A 3nm CMOS FinFlex™ Platform Technology with Enhanced Power Efficiency and Performance for Mobile SoC and High Performance Computing Applications

Shien-Yang Wu, C.H. Chang, M.C. Chiang, C.Y. Lin, J.J. Liaw, J.Y. Cheng, J.Y. Yeh, H.F. Chen, S.Y. Chang, K.T. Lai, M.S. Liang, K.H. Pan, J.H. Chen, V.S. Chang, T.C. Luo, X. Wang, Y.S. Mor, C.I. Lin, S.H. Wang, M.Y. Hsieh, C.Y. Chen, B.F. Wu, C.J. Lin, C.S. Liang, C.P. Tsao, C.T. Li, C.H. Chen, C.H. Hsieh, H.H. Liu, P.N. Chen, C.C. Chen, R. Chen, Y.C. Yeo, C.O. Chui, W. Chang, T.L. Lee, K.B. Huang, H.J. Lin, K.W. Chen, M.H. Tsai, K.S. Chen, X.M. Chen, Y.K. Cheng, C.H. Wang, W. Shue, Y. Ku, S. M. Jang, M. Cao, L.C. Lu, T.S. Chang

Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan, R.O.C., email: shien-yang_wu@tsmc.com

2022 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-8959-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IEDM45625.2022.10019498

***Abstract* —** The industry fastest time-to-manufacturability 3nm CMOS platform technology is presented. FinFlex™ with standard cells consisting of different fin configurations is introduced for the first time to offer the critical design flexibility for better power efficiency and performance optimization compared to traditional FinFET technologies. An aggressive scaling of ~1.6X logic density increase, 18% speed improvement and 34% power reduction are achieved over our previous 5nm CMOS process. This FinFlex™ platform technology provides the best-in-class PPAC values to fully unleash product innovations in 5G and HPC applications.

I. Introduction

The proliferation of artificial intelligence applications and 5G deployment in recent years have been the driving forces for high performance computing at data center as well as low-power networking and processing capability at edge devices [1]. As machine learning is rapidly adopted in wide spectrum of industries that require big data processing with speed and accuracy, HPC is picking up the momentum as the next critical growth driver. Advanced CMOS logic technology with highest performance and best power efficiency is more important than ever to unleash innovations that would shape every aspect of our daily life and society.

This paper presents the state-of-art 3nm platform technology with on-target device performance as well as scaling innovation from standard cell design and critical ground rules. In addition to successfully extend bulk FinFET into 3nm node, FinFlex™ standard cell innovation offers the much-desired design flexibility of the multi-cell architecture. Combining with 6 Vt offerings spanning across 200mV, this technology provides an unprecedented design flexibility to satisfy a wide spectrum of power efficient SoC requirement and high-performing demand of HPC applications at the most competitive logic density. This process has been verified on our development test vehicle consisting of high-density and high-current SRAM macros and logic test chip.

II. Design Flexibility – FinFlex™ and Multiple Vt

FinFlex™, an innovative standard cell architecture with different fin configurations, is introduced for the first time in this 3nm technology. Accompanied with traditional pitch scaling at critical layers, it achieves the logic density increase of a full node. To further reduce FinFET footprint, fin pitch scaling and fin depopulation are the typical approaches adopted by the industries. With fin pitch already at below 30nm and fin number reduction to single fin, process variation and insufficient device drivability become the major obstacles for further scaling. FinFlex™ offers several configurations as in Fig.1 to address this trade-off between scaling and performance. The 2-1 fin configuration achieves area reduction without sacrificing performance for power sensitive applications. The 2-fin device within can be used in the critical path to leverage its higher current while the single fin be used for leakage reduction. The result is the highest density standard cell to date with the lowest power consumption. Similarly, the 3-2 fin configuration, equipped with 3-fin for higher drive current, is well suited for performance demanding applications. The conventional 2-2 fin configuration can be applied where good balance between performance, power and density is desired. Contrary to the simple fin cut in a regular standard cell with only transistor level capacitance reduction, FinFlex™ offers cell level area scaling as well as chip level capacitance reduction through co-optimization of BEOL place-and-route. Moreover, with 6 different Vt offerings in this technology, designers can choose various combinations of fin number and Vt for individual N/PMOS to satisfy wide range of speed and leakage requirement on the same chip. Fig. 2 shows ARM Cortex-72 CPU performance and area improvement of this 3nm FinFlex™ technology compared to our 5nm node [2]. The power-efficient 2-1 cell demonstrates 30% power reduction with 11% speed gain at 0.64X area; high-performing 3-2 configuration 33% speed gain with 12% power reduction at 0.85X area; and the balanced 2-2 cell 23% speed gain with 22% power reduction at 0.72X area. This innovation is one of the critical components to successfully extend the lifetime of FinFET architecture for another full technology node.

III. Process Architecture

In addition to the novel standard cell feature, critical ground rule scaling is adopted to achieve ~1.6X logic density improvement than previous 5nm node. Careful fin width and profile optimization across various fin arrangements maintains the required short channel effect at reduced gate length. Low-K spacer is implemented to reduce parasitic capacitance between contact and gate without affecting yield and reliability. Raised

source/drain with dual epitaxy process is optimized to provide channel strain and to reduce source/drain (S/D) resistance. The 6th generation high-K metal gate (HK/MG) RMG process supports core and I/O devices. Novel contact schemes and process solutions in middle-end-of-line reduces the parasitic resistance for the tight CPP scaling while maintaining healthy yield and reliability. Advanced Cu/low-k interconnect scheme with aggressively scaled minimum metal pitch process is also developed. Innovative barrier and liner engineering as well as patterning optimization have kept BEOL metal and via RC on track without impacting chip performance due to scaling.

IV. Transistor Performance

The 2-1 fin configuration of this 3nm technology offers 18% iso-power speed gain or 34% power reduction at the same speed over our 5nm technology based on Figure of Merits (FOM) consists of Inverter, NAND, and NOR circuitry with a fan-out of 3 (F.O. = 3) as shown in Fig.3. Fin width and profile are carefully optimized to obtain ~50mV/V DIBL at the target scaled Lg (Fig.4), proving that FinFET is still a viable architecture at 3nm node [3, 4]. FOM performance as well as NMOS and PMOS devices have achieved the targeted performance for this technology as illustrated in Fig.5 and Fig.6, respectively. To fully realize the promised benefit of FinFlex™, it is crucial to eliminate fin number difference induced loading effect that may degrade the intrinsic fin performance. Single fin device is especially vulnerable because many process steps, such as etch and epitaxy, naturally deviate from those experienced by multi-fin structures. Fig.7 shows that after process optimization, the single fin device in the 2-1 fin configuration performs as-design with ~50% active power of its 2-fin counterparts. For high-speed applications, 3-2 fin configuration has more than 9% speed increase as observed in Fig.8. Six different Vt options with >200mV range (Fig.9) are ready for selection to further provide the design flexibility for power-performance trade-off. Since device variation has become increasingly important in budgeting the design margin, process improvement specifically to combat variation is also implemented to reduce device Vt mismatch (AVt) by 20% for both NMOS and PMOS in Fig. 10. For I/O device, LDD implant optimization in Fig. 11 reduces Iboff by more than 2 orders based on the required fin profile for SCE control.

V. Interconnect Technology

Interconnect process has played an ever more important role in determining the overall chip performance. For this 3nm technology, minimum metal pitch at 23nm is used to enable the scaling of FinFlex™ 2-1 fin configuration while providing the needed routing efficiency. To our knowledge, this is the tightest metal pitch reported so far in the advanced nodes. Innovative liner for Cu is incorporated to reduce the RC of this minimum pitch by 20% for nominal metal width and 30% for structure with 2X metal width shown in Fig. 12. Significant Via Rc reduction by ~60% based on an innovative barrier process in Fig. 13 is an essential component to enable such aggressive pitch scaling. By examining metal resistance of line A vs. line B for M0 and Mx layers, process robustness is demonstrated by the comparable distribution between line A and line B in Fig.14. Barrier thickness reduction at upper relaxed metal pitches as well as ELK dielectrics are deployed to minimize overall BEOL RC delay. Fig.15 shows the cross-section view of 15-level Cu/low-k metal stacks. The tight Rc distributions of stacked contact to via chain for 6-level and 15-level metals prove the stability of this package. The reliability of BEOL process integration is also examined. Excellent EM performance of Vx/Mx and Vx/Mx+1 for the minimum pitch metal, and interconnect SM stability are verified in Figure 16(a) and 16(b), respectively. The resistance shift percentage of Kelvin Rc structures with on-rule and wide metal is negligible after stressed for 500 hours. Furthermore, three critical layers requiring EUV double patterning in previous generation are replaced by single EUV patterning, which reduces process complexity, intrinsic cost and cycle time.

VI. Yield and Reliability

Competitive HD and HC SRAM cells are offered for low leakage and high-performance applications. A yield learning vehicle consisting of both HD and HC 6-T SRAM 256Mb macro as well as logic test chip with CPU/GPU/SoC blocks is used for technology development. The butterfly curves of the $0.021um^2$ HD SRAM cell are shown in Fig.17 where cell stability down to 0.3V is demonstrated. The static noise margin (SNM) of 97mV and 124mV are achieved for 0.45V and 0.6V operation, respectively. The Shmoo plot of the 256Mb HD SRAM macro in Fig. 18 illustrates full read and write capability down to 0.5V. The 256Mb HC/HD SRAM macros and product-like logic test chip have consistently demonstrated healthier defect density than our previous generations at the same development phase. Furthermore, both 256Mb HC/HD SRAM macros passed HTOL 1000hrs qualification with margin as in Fig.19, and the logic test chip passed the Vmin-power specification as illustrated in Fig.20 for the CPU.

VII. Conclusion

We introduce the industry leading 3nm FinFlex™ CMOS foundry technology equipped with innovative design flexibility and wide range of Vt options. With this new DTCO feature, product designs with different functional blocks optimized for performance, power, and/or area targets can be integrated on the same chip if needed. Together with critical ground rule scaling and an aggressive minimum metal pitch at 23nm, this technology offers the best-in-class logic performance, power efficiency and low Vmin SRAM at the highest density to date. With device performance achieving design target and process induced variation properly addressed, the demanding requirements of high performing HPC applications as well as power sensitive SoC products can all be well satisfied. Technology maturity and readiness for high volume production of various 5G mobile and AI/HPC applications is well demonstrated with stable yield and robust manufacturability and is proven with rigorous technology qualification.

REFERENCES

[1] S.-Y. Wu, IEDM Tech. Dig., p.36.3, 2019.
[2] G. Yeap et al., IEDM Tech. Dig., p.36.7, 2019.
[3] S.-Y. Wu et al., IEDM Tech. Dig., p.2.6, 2016.
[4] S.-Y. Wu et al., IEDM Tech. Dig., p.9.1, 2013.

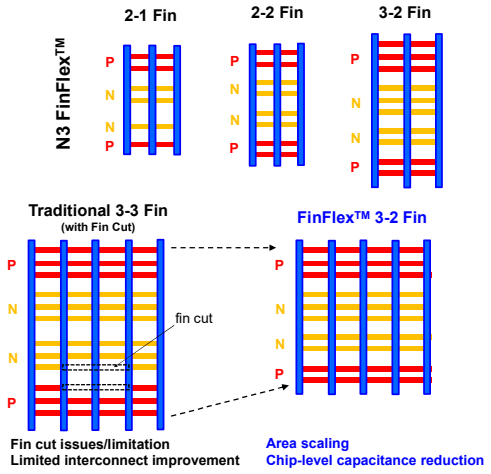


Fig. 1. FinFlex™ schematics and comparison to traditional fin depopulation. Area reduction and significant chip level capacitance reduction are the major benefits of this innovation over traditional FinFET designs.

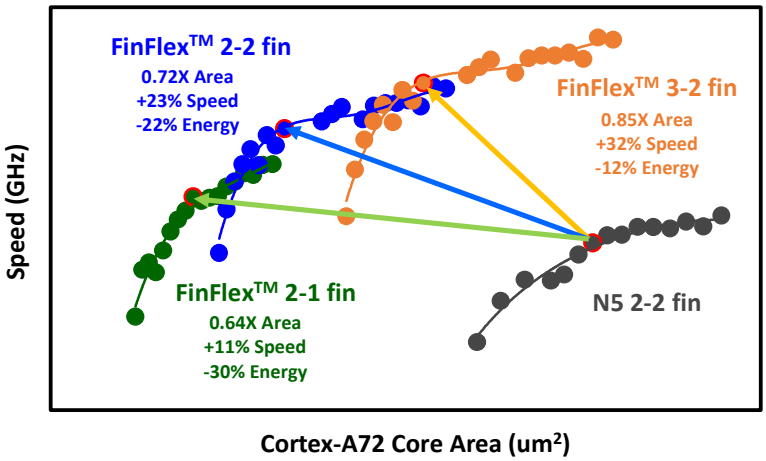


Fig. 2. FinFlex™ improvement in ARM Cortex-A72. FinFlex™ 2-1 fin configuration targets ultra power efficiency, 2-2 fin efficient power and 3-2 fin ultra high performance. Different area, speed and power efficiency improvements from our N5 technology are shown for each configuration.

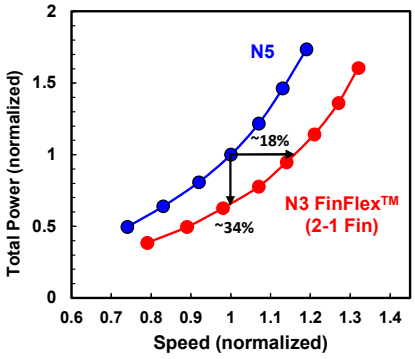


Fig. 3. FinFlex™ 2-1 cell provides 18% SPD gain @ fixed power or 34% power reduction @ fixed speed

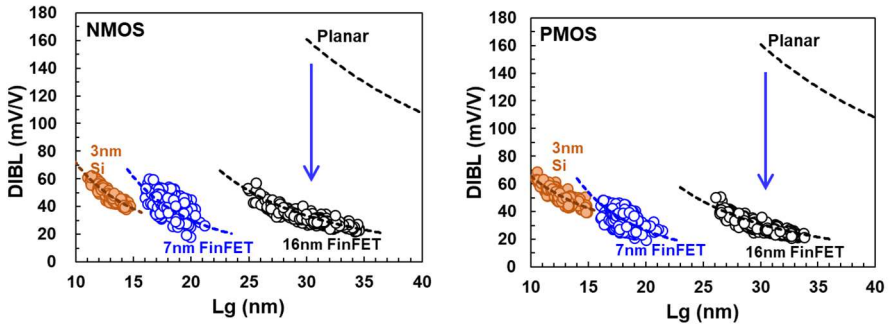


Fig. 4. FinFET SCE improvement continue to support the required Lg scaling for 3nm technology.

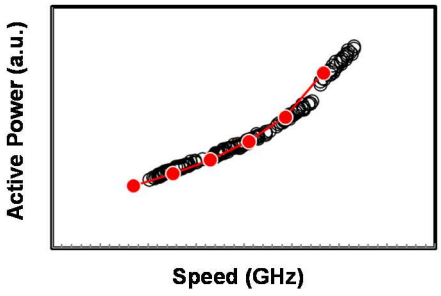


Fig. 5. Figure-of-merit (FOM) structure achieves on-target power-speed performance for all Vts.

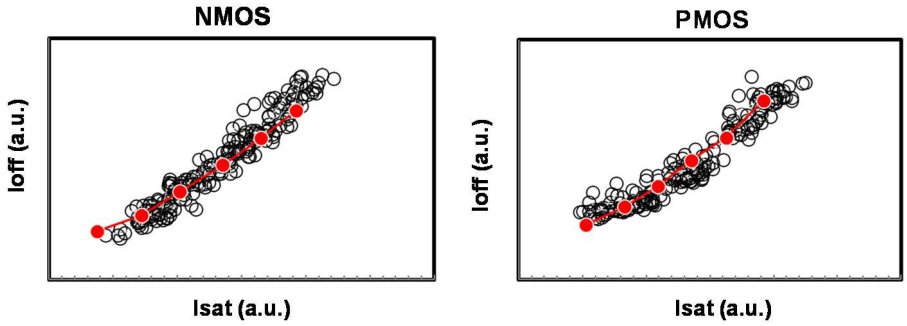


Fig. 6. Both NMOS and PMOS devices demonstrate on-target performances.

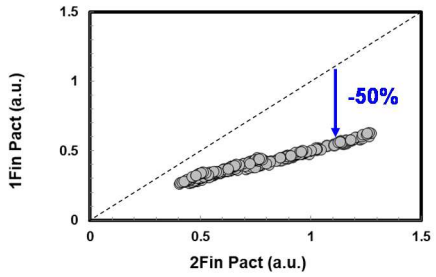


Fig. 7. 1-fin device demonstrates 50% active power reduction w/o process loading induced degradation.

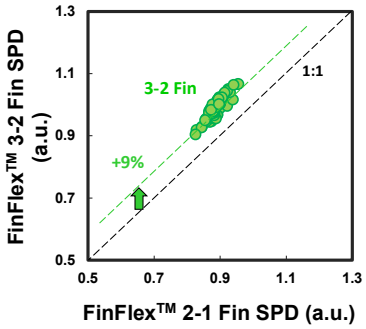


Fig. 8. FinFlex™ 3-2 fin has additional 9% SPD gain.

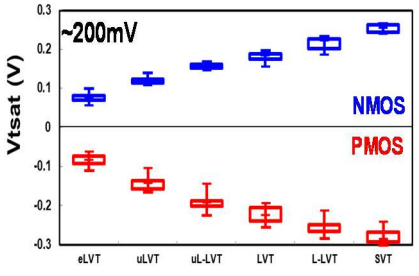


Fig.9 Six different Vt options with ~200mV span.

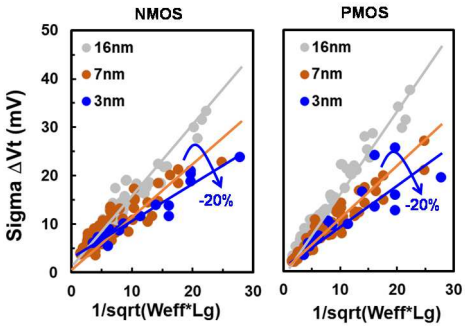


Fig.10 Superior mismatch performance demonstrated.

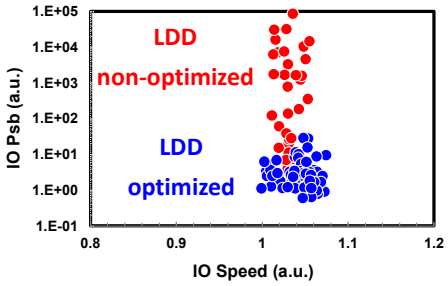


Fig.11 I/O device Psb vs. speed. With LDD optimization, Iboff is significantly reduced.

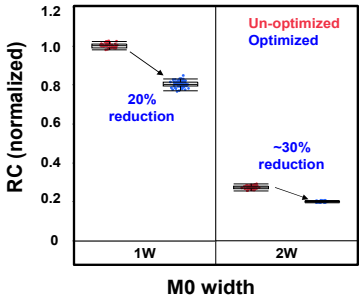


Fig.12 Pitch 23nm metal line RC increase is contained by innovative Cu liner process.

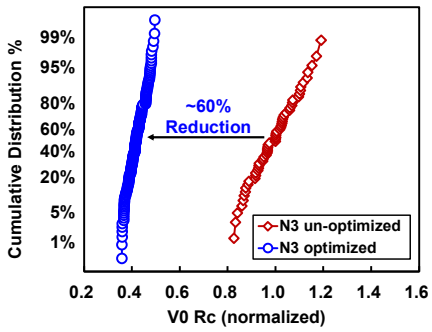


Fig.13 Significant via Rc reduction at the tightest pitch by an innovative barrier process.

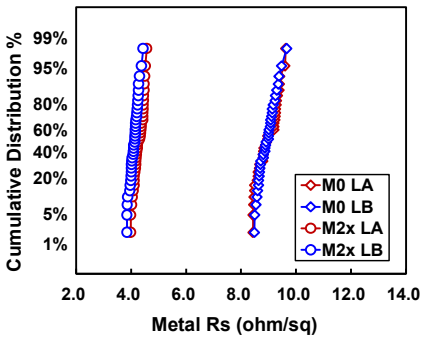


Fig.14 M0/Mx metal resistance distributions of line-A and line-B at aggressively scaled pitch.

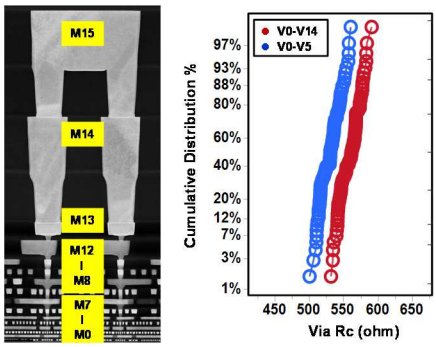
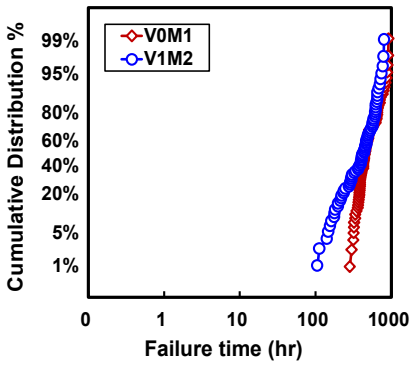


Fig.15 TEM image of 15-level of metal stacks and tight distribution of the Via Rc stack.



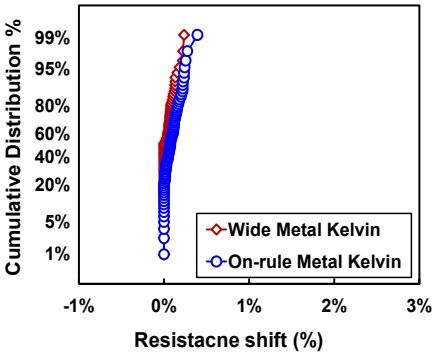


Fig.16 (a) EM performance of the minimum pitch metal; (b) SM of kelvin structures.

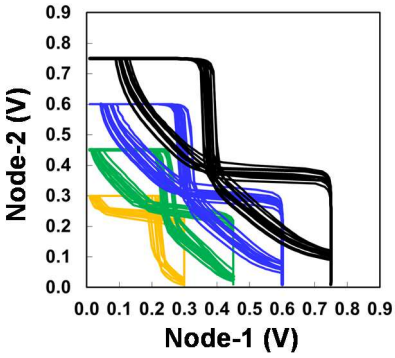


Fig. 17. SNM of a 0.021um^2 high density 6-T SRAM cell.

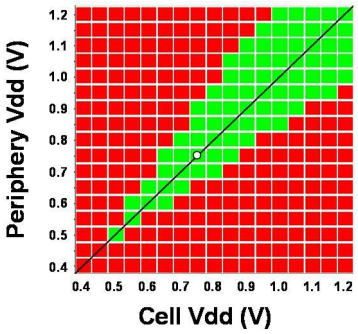


Fig. 18. Schmoo of a 0.021um^2 HD 256Mb SRAM macro with full read/write function down to 0.5V.

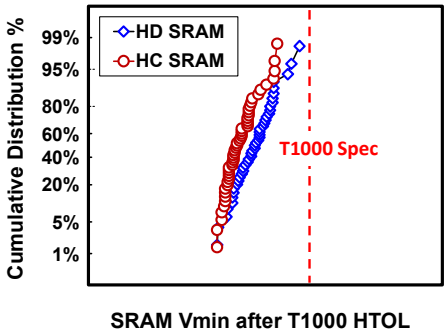


Fig. 19. Both HC/HD 256Mb SRAM pass HTOL 1000hrs spec.

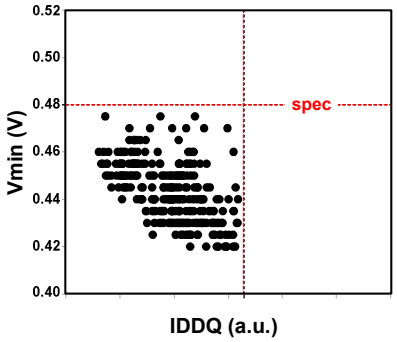


Fig. 20. Vmin vs. IDDQ of the CPU block in the logic test chip.