Challenges in Design and Modeling of Cold CMOS HPC Technology

Victor Moroz, Jamil Kawa, Xi-Wei Lin, Andrew R. Brown, Plamen Asenov, Jaehyun Lee, Mohit Bajaj, Tyler Michalak, Craig Riddet, Alexei Svizhenko, Renato Hentschke, and Søren Smidstrup
Synopsys, Mountain View, California, USA; Glasgow, UK; Bangalore, India; Hillsboro, Oregon; and Copenhagen, Denmark

Introduction

There are two fundamental trends observed in HPC (High Performance Computing) server farms that are becoming increasingly worrying. One is that their power consumption is becoming a major concern [1]. Another is that there is an increasing percentage of transistors on the chip that cannot be used simultaneously due to overheating. This is often referred to as dark silicon [2] (Fig. 1).

There is a hope that introducing cryogenic cooling into HPC can address both problems simultaneously [3]. This hope is based on the observation that transistor subthreshold slope gets remarkably steeper at cryogenic temperatures [4], which can enable transistor operation at a much lower power supply voltage (Vdd) compared to room temperature (Fig. 2).

In this work, we perform holistic DTCO (Design Technology Co-Optimization) analysis of cold CMOS technology, spanning from characterization of transistor and interconnect behavior at cryogenic temperatures all the way to power-performance evaluation of ring oscillators and logic blocks.

Methodology

We apply a holistic DTCO tool flow [5,6] to a typical 7nm FinFET technology (Fig. 3). We evaluate this technology at room temperature (300 K), at 150 K, and at 77 K. The 77 K temperature corresponds to liquid nitrogen cooling, and 150 K can be an attractive intermediate point between room temperature and liquid nitrogen.

Transistor models in the target range of cryogenic temperatures are tuned to reproduce experimentally observed subthreshold slope behavior [4] and carrier transport properties [7].

Results

At 77 K, transistor performance increases dramatically due to the much steeper subthreshold slope and due to carrier higher mobilities. When we match the on-state and off-state currents of NMOS and PMOS transistors to their 300 K values, we achieve this by lowering Vt by approximately 300 mV and by lowering Vdd from 0.95 V to 0.6 V.

Figures 4 and 5 show the on-state and off-state currents of NMOS and PMOS transistors at 300 K and at 77 K. Whereas at room temperature NMOS driving strength is stronger than PMOS by ~15%, we see a reversal at 77 K, with PMOS becoming ~30% stronger. The global on-state current variability at 77 K approximately doubles compared to 300 K, and approaches the level of variability that was observed at 28nm planar MOSFETs.

For the off-state currents, global variability explodes, especially for NMOS, from 20x at room temperature to 10,000x at 77 K. Practically, this is manageable, because what really matters in terms of leakage is the leakage budget of the entire chip that tends to contain billions of transistors. Considering the wildly wide Ioff statistics, we integrate the chip-wide off-state current distribution and shift the Vt such that the mean Ioff value meets the spec (Fig. 6).

The main reason for the dramatic variability increase at 77 K is the stronger transistor sensitivity to the gate and drain biases, which improves driving strength, but backfires in terms of the off-state current variability. However, once we meet the integrated chip-wide leakage budget by adjusting the Vt's upwards, this technology looks viable at 77 K operation.

978-1-6654-0685-7/21/$31.00 ©2021 IEEE

Running an inverter-based ring oscillator (RO) at 77 K shows that the SS (Slow PMOS and Slow NMOS) process corner achieves about 7x power reduction at iso-frequency, with Vdd of 0.3 V at 77 K versus Vdd of 0.8 V at 300 K (Fig. 7). At an intermediate temperature of 150 K, iso-frequency is achieved at 0.4 V power supply voltage with power reduction of 4x.

Alternatively, instead of power reduction, cold CMOS can be employed to achieve about 40% performance gain at iso-power conditions for HPC regimes (Fig. 7).

The encouraging results depicted in Figures 4 through 7 require reducing both the NMOS and the PMOS Vt's by about 310 mV, which is a challenging issue in terms of finding an appropriate HKMG (High-K Metal Gate) process flow. One potential solution is to use La-Hf dipoles for NMOS and Al-Hf dipoles for PMOS. Ab-initio analysis shows that the target Vt shifts can be achieved by dipole engineering, with 310 mV PMOS Vt shift enabled by introducing 8.7e13 cm^{-2} aluminum dipole density into the HKMG stack (Figs. 8 – 11).

For a typical 7nm FinFET technology, global and local transistor variations contribute approximately 30% and 70%, respectively. Global variation includes some NMOS-PMOS correlation, whereas local variability is uncorrelated (Fig. 12).

When we apply local and global variabilities to an 11-stage ring oscillator that represents behavior of a typical logic circuit, we see that the global variability dominates in terms of defining critical points at which 77 K circuits need to be evaluated (Fig. 13). Based on these results, it is sufficient to evaluate logic circuits at the SS, TT, and FF global process corners, as local variability contribution self-averages out.

To explore cold CMOS on a larger scale than an RO, we build a mini-library with about 50 logic cells and then use virtual PDKs (Process Design Kits) for 300 K, 150 K, and 77 K to build a logic block with about 15,000 cell instances (Fig. 14).

Power consumption of such a logic block exhibits similar behavior to the ring oscillator, with iso-frequency power consumption reducing by 4x at 150 K and by 7.5x at 77 K (Fig. 15). Two thirds of the power is consumed by capacitive switching, and the remaining third – by internal dissipation within cells.

The power consumption trend appears to follow a simple exponential between 300 K and 77 K. Extrapolating that trend towards lower temperatures suggests that a 10x power reduction might be possible within the sub-10 K temperature range.

Conclusions

Holistic DTCO analysis of cold CMOS shows a stronger FinFET driving strength that is offset by higher variability. Considering all factors, both the RO and the logic block show considerable power and/or performance gains at 150 K and 77 K CMOS.

References

[1] Nicola Jones, "How to stop data centers from gobbling up the world's electricity", Nature (2018)
[2] Greg Yeric, "Moore's law at 50: Are we planning for retirement?", IEDM Proceedings, pp. 1-8, (2015)
[3] DARPA LTLT (Low Temperature Logic Technology) Program – see darpa.mil (2021)
[4] Arnout Beckers, Farzan Jazaeri, & Christian Enz, "Theoretical Limit of Low Temperature Subthreshold Swing in Field-Effect Transistors", Electron Device Letters, v. 41, n. 2, pp. 276 – 279 (2020)
[5] Victor Moroz, Xi-Wei Lin, and Thuc Dam, "Logic Block Level DTCO is the New Moore's Law", EDTM Proceedings (2020)
[6] Victor Moroz, Xi-Wei Lin, Plamen Asenov, Deepak Sherlekar, Munkang Choi, Binjie Cheng, Suketu Parikh, Po-Wen Chan, and J. J. Lee, "Can We Ever Get to a 100nm Tall Library? Power Rail Design for 1nm Technology Node", VLSI Technology Proceedings (2020)
[7] Ken Uchida, Junji Koga, and Shin-ichi Takagi, "Experimental Study on Electron Mobility in Ultrathin-Body Silicon-on-Insulator Metal-Oxide-Semiconductor Field-Effect Transistors", J. Appl. Phys., v. 102, 074510 (2007)

978-1-6654-0685-7/21/$31.00 ©2021 IEEE

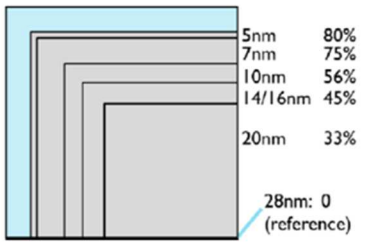


Fig. 1. Percent of dark silicon (with no design mitigation). Source: [2].

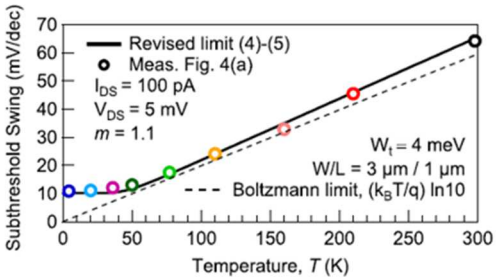


Fig. 2. Observed steeper subthreshold slope at cryogenic temperatures. Source: [4].

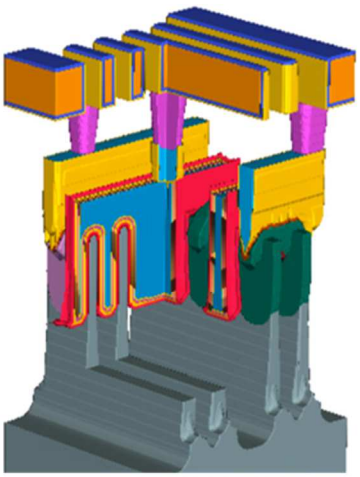


Fig. 3. 3D view of an inverter built with FinFET technology at 7nm design rules

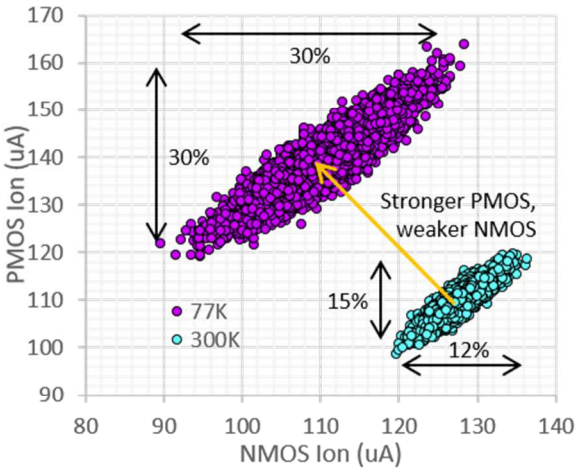


Fig. 4. On-state current at 300K at Vdd of 0.95 V and at 77K at Vdd of 0.6 V, including global variability. The Ion variability at 77K approximately doubles wrt 300K mainly due to the steeper subthreshold slope.

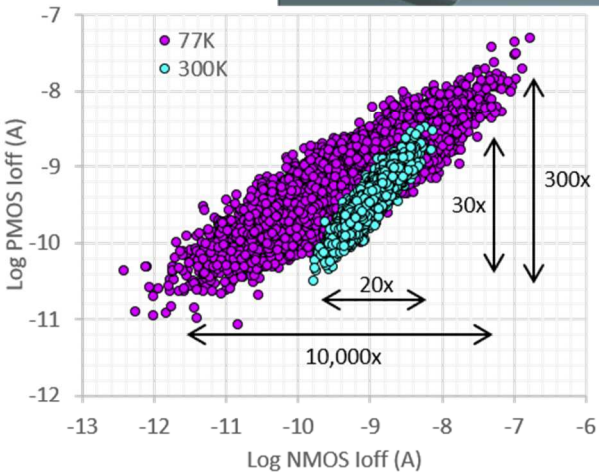


Fig. 5. Off-state current at 300K at Vdd of 0.95 V and at 77K at Vdd of 0.6 V, including global variability. The Ioff variability at 77K is hugely wider wrt 300K mainly due to the steeper subthreshold slope.

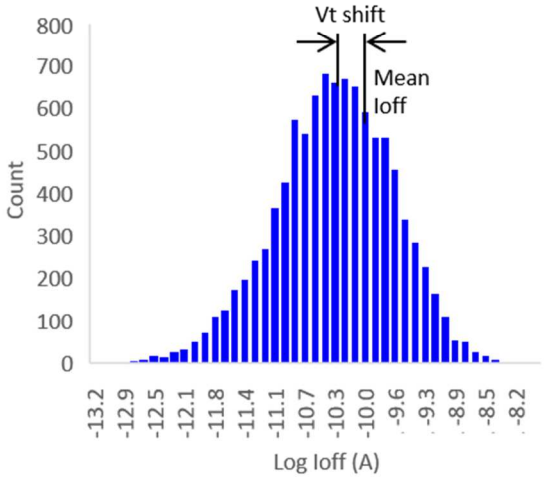


Fig. 6. Histogram of the off-state current at 77K, requiring 13 mV Vt increase for NMOS and 4 mV Vt increase for PMOS to get the target mean Ioff.

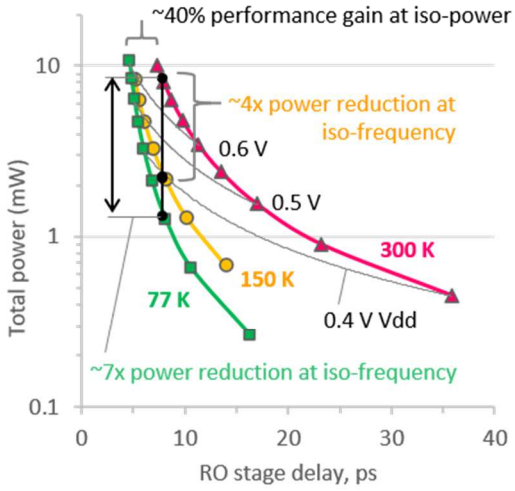


Fig. 7. Ring oscillator power-performance comparison at 300K (red) vs 150K (yellow) and 77 K (green) at SS process corner.

978-1-6654-0685-7/21/$31.00 ©2021 IEEE

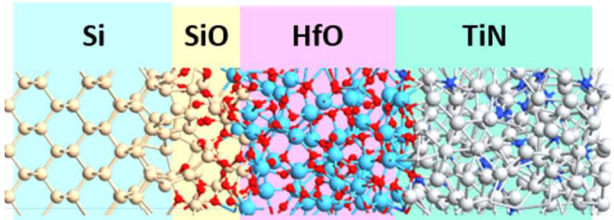


Fig. 8. PMOS HKMG atomistic structure with crystalline Si and amorphous SiO, HfO, and TiN. Generated via Molecular Dynamics by a sequence of ALD processes, and then connected to Ti-rich TiN. The final structure is relaxed using density functional theory in QuantumATK.

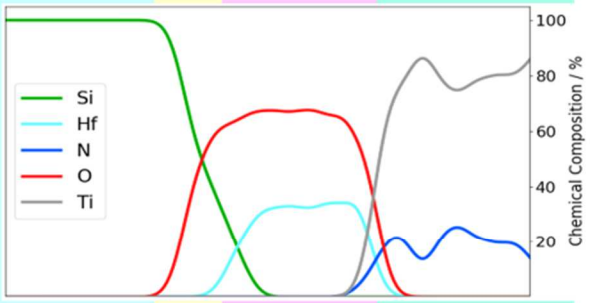


Fig. 9. The resulting chemical compositions extracted from atomistic HKMG structure depicted on Fig. 8. There is no stoichiometric SiO_2 observed, but instead there is a bridge of SiO_x between Si and HfO.

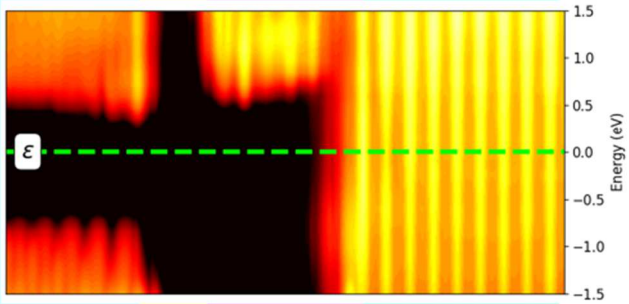


Fig. 10. Band diagram of PMOS HKMG without the presence of aluminum calculated using density functional theory and the hybrid functional HSE06.

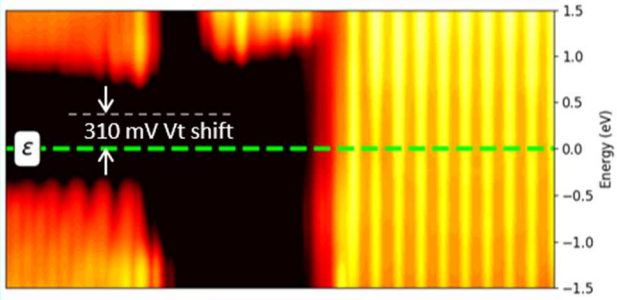


Fig 11. Addition of 8.7e13 cm^{-2} aluminum dipoles in the PMOS HKMG stack reduces PMOS Vt by 310 mV.

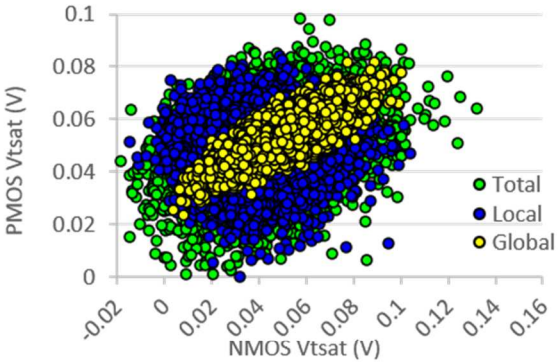


Fig. 12. Global (somewhat correlated), local (un-correlated) and total transistor variabilities at 77K.

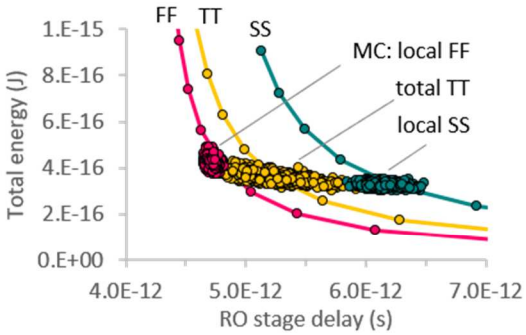
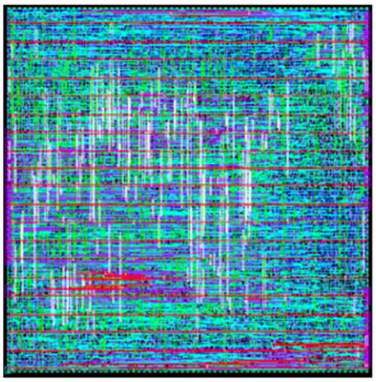


Fig. 13. Local, global (i.e. process corners) and total 11-stage ring oscillator variabilities at 77 K.

Fig. 14. Logic block with 15k instances after place-and-route step. The utilization is around 80%.



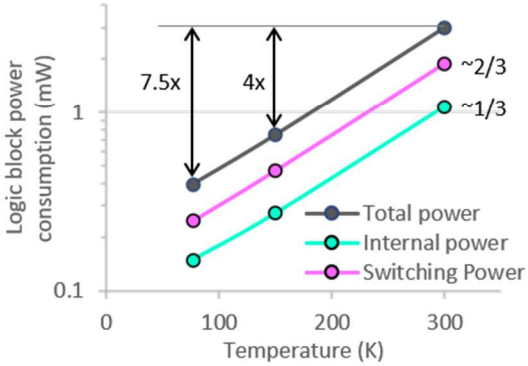


Fig. 15. Logic block power consumption vs temperature at iso-frequency for TT corner.

978-1-6654-0685-7/21/$31.00 ©2021 IEEE