2nm Platform Technology featuring Energy-efficient Nanosheet Transistors and Interconnects co-optimized with 3DIC for AI, HPC and Mobile SoC Applications

Geoffrey Yeap, S.S. Lin, H.L. Shang, H.C. Lin, Y.C. Peng, M. Wang, PW Wang, CP Lin, KF Yu, WY Lee, HK Chen, DW Lin, BR Yang, CC Yeh, CT Chan, JM Kuo, C-M Liu, TH Chiu, MC Wen, T.L. Lee, CY Chang, R. Chen, P-H Huang, C.S. Hou, YK Lin, FK Yang, J. Wang, S. Fung, Ryan Chen, C.H. Lee, TL Lee, W. Chang, DY Lee, CY Ting, T. Chang, HC Huang, HJ Lin, C. Tseng, CW Chang, KB Huang, YC Lu, C-H Chen, C.O. Chui, KW Chen, MH Tsai, CC Chen, N. Wu, HT Chiang, XM Chen, SH Sun, JT Tzeng, K. Wang, YC Peng, HJ Liao, T. Chen, YK Cheng, J. Chang, K. Hsieh, A. Cheng, G. Liu, A. Chen, HT Lin, KC Chiang, CW Tsai, H. Wang, W. Sheu, J. Yeh, YM Chen, CK Lin, J. Wu, M. Cao, LS Juang, F. Lai, Y. Ku, S.M. Jang, L.C. Lu

Global R&D Center, Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu, Taiwan, R.O.C.

Abstract

A leading edge 2nm CMOS platform technology (N2) has been developed and engineered for energy-efficient compute in AI, mobile and HPC applications. This industry-leading N2 logic technology features energy-efficient gate-all-around nanosheet transistors, middle-of-line and backend-of-line interconnects with densest SRAM macro of ~38Mb/mm^2. N2 delivers a full node benefit from previous 3nm node [4] in offering 15% speed gain or 30% power reduction with >1.15x chip density increase. N2 platform technology, equipped with new Cu scalable RDL, flat passivation and TSVs, co-optimizes holistically with 3DFabric™ technology enabling system integration/scaling for AI/mobile/HPC product designs. N2 successfully met wafer-level reliability requirements and passed 1000hrs HTOL qual with high yielding 256Mb HC/HD SRAM, and logic test chip (>3B gates) consisting of CPU/GPU/ SoC blocks. Currently in risk production, N2 platform technology is scheduled for mass production in 2H'25. N2P, 5% speed enhanced version of N2 with full GDS compatibility, targets to complete qualification in 2025 and mass production in 2026.

Introduction

Advanced CMOS technology has been the key enabler for semiconductor product innovations. Since the generative AI break-through moment in Q1'23, AI together with 5G-advanced mobile and HPC have ignited the industry with an insatiable appetite for best-in-class advanced energy-efficient logic technology [1]. Our industry leading 2nm platform technology (N2) is one such advanced logic technology. This paper describes the state-of-art N2 technology successful transition into NS platform technology and acceleration of >140x energy-efficient compute from 28nm to N2 as shown in Fig. 1. We also present system technology co-optimization (STCO) innovation in design rules, standard cell, SRAM and interconnects co-optimization with 3DFabric™. N2 technology has been verified on our development/qual test vehicle. N2 met all the wafer-level reliability requirements and completed the full 1000hours HTOL qualification with high yielding 256Mb HD/HC SRAM and logic test chip (>3B gates). Now in risk production, N2 is on track for mass production in 2H'25. N2P with 5% additional speed and full GDS compatibility targets to complete qual in 2025 and mass production in 2026.

N2 NanoFlex™ [3] Technology Architecture

The N2 2nm platform technology is defined and developed to meet PPACt (Power, Performance, Area, Cost, and Time-to-market) [2]. STCO is emphasized with smart scaling features instead of brute-force design rule scaling which drastically increases process cost and inadvertently causes critical yield issues. Extensive STCO coupled with smart scaling of major design rules (e.g., gate, nanosheet, MoL, Cu RDL, passivation, TSVs) was performed in optimizing this 2nm technology to achieve target PPA. This development also involves co-optimization with 3DFabric™ SoIC 3D-stacking and advanced packaging technology (INFO/CoWoS variants) thereby accelerating system integration/scaling for AI/mobile/HPC product designs.

N2 NanoFlex™ [3] standard cell innovation offers not only nanosheet width modulation but also the much-desired design flexibility of the multi-cell architecture. N2 short cell lib for area and power efficiency. Selective use of tall cell lib lifts frequency to meet design target Combining with six-Vt offerings spanning across 200mV, N2 provides unprecedented design flexibility to satisfy a wide spectrum of energy-efficient compute applications at the best logic density. N2 delivers a full node scaling with attractive PPA values at projected cost and time-to-market: ~15% speed gain or ~30% power reduction with >1.15x chip density scaling (Fig. 2-3).

Energy-efficient Nanosheet Transistors, MoL and BEOL Interconnects

Multiple generations of Si FinFet with fin depopulation were in use from 16nm to 7nm (2-fin) node. High mobility channel transistors with industry-first zero-thickness dipole based true multi-Vt (7-Vt), cut metal-gate and gate-contact over-active innovations extended FinFet architecture into N5 node [2]. FinFlex™ DTCO coupled with other key enhancements successfully extracted another full node PPA benefits in N3, last FinFet node [4].

N2 platform technology successfully completes the transition from FinFet into energy-efficient nanosheet technology. Figure 4 shows optimized nominal gate-length NS transistors with excellent DIBL and subs-threshold swings. Long gate-length NS transistors achieve near-ideal 60.1mV/dec swings. Fig. 5 shows the six-Vt's ranging from the extreme low-Vt to standard-Vt in ~200mV span for N2 N/P FETs. Si data is very close to matching ring speed@standby-power at all six Vts. This multi-Vt capability is enabled with 3rd-generation (since N5) dipole-based multi-Vt integration with both n-type and p-type dipoles.

Much process and device enhancements are focused on engineering not just the transistor drive currents through sheet interface/thickness, junction engineering, dopant diffusion/activation and stress engineering, but more on Ceff reduction to drive best-in-class energy efficiency. All these enhancements lead to much improved I/CV speed gain of 70% and 110% respectively for NS N/P FETs. N2 nanosheet technology exhibits substantially better Perf/Watt than FinFET at low Vdd range of 0.5V-0.6V (Fig. 7). Emphasis is placed on low Vdd perf/watt uplift through process and device continuous improvements resulting in 20% speed gain and 75% lower stand-by power at 0.5V operation. N2 NanoFlex coupled with multi-Vt provides unprecedented design flexibility to satisfy a wide spectrum of energy-efficient compute applications at the most competitive logic density.

Overall technology energy efficiency and performance are also critically dependent on MoL, backend and far-backend interconnects. With innovative materials and processing, VG Rc reduces significantly by 55% with barrier-less all-tungsten MoL. The low resistance MoL combined with capacitance reduction features achieve a total of ~6.2% INV D4 ring oscillator speed gain (Figure 8). Optimized M1 with novel 1P1E EUV patterning led to close to 10% std cell capacitance reduction and a saving of multiple EUV masks. Substantial My RC and Vy Rc reductions seen on the tightest 193i 1P1E workhouse metal/via layers (Fig. 12). In summary, N2 MoL and BEOL RC reduce by ~>20% contributing significantly to energy-efficient compute.

Seamless Integration with 3DFabric Tech

This 2nm platform technology, including the new Cu RDL with flat passivation and TSVs, co-optimizes holistically with 3DIC enabling system integration/scaling for AI/mobile/HPC product designs (Fig. 11-12). Attention is paid to optimize materials and processing in backend/far-backend for global warpage and local planarity for robust integration with 3D stacking. N2 also optimizes pTSV/sTSV (for power/signal) in terms CD/pitch/density for F2F/F2B stacking with SoIC bond pitch scaling from 9µm/6µm down to 4.5µm.

SRAM, Logic Test Chip and Qual/Reliability

For advanced nodes, SRAM bitcell scaling has become a challenge. With N2 NanoFlex and improved on-off current, DTCO is employed to maximize #bitcell/bitline, bitline loading and SRAM peripheral layout efficiency resulting in densest 2nm SRAM macro density ~38Mbmm^2 (Fig. 13). N2 HC/HD pull-down Nfet with better Vt-sigma than FinFet resulting in ~20mV lower HC Vmin and 30~35mV lower HD Vmin (Fig. 14). HD 256Mb SRAM shmoo plot in Fig. 15 illustrates full read and write down to ~0.4V. With innovative well engineering and junction isolation, N2 has better latch-up trigger voltage for both logic and SRAM than FinFet (Fig. 15). Higher Vtrig in N2 leads to additional logic density and more effective DVS screening for product quality. N2 test chip demonstrates healthy CPU/GPU functionality and passed the GPU Vmin-power spec shown in Fig. 16.

N2 256Mb HC/HDSRAM consistently demonstrated healthy defect density resulting in >80% / >90% avg/peak yields (w/o repairs). Fig. 20 shows 256Mb SRAM passed 1000hour HTOL qualification with ~110mV margin.

Additional HPC features such as super high performance MiM (SHP-MiM) with ~>200fF/mm2 capacitance density is offered for higher Fmax by minimizing transient drooping voltage. High-speed SerDes test chip also demonstrated fully functioning 14Gb/s LPDDR6 and 10Gb/s HBM3E interfaces.

References:
[1] Y.J. Mii, IEDM, plenary, 2024.
[2] G. Yeap et al., IEDM, 2019.

[3] "TSMC 30th North America Tech Sym. with Innovations Powering AI with Silicon Leadership," TSMC PR, 2024/04/24
[4] S.-Y. Wu et al., IEDM, 2022.

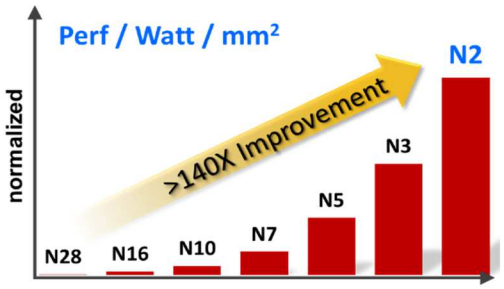

Fig.1 N2 NanoFlex™ accelerates energy-efficient compute

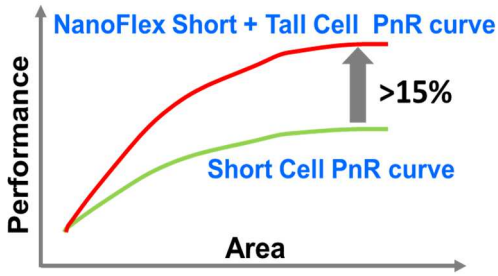

Fig.2 N2 NanoFlex innovation modulates NS width for best PPA. Short cell library for area and power efficiency. Selective use of tall cell library lifts frequency to design target

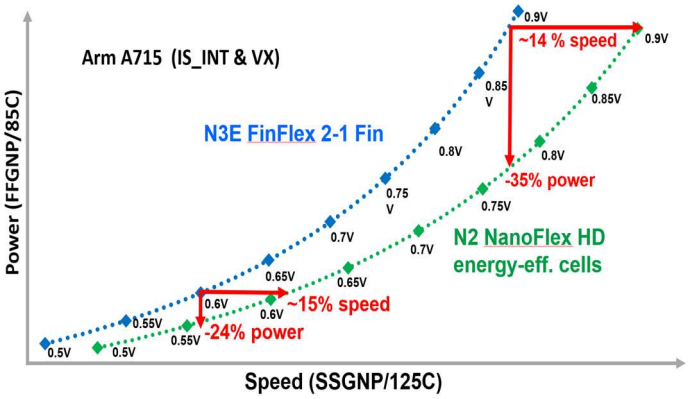

Fig.3 N2 NanoFlex HD cells gain 14~15% speed@power vs. N3E FinFlex 2-1 fin cell across Vdd range: 35% power saving at higher voltage and 24% power saving at lower voltage

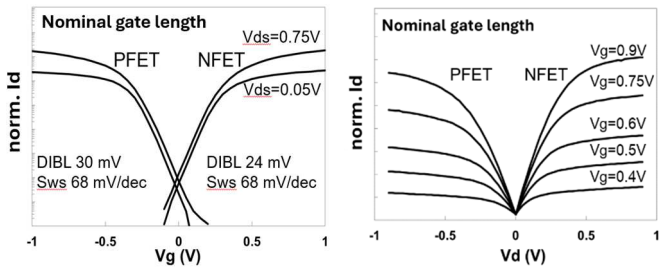

Fig.4 N2 gate & drain characteristics w/ excellent DIBL/Sws

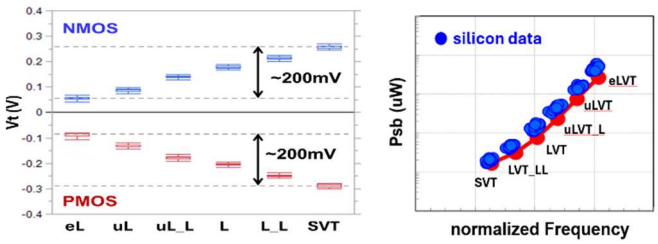

Fig.5 Six-Vt's with ~200mV range for low leakage and high-perf. optimization. Si data closes to matching ring speed @standby-power

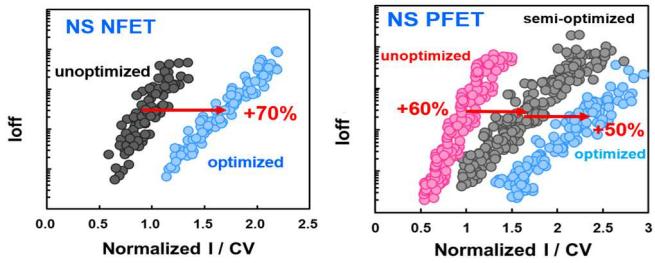

Fig.6 Not only drive current/mobility enhancement, more so on Ceff reduction: N/P +70% and +110% gain in I/CV speed

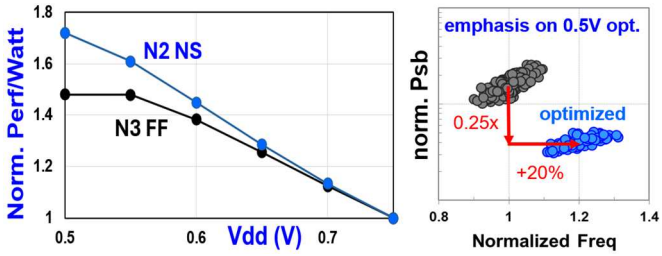

Fig.7 NS vs FF: better Perf./Watt at 0.5V~0.6V. Sp. emphasis at low-Vdd: +20% speed and 75% lower std-by power

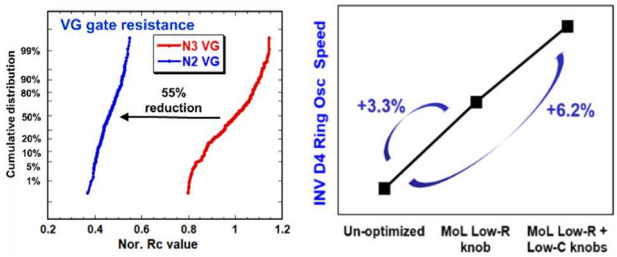

Fig.8 MoL low R (VG, VD and MD) and lower Ceff optimization leading to 6.2% speed gain

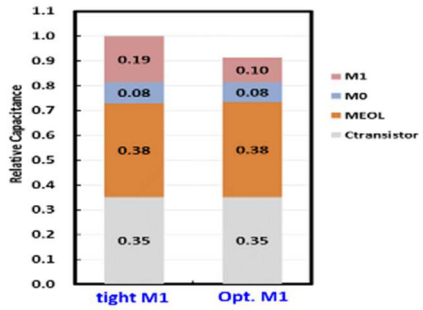

Fig.9 Optimized M1 with novel 1P1E EUV patterning leads to 9% Ceff reduction and a saving of multiple EUV masks

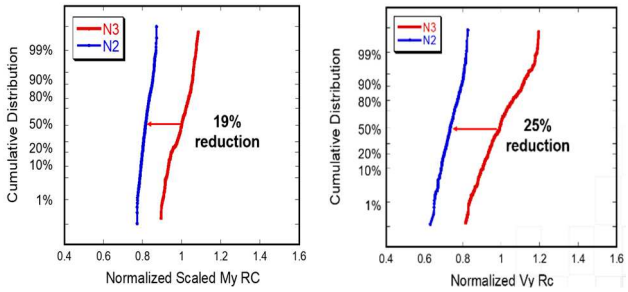

Fig.10 Significant My RC and Vy Rc reduction on N2 tightest 193i 1P1E workhouse metals/vias

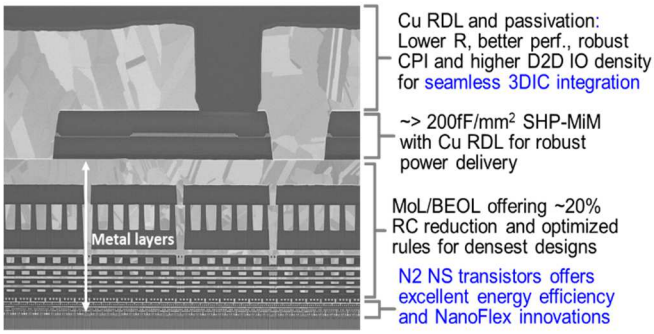

Fig.11 N2 new Cu RDL and passivation provide seamless integration with 3DFabric™ (SoIC, INFO and CoWoS) tech

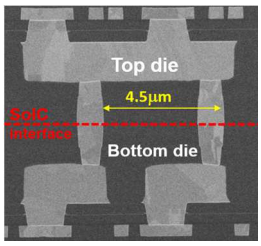

Fig.12 N2 optimizes pTSV and sTSV for F2F/F2B stacking w/ SoIC bond pitch →4.5 µm

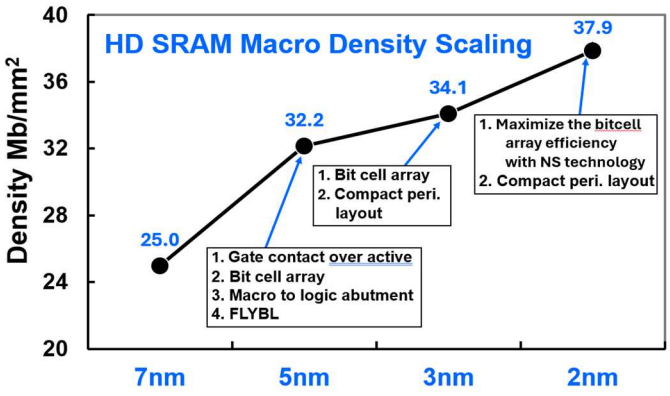

Fig.13 N2 offers highest SRAM macro density ~38Mb/mm²

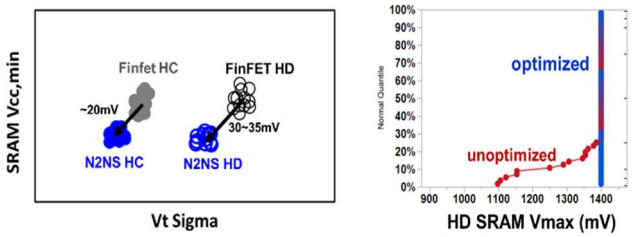

Fig.14 N2NS with better DIBL and Vt-sigma leading to lower Vmin for more energy-efficient compute. HD Vmax>1.4V

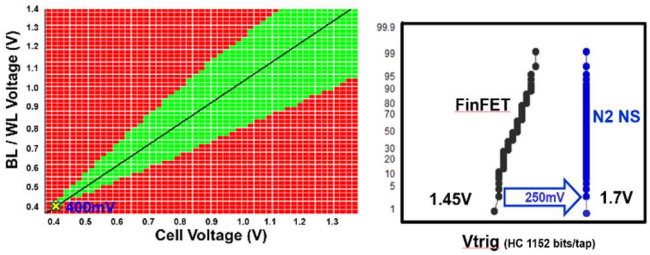

Fig.15 256Mb HD SRAM shmoo to 0.4V. N2 >1.7V Vtrig: higher logic density and effective DVS for product quality

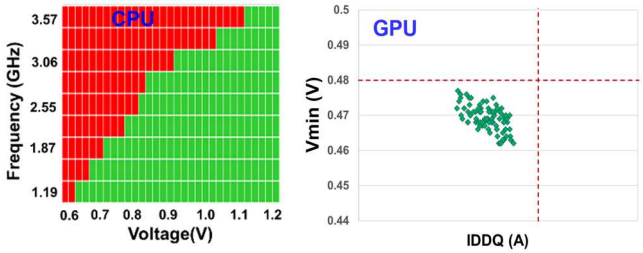

Fig.16 Shmoo plots of CPU block and GPU Vmin vs. IDDQ in the high yielding logic test chip in N2 qual vehicle

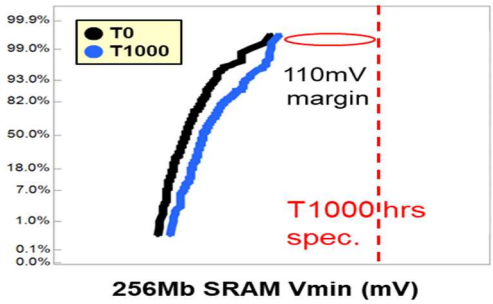

Fig.17 N2 technology met all wafer-level reliability requirements and passed 1000hrs HTOL spec

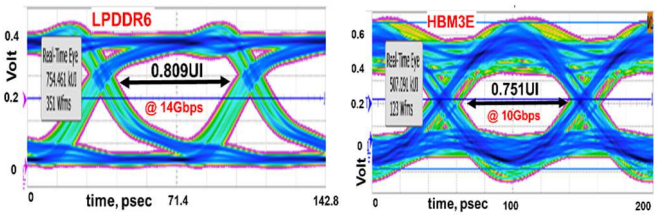

Fig.18 High-speed SerDes test chip in N2 vehicle showing fully functioning LPDDR6 @14Gb/s and HBM3E @10Gb/s