2025 IEEE 68th International Midwest Symposium on Circuits and Systems (MWSCAS)
August 10-13, 2025. Lansing, Michigan, USA

Design and Optimization of On-Chip Interconnects for Cryogenic Operation

Ali H. Hassan[1], Nathan Lang[1], Puneet Gupta[1], Sudhakar Pamarti[1], and Chih-Kong Ken Yang[1]
[1]Department of Electrical and Computer Engineering, University of California, Los Angeles, USA
Email: {aehassan@ucla.edu, njlang@g.ucla.edu, puneetg@ucla.edu, spamarti@ee.ucla.edu, and yang@ee.ucla.edu}

2025 IEEE 68th International Midwest Symposium on Circuits and Systems (MWSCAS) | 979-8-3315-8934-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/MWSCAS53549.2025.11244410

***Abstract*—Digital processing implemented in a technology optimized for reduced temperatures, such as 77K, has significantly improved power dissipation. Operating at cryogenic temperatures benefits both active devices and the properties of passive elements, such as interconnects. With over 30% gate delay and power dependent on interconnect, this paper focuses on optimizing on-chip metal interconnects based on a design technology co-optimization (DTCO) methodology for cryogenic environments. By refining the metal stack architecture, we achieve substantial improvements, yielding power savings of >10% for synthesized logic and >80% for repeaters at 77K. The results are validated in 14-nm-class FinFET technology.**

Keywords—Cryogenic Computing, Low Temperature, Inverter-Based Repeater, BEOL Interconnects, RC delay, DTCO.

I. Introduction

The doubling of transistor density every 18 months over the past half-century, as predicted by Moore's law, has driven the rise of high-performance computing (HPC). This progress has enabled breakthroughs in areas such as artificial intelligence, deep learning, cloud computing, and real-time photorealistic graphics for gaming, audio/video signal processing, and scientific visualization. However, the rapid scaling of CMOS technology, which powered this revolution, has significantly slowed. One of the most concerning consequences is that the scaling of transistor performance and power has stagnated since the 10-nm node [1]. Additionally, the impact of interconnects has worsened, with parasitic effects even more dominant as technology nodes shrink. Cryogenic computing is emerging as a potential solution to overcome these challenges, offering new opportunities for power and performance [1-2].

Performance enhancements at 77K have been widely investigated across various aspects of computing, including interconnect optimization [3], core performance improvements [4], and cryogenic-based memory architectures [5]. Back-End-of-Line (BEOL) interconnects remain a significant performance bottleneck in modern processors, especially as technology scales and parasitic effects worsen [6]. At cryogenic temperatures, while transistor performance improves due to reduced thermal noise and enhanced carrier mobility, interconnect resistance and capacitance behavior also change, necessitating novel design approaches to optimize overall system efficiency. This work presents an interconnect optimization strategy that redefines the metal-stack architecture through metal thinning, addressing Back-End-of-Line (BEOL) bottlenecks and enhancing energy efficiency at 77K. The proposed approach reduces interconnect capacitance while maintaining resistivity comparable to room temperature, ensuring lower power consumption without degrading signal integrity. By strategically thinning metal layers, this method optimizes interconnect performance, balancing capacitance reduction and resistance management to achieve significant energy savings in cryogenic computing environments. The optimization process begins with stack refinement for buffer-based interconnects, followed by adjustments in stack characteristics within a place-and-route (PnR) design flow to evaluate the resulting impact on a processor core. By integrating these modifications into a 77K-optimized design framework, this methodology demonstrates the potential for significant energy savings and improved metal utilization, offering a promising direction for cryogenic computing and future high-performance applications.

The rest of the paper is organized as follows. The metal-stack optimization flow is discussed in Section II. Circuit analysis and implementation are presented in Section III. Finally, a conclusion is drawn in Section IV.

II. Proposed Metal-Stack Optimization

Fig. 1 shows a standard technology BEOL metal stack, including representative layer thicknesses. The temperature-dependent resistivity coefficients of different metal layers vary based on their material composition, fabrication process, and geometry. Fig. 2 portrays resistance measurements for this technology, where these variations lead to a 2X reduction in resistance for lower-level metal layers and up to a 5X reduction for higher-level metal layers, primarily due to differences in electron-scattering mechanisms [7].

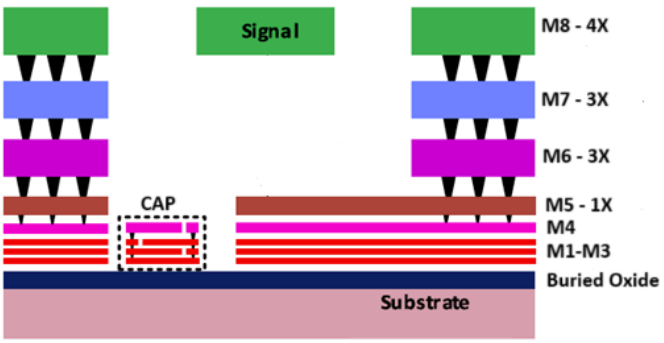

Fig. 1. *Cross section of Metal Back End of Line (BEOL) Stack*

At 77K, metal resistivity continues to scale almost linearly with temperature, making liquid nitrogen cooling a practical solution for high-performance computing applications that require energy-efficient operation [7]. The significant reduction in resistivity at cryogenic temperatures provides an opportunity for further energy savings by thinning wire height, which reduces capacitance while maintaining manageable resistance levels. Although lowering interconnect resistance improves RC delay, our findings demonstrate that capacitance reduction has a far greater impact on energy efficiency, making it the primary optimization target in cryogenic environments.

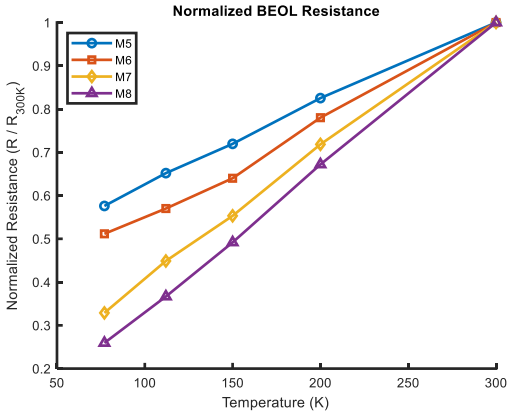

Fig. 2. Measured Normalized Resistance Variations Versus Temperature

III. Circuit Analysis and Implementation

The impact of wire thinning on various circuit scenarios is systematically analyzed and compared between low-temperature (LT) and room-temperature (RT) environments using an LT-optimized technology. The analysis begins with the design of a 31-stage ring oscillator featuring an enable signal, which serves as a representative testbench to emulate the logic depth of a large digital core with minimal interconnect distances. This configuration illustrates a baseline for evaluating logic performance. The circuit schematic of the implemented ring oscillator is shown in Fig. 3(a), while its physical layout is depicted in Fig. 3(b).

The findings are summarized in Table I, highlighting three key performance comparisons: (1) the baseline performance of the circuit schematic at RT versus LT, (2) the impact of parasitic capacitance, with modest wire loading, on performance at both RT and LT, and (3) the performance improvements achieved through a 30% reduction in metal thickness at LT, simulating the effects of lower metal interconnects.

A critical aspect of the LT-optimized technology is the use of transistors with reduced supply voltage (V_{dd}) and threshold voltages (V_{th}), ensuring that the oscillation frequency is constant, and the leakage power remains a constant proportion of active power relative to RT conditions [8-9]. The simulations show a 10X power improvement with or without some additional wiring capacitance added to the circuits. The power ratio improves to 13.3X when the capacitance is reduced by 30%. Note that a lower-V_{dd} is used to equalize the frequency with the reduced capacitance. The energy benefit is from the combination of the lowered capacitance and the reduced supply to compensate for the improved delay and equalize the frequency.

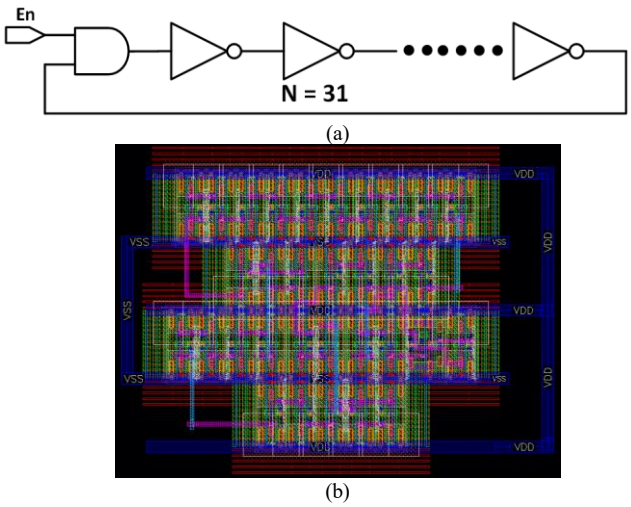

Fig. 3. A 31-Stage Ring Oscillator Circuit (a) Schematic and (b) Layout

TABLE I.
Ring Oscillator Performance Summary

Cell Setup	Vdd (V)	Freq (GHz)	Pavg (µW)	Poff (µW)	P – Ratio
RO @ RT	0.8	3.96	227.1	0.82	Ref
RO @ RT*	0.8	2.65	228.65	0.82	Ref
RO @ LT	0.25	3.94	21.3	0.1	10.66X
RO @ LT*	0.25	2.63	23.05	0.1	9.9X
RO @ LT**	0.23	2.59	17.14	0.085	13.34X

*Cell 1 – Model a cap of 1.2 fF at each node for 10 µm trace
**Cell 2 – Model a 30% cap reduction (0.84 fF)

Interconnect power savings are most evident in repeaters or buffers that buffers the digital signals over longer routes to maintain signal integrity and improve delay. Data movement can lead to substantial power dissipation due to the interconnect capacitance. Using LT and metal stack thinning can greatly improve the power cost of interconnect buffering and shows the upper bound of the amount power savings. Table II shows a power comparison of optimized LT in comparison with RT. The table optimizes the buffer insertion and sizing for a long 4mm interconnect on a higher metal layer with lower resistivity. We applied 50% thinning to maintain the same resistance between LT and RT. The set of simulations represented in the table minimizes energy-delay of the repeater chain in driving the total distance. The driver size (Drv Size) is shown as a multiplier on a minimum-sized buffer. N is the number of segments. The top row of the table shows the optimization result of buffer insertion in the RT condition. The second row performs the same optimization in LT without any interconnect thinning. Note that with the LT device threshold is previously defined to match the gate delay between LT and RT. In this optimization, V_{dd} is optimized to the nominal LT supply of 0.25V. Because of the reduced resistance of the LT interconnect, both the driver size and the number of segments is reduced. The reduced supply voltage enabled the design to have a substantial energy savings of 10.5X. The third row shows the delay improvement when the interconnect metal is thinned by 50% and the same parameters as the first row are used except under LT operating conditions. There is a substantial improvement in delay from the reduced

capacitance. In the fourth row, the same optimization as the second row is performed for N, DrvSize, and V_{dd}. The full potential of interconnect thinning is highlighted in this result of 57X energy reduction.

TABLE II. REPEATER DESIGN PERFORMANCE SUMMARY – LT OPTIMIZED TECHNOLOGY

Cell Setup	Vdd (V)	Delay (ps)	Nseg	Engr (fJ)	DrvSize	E-Ratio
Repeat-RT	0.8	641.32	6	427.84	36	Ref
Repeat-LT*	0.25	637.56	5	40.6	28	10.53X
Repeat-LT**	0.25	377.22	6	23.13	36	18.5X
Repeat-LT***	0.15	641.7	5	7.44	38	57.5X

*No thinning of interconnect with optimized Nseg and DrvSize for constant delay
**With thinning of interconnect with same Nseg and DrvSize as RT
***With thinning of interconnect with optimized Vdd, Nseg and DrvSize

This thinning of interconnect can be evaluated for a place and route (PnR) synthesis flow. Fig. 4 illustrates the design flow for optimizing a standard metal stack, showcasing the steps involved in achieving energy-efficient interconnect performance in an LT environment. This paper uses this process to evaluate the benefit of interconnect thinning,

The process begins with the foundry's standard metal stack configuration, where an encoded interconnect technology (ICT) file is modified to implement metal thinning and include the impact of this thinning on the resistance. The results shown in the two prior exemplar circuits show that any performance impact of increased resistance is far overshadowed by the reduced energy of lower capacitance. The modified ICT file is processed using a commercially available electronic design automation (EDA) tool, Cadence Techgen. This tool generates an updated capacitance table file (QRC file) that can be directly used in a PnR flow, enabling accurate modeling of interconnect behavior in the LT environment. While this flow is limited by often conservative estimates of capacitance in a QRC file, it is compatible with LT-optimized device models that are incorporated in the process design kit (PDK).

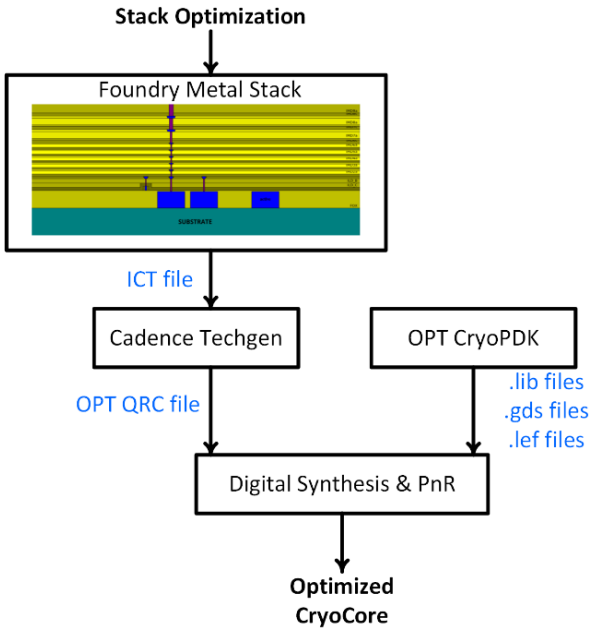

Fig. 4. *PnR Stack Optimization Flow*

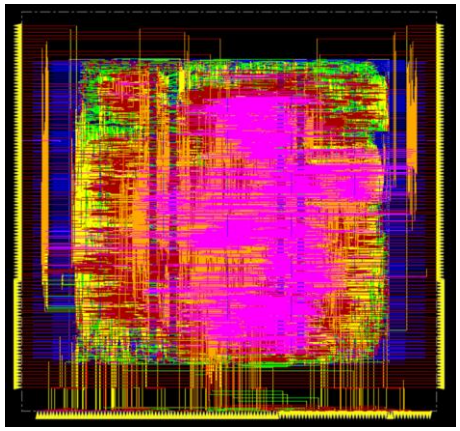

Fig. 5. Cortex-M3 Layout Implementation in an LT-Environment

To evaluate the impact of interconnect thinning on a digital processor design, the result of applying this methodology on a Cortex-M3 processor core is depicted in Fig. 5 with 20,793 gates and 200 µm x 190 µm area. The caches are not included in this evaluation. This implementation using LT-specific standard cells without interconnect thinning shows approximately 10X reduced power in comparison to an RT design of the same target cycle time primarily due to the V_{dd} of 0.25V.

Fig. 6 illustrates the extracted wire lengths from the implementation after PnR, highlighting the utilization patterns of the metal layers in the routing process. The analysis reveals that, as expected, the lowest metal layer, M1, is minimally utilized, with only 1% of its capacity engaged in routing. Similarly, the higher metal layers, such as M7 and M8, are used sparingly within the implemented core. Hence thinning and the use of upper metal layers do not have a substantial impact on the power. Meanwhile, maintaining the thickness of M7 and M8 and not including them for logic routing can benefit other aspects of design such as improved clock routing, and power and ground loss, or cost saving. The intermediate metal layers (M2-M6) demonstrate significant utilization, with varying ratios ranging from 20% to 24% and hence are important for thinning.

The improvements in switching power consumption from metal thinning across all layers, including M1, are illustrated in Fig. 7. Based on place-and-route (PnR) power reports, a 15.4% reduction in switching power is observed compared to the LT baseline without thinning, validating the effectiveness of the proposed interconnect optimization. Since switching power constitutes roughly half of the total power budget, this translates to a 7.8% reduction in overall power consumption. Moreover, as internal cell power also depends on local routing—primarily in M1, which is included in the thinning process—an additional 3.8% power reduction is achieved through decreased parasitic capacitance within standard cell layouts. Collectively, these contributions lead to a total core power improvement of 11.6%, underscoring the impact of uniform metal thinning on both global interconnect and local routing domains in LT environments.

Fig. 7 presents the percentage reduction in capacitance (top) and switching power (bottom) for metal layers M1 to M8, along with the overall impact on the total interconnect. The observed

capacitance reduction is achieved through metal thinning, which directly contributes to a corresponding decrease in switching power. It is important to note that the results in Fig. 7 only account for the reduction in interconnect capacitance. As shown previously for the oscillator and repeater designs, thinning also leads to an improved delay. For the Cortex-M3 processor, timing analysis reveals that the minimum cycle time along the critical path is reduced by 13% as a result of the proposed metal-stack optimization. This performance gain creates a valuable timing margin that can be exploited to reduce the supply voltage (V_{dd}) while preserving the original target cycle time. Our evaluation confirms that a 13% reduction in V_{dd} maintains equivalent timing, indicating an approximately 1:1 sensitivity between delay and supply voltage in this LT environment. When combined with the reduction in interconnect capacitance from metal thinning, this voltage scaling contributes to a cumulative power savings of 26%, demonstrating the effectiveness of coordinated electrical and physical-level optimizations for energy-efficient digital design at 77K.

It is important to note that in an LT environment, along with power dissipation, the signal current is reduced. While the thinning of interconnect metals may raise reliability concerns, the reduced temperature and current mitigate much of that concern. The primary challenge lies in the manufacturability constraints associated with extreme metal thinning. However, reducing the BEOL metal thickness by 30-50% is not expected to pose significant fabrication challenges, making it a feasible approach for improving interconnect efficiency in LT applications.

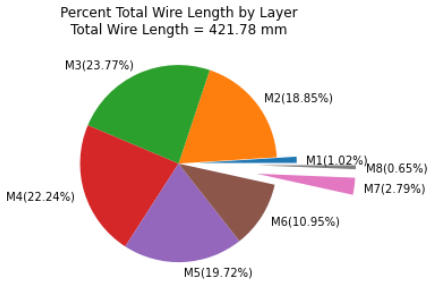

Fig. 6. *Metal Utilization for Cortex − M3 Core Layout*

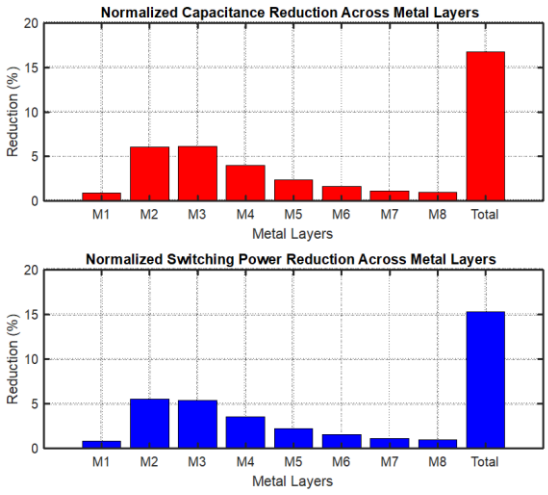

Fig. 7. *Capacitance and Switching Power Reduction Across Metal Layers*

IV. Conclusion

Cryogenic computing presents a promising and energy-efficient alternative to conventional technologies, especially for high-performance applications. Device optimization for a 14-nm FinFET technology in low-temperature environments has been shown to dramatically enhance performance by permitting the use of substantially reduced V_{dd} and threshold. These adjustments have been shown to improve the energy by more than 10X. This work further shows that the passive interconnects can be optimized by thinning the metals to lead to further energy or power reduction. In comparison to non-thinned interconnects, the proposed approach can improve power performance in a repeater structure by 57X, which can be considered an upper bound of the achievable improvement. In a more generalized logic, such as a ring oscillator or a PnR of a processor core, the improvement is 10- 26X. This work underscores the critical role of interconnect power optimization for LT operation and the overall potential of cryogenic computing to deliver superior energy efficiency and performance to address the limitations of conventional CMOS scaling at advanced technology nodes.

Acknowledgment

This work is funded by DARPA and ARL through the Low Temperature Logic Technology (LTLT) Project under Award Number W911NF2220019. Any opinions, findings, or conclusions expressed in this material are those of the authors and do not necessarily state or reflect those of the United States Government or any agency thereof.

References

[1] I. Byun, D. Min, G. Lee, S. Na, and J. Kim, “A next-generation cryogenic processor architecture,” IEEE Micro, vol. 41, no. 3, pp. 80–86, 2021.

[2] H.-L. Chiang, J.-J. Wu, C.-H. Chou, Y.-F. Hsiao, Y.-C. Chen, L. Liu, J.-F. Wang, T.-C. Chen, P.-J. Liao, J. Cai, X. Bao, A. Cheng, and M.-F. Chang, “Design technology co-optimization for cold cmos benefits in advanced technologies,” in 2021 IEEE International Electron Devices Meeting (IEDM), 2021, pp. 13.2.1–13.2.4.

[3] D. Prasad, et al., “Cryo-computing for infrastructure applications: A technology-to-microarchitecture co-optimization study,” in 2022 International Electron Devices Meeting (IEDM), 2022, pp. 23.5.1–23.5.4.

[4] H. L. Chiang, et al., “Cold cmos as a power-performance-reliability booster for advanced finfets,” in 2020 IEEE Symposium on VLSI Technology, 2020, pp. 1–2.

[5] E. Garzon, Y. Greenblatt, O. Harel, M. Lanuzza, and A. Teman, “Gain cell embedded dram under cryogenic operation-a first study,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 29, no. 7, pp. 1319–1324, 2021.

[6] M. Bhushan, A. Gattiker, M. B. Ketchen, and K. K. Das, “Ring oscillators for CMOS process tuning and variability control,” IEEE Trans. Semicond. Manuf., vol. 19, no. 1, pp. 10–18, Feb. 2006.

[7] C. L. Gan and C. Y. Huang, "Interconnects Reliability for Future Cryogenic Memory Applications," in Interconnect Reliability in Advanced Memory Device Packaging, Springer, 2023, pp. 185–207.

[8] Z. Chen, et al., "A Comparative Analysis of Low Temperature and Room Temperature Circuit Operation," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 33, no. 1, pp. 102-113, Jan. 2025.

[9] A. H. Hassan, et al., "Cryogenic Alternative: CMOS Versus Dynamic-Based Logic," 2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS), Springfield, MA, USA, 2024, pp. 1007-1010.