performance (speed-at-iso-leakage) due to severely degraded R_{EFF} (shown in Fig. 3b). This is essentially due to S_{SAT} degradation. As T_{SP} scales reduction in L_{EFF} (definition in Fig. 2c) leads to short-channel degradation even at constant L_G (see [12] for more details). Fig. 3c shows the corresponding C_{EFF} (which includes C_{GG}, $C_{GATE-EPI}$, C_{OV} and C_{MOL} components). As expected C_{EFF} degrades as T_{SP} scales. This is because C_{MOL} and C_{OV} degrade. C_{ov} degrade as junction overlap increases (Fig. 2c). R_{EFF}, C_{EFF}, device trade-off for T_{SP}, L_G, L_{CNT} scaling is discussed in detail in [12]. Improving junction gradient δ can recover performance as both S_{SAT} and C_{OV} improves. Better junction results in larger L_{EFF} which improves short-channel effects, while C_{OV} is improved due to overlap length reduction, for the same gate-length. Note that R_{EFF} sensitivity also improves as δ-improves. In Fig. 4, combined impact of L_G and T_{SP} is shown. Only scaling L_G or T_{SP} for fixed T_{NS} (5nm) and δ (1.75nm.dec^{-1}) degrades performance severely. While S_{SAT} and R_{EFF} degrade (4b), C_{GG} improves (4c). However, a marginal trade-off appears where it is possible to scale L_G if T_{SP} can be *increased* (or vice-versa). This is because higher L_{EFF} (Fig. 2c) results in improved S_{SAT}, C_{OV} and C_{MOL} (Fig. 4c). Further, degradation due to T_{SP} scaling is

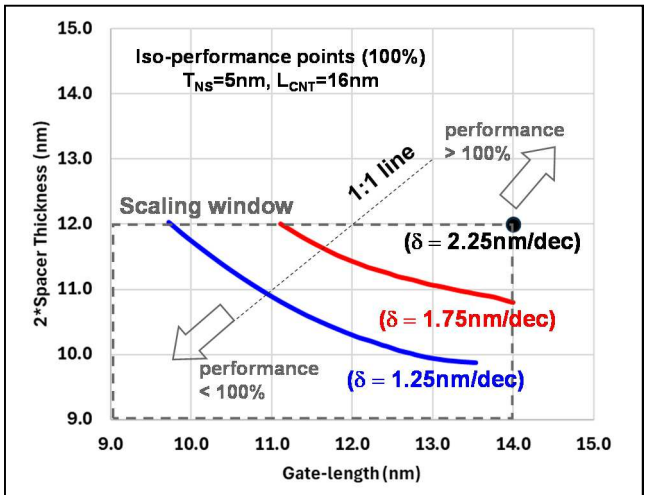

Fig. 5. T_{SP} and L_G requirements (T_{NS}=5nm) as a function of junction gradient δ, at iso-performance. L_G scaling provides better CPP scaling path compared to T_{SP} (>2nm L_G scaling per 1nm 2T_{SP} scaling) because of large increase in C_{EFF} (both C_{MOL} and $C_{GATE-EPI}$ degrade).

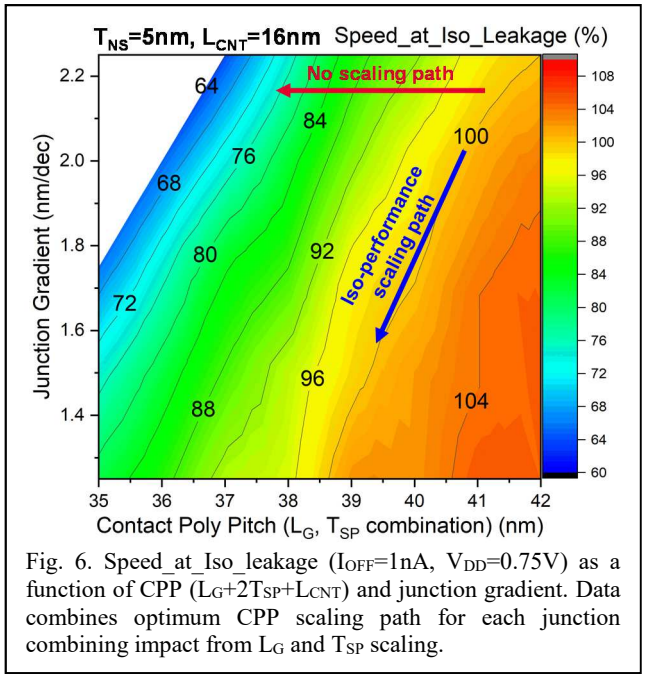

Fig. 6. Speed_at_Iso_leakage (I_{OFF}=1nA, V_{DD}=0.75V) as a function of CPP (L_G+2T_{SP}+L_{CNT}) and junction gradient. Data combines optimum CPP scaling path for each junction combining impact from L_G and T_{SP} scaling.

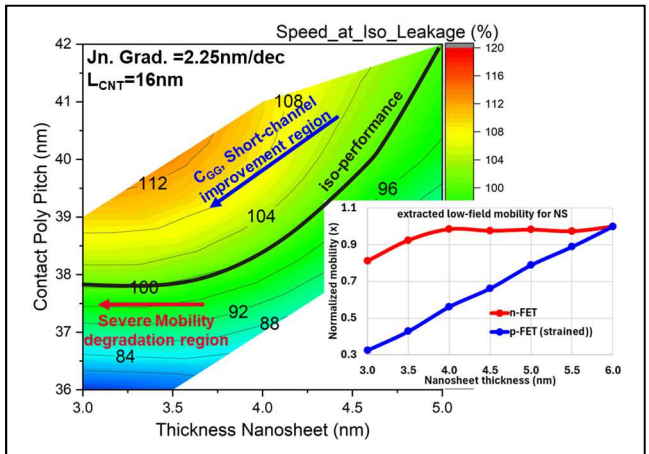

Fig. 7. Speed-at-iso-leakage as a function of CPP (L_G)and T_{NS}. T_{SP}=6nm as L_G allows maximum scaling advantage. As can be seen, T_{NS} scaling up-to ~3.5nm is feasible for minimum CPP of 37.5nm. Severe mobility (*inset*: extracted for n/p-NSs) degradation limits further CPP gain with T_{NS} scaling.

larger than L_G scaling. This is shown in Fig. 5 where iso-performance (100% at baseline) points are extracted for each δ-case. δ improvement allows for each L_G and T_{SP} scaling albeit scaling T_{SP} limits L_G scaling. As can be seen, for each 1 nm total T_{SP} thickness scaling (2 sides), L_G can be scaled by almost 3 nm maintaining iso-performance. Thus, for effective CPP scaling, it would be desirable to scale L_G keeping T_{SP} fixed.

Fig. 6 shows the combined effect of L_G, T_{SP} and δ on performance. At fixed CPP, without improvement in δ, severe degradation in performance results. Aggressively improving δ allows for CPP scaling by up to 5 nm at iso-performance. The exercise is repeated for T_{NS} scaling from 5 nm to 3 nm. Fig. 7 shows the impact of T_{NS} scaling on CPP scaling for a fixed δ of 2.25nm/dec. Interestingly, as can be seen, as T_{NS} scales from 5 nm to 4 nm, significant gain in CPP scaling is possible driven by S_{SAT} improvements as well as balancing of n- and p-FET mobility changes [7-9]. However, at T_{NS} scales below 3.5 nm, CPP scaling saturates as gains from S_{SAT}, C_{GG}, C_{MOL} and C_{OV} cannot be translated to performance due to both n/p-FET mobility degradation (inset Fig. 8), thus, putting a limit to T_{NS} scaling of ~3.5 nm. The results are summarized in Fig. 8 which shows minimum CPP as a function of δ and T_{NS} at L_{CNT} 16nm. At fixed T_{SP} (6nm), minimum CPP of ~34nm (L_G ~6nm) is possible with aggressive δ engineering and scaling T_{NS} to ~3.5nm, consistent with [6]. If δ-cannot be improved, minimum CPP is limited to ~38nm (L_G ~10nm).

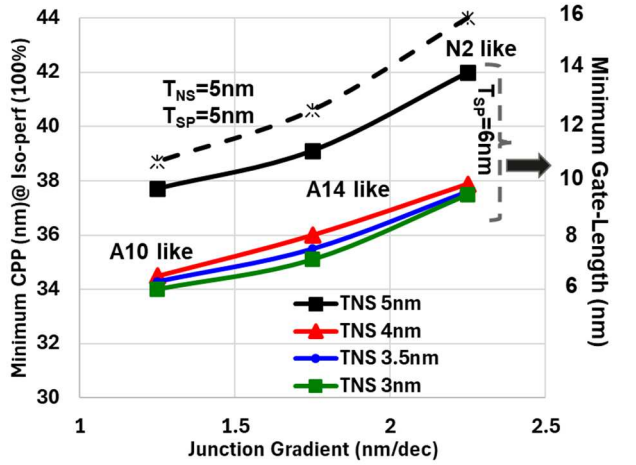

Fig. 8. Minimum CPP and L_G as a function of scaling T_{NS} and Junction-gradient for fixed T_{SP}=6nm (as L_G scales better than T_{SP}). T_{NS} 5nm allows CPP scaling to 38nm if junction can be improved; T_{NS} 3.5nm allows CPP 38~34nm; while little gain with further T_{NS} scaling (due to severe mobility degradation).

CONCLUSIONS

We have investigated intrinsic (n/p mobility, S_{SAT}, C_{GG}) and extrinsic (R_{EXT}, C_{MOL}, C_{OV}) trade-off scenarios for CPP (L_G, T_{SP}) scaling by junction and T_{NS} scaling, using a quantum mechanics based DTCO flow. Scaling trade-off between L_G, T_{SP} exists, L_G outperforming T_{SP} due to C_{MOL} contribution. Si-T_{NS} can be scaled up to ~3.5 nm maintaining iso-performance beyond which both n/p-FET mobility degradation negates C_{GG}, S_{SAT} gains dampened by MoL parasitic. Thus, a minimum CPP of up to 34nm is possible for a NS based devices (GAA NSs/C-FETs) leading towards A10 node. Further T_{NS} scaling would require mobility improvements by surface-roughness improvement, strain engineering as well as channel orientation optimization.

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] S. Mochizuki, N. Loubet, P. Mirdha, C. Durfee, H. Zhou, G. Tsusui, et. al., "Evaluation of (110) versus (001) Channel Orientation for Improved nFET/pFET Device Performance Trade-Off in Gate-All-Around Nanosheet Technology," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023, pp. 1-4.

[6] 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.

[7] K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata and S. Takagi, "Experimental study on carrier transport mechanism in ultrathin-body SOI n- and p-MOSFETs with SOI thickness less than 5 nm," Digest. International Electron Devices Meeting, San Francisco, CA, USA, 2002, pp. 47-50.

[8] R. Kaur and N. R. Mohapatra, "Exploring Sheet Thickness Scaling and Substrate Orientation for Maximizing Nanosheet pFET Performance," in IEEE Transactions on Electron Devices, vol. 71, no. 9, pp. 5754-5760, Sept. 2024.

[9] L.-C. Hung, G. Rzepa, M. Kampl, C.-M. Tsai, et. al., "Hierarchical Transport Modeling for Path-Finding DTCO," 2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), San Jose, CA, USA, 2024, pp. 1-4.

[10] Z. Stanojević et al., "Nano Device Simulator—A Practical Subband-BTE Solver for Path-Finding and DTCO," in IEEE Transactions on Electron Devices, vol. 68, no. 11, pp. 5400-5406, Nov. 2021.

[11] K. K. Bhuwalka et al., "Design Decoupling of Inner-and Outer-Gate Lengths in Nanosheet FETs for Ultimate Scaling," in IEEE Journal of the Electron Devices Society, doi: 10.1109/JEDS.2024.3492722.

[12] 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.