A 64-Bit Arm CPU at Cryogenic temperatures: Design Technology Co-Optimization for Power and Performance

Rakshith Saligram[1], Divya Prasad[2], David Pietromonaco[2], Arijit Raychowdhury[1], Brian Cline[2]

[1]Georgia Institute of Technology, [2]Arm Inc

2021 IEEE Custom Integrated Circuits Conference (CICC) | 978-1-7281-7581-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/CICC51472.2021.9431559

Compute demand has grown over 100X within the last decade and has well surpassed the growth in classical Moore's Law transistor density (Fig. 1 (a) [1]). Plateaued dimension scaling even with fin depopulation, short channel effects and shrinking wire dimensions (leading to exponential rise in resistance) have made matters worse. Modern microprocessor designs are now equally limited by transistor and wire performance (e.g., Fig. 1 (b)). Thus, advancements in both transistors and interconnects are needed alongside new architectures to meet the datacenter and High-Performance Computing (HPC) demands. Low-temperature CMOS has emerged as a potential way to provide the needed process technology advancement [2], [3]. Operating CMOS at low temperatures (down to ~77K) improves transistor carrier mobility (μ), subthreshold swing (*SS*) and source/drain resistances, but also increases transistor threshold voltage, V_{th}, (due to Fermi potential shift and bandgap widening [4]). Despite the increased V_{th}, overall performance improvements with low temperature CMOS have been demonstrated with appropriate process changes [5]. Additionally, bulk resistivity of wires also improves with temperature reduction [10].

In this paper, we benchmark the potential benefit of low-temperature operation by performing a design and co-optimization implementation study of the 64-bit high-efficiency Arm© Cortex©-A53 (Design I). The key contributions are (i) Gen-I FinFET low-temperature models calibrated to Room-Temperature (RT) foundry data and complemented with its low-temperature dependencies with projected V_{th} tuning, (ii) low-temperature wire resistance models which are RT foundry calibrated, (iii) production-grade RT standard-cell libraries (SC) re-characterized across temperatures (*T*=300K, 200K, 150K and 100K) and supply voltages (V_{DD}=0.8V, 0.6V, 0.4V), (iv) Performance-Power-Area (PPA) analysis on the Arm core implementations across V_{DD} and *T*, and lastly (v) analysis on self-heating effects across temperatures for a high-performance Arm core (Design II) implemented on Gen-II FinFET.

The design and process assumptions are stated in Table I. In particular, for the device model, we utilize an open-source MIT Virtual Source (MVS) framework [6], calibrate to foundry Gen-I FinFET (Gen-1FF) I-V/C-V data at RT, and integrate its low temperature dependencies obtained from measurements in [3]. At each *T*, V_{th} is tuned (lowered at lower temperatures) such that devices across temperatures exhibit similar leakage currents to RT (300K) devices (Fig. 1 (c)). At 100K, the simulated V_{th} tuned devices achieve up to 68% higher saturation current as compared to RT at 0.8V V_{DD} (Fig. 1 (d)); note that this is >3X the typical 20% improvement in saturation current expected from node-to-node scaling by Moore's law [13]. Open source RT wire resistivity models that capture copper size effects [7-8] are calibrated to foundry RT resistance data for each metal layer of an 11-layer Back-End-Of-Line (BEOL) stack for the Gen-I FinFET foundry PDK. Total copper resistivity, ρ_{Cu}, comprises of (i) temperature-independent resistivity due to size effects (ρ_{size}) including surface scattering and grain bound scattering and (ii) temperature-dependent bulk resistivity of copper (ρ_{bulk}) [9]. Higher metal layers have smaller ρ_{size} component in ρ_{Cu} than lower layers (Fig. 2 (b)); thus, a larger improvement in resistivity is seen for global metals (21% at 100K) compared to local metals (17% at 100K).

Production-grade 9-track SC libraries (Low V_{th} flavor) are re-characterized at multiple *T* and V_{DD} using industry-standard tools (Fig. 2 (d)). At V_{DD} = 0.8V, the Gen-IFF logic gates exhibit 38% improvement in average rise/fall delay at 100K compared to RT, (Fig. 2 (c)) which is ~3X more than that obtained from Gen-IFF to Gen-IIFF scaling. With V_{th} reduced at low temperatures to keep leakage constant and increase saturation current, a larger overdrive voltage (V_{DD}-V_{th}) is created leading to an increase in charge inversion at the channel and hence increased input gate capacitance (Fig. 3 (a)). Considering the increased transistor peak current at lower temperatures countered by reduced switching time, the short circuit energy is similar across *T* with the same input slew; however, with large input slews, short circuit increases drastically at lower *T*. Utilizing the aforementioned low-temperature BEOL technology and SC, an Arm core is implemented using industry-standard EDA tools (Table I, Fig. 3 (b)) without its L1/L2 caches, independently, across *T* and V_{DD}.

First, we evaluate the peak performance achieved across *T* at V_{DD}=0.8V; up to a 1.56X performance boost is observed at 100K compared to RT; this, however, comes with a non-linear increase of >2X in power dissipation (Fig. 4 (a)-(b)). This is largely due to (i) increase in input gate capacitance of lower V_{th} devices, (ii) the upsizing of gates (resulting in larger input gate capacitance) and more buffering for hold fixing; all of which lead to higher switching and short-circuit power. It is expected that if this design is complemented with multi-V_{th} SC for non-critical paths, the power can be improved. The performance improvement at the core level surpasses that at the logic gate level (1.38X) which is attributed to the improvement in BEOL RC and the EDA tool optimizations. Our second approach is to take advantage of the increased voltage overdrive at low *T* and reduce V_{DD} whilst achieving the same 0.8V-RT-frequency to lower dynamic power. Iso-V_{DD} comparisons across *T* at 0.6V exhibit 1.87X performance boost at 100K compared to RT (0.6V RT frequency < 0.8V RT frequency); and at 0.4V V_{DD}, the 100K design is able to meet the target performance of the design at RT with 0.8V V_{DD}, while the RT design is not functional at 0.4V. The frequency-versus-power sensitivity across (*T*, V_{DD}) in Fig. 4 (c)-(d) showcases that at Iso-0.8V-RT-frequency, the 0.4V-100K design implementation dissipates 3.7X lower power. Interestingly, the 0.8V-RT-frequency can be achieved at 200K with 2X lower power, at 150K with 3X lower power, and at 100K with 3.7X lower power (observed at iso-performance line in Fig. 4 (d)).

In additional to the discussed low-*T* PPA benefits, we also quantify the material-to-system level thermal implications on a Gen II FinFET, server-market Arm core (Table II). Modern RT designs are performance limited due to the substantial increase in junction temperatures mandating frequency throttling for reliable operation. The thermal conductivity of silicon bulk (where most of the rise in *T* happens), improves by almost ~10X at 100K, i.e., conducts heat as well as copper metal, alleviating self-heating effects on die [11]. To study this, a single high-performance core is selected (Design II--to better capture self-heating effects than Design I) and benchmarked with the maximum-power workload at ambient temperatures of 298K and 100K with no design changes but only changes to material thermal conductivity values resulting due to *T*. Heat maps are extracted for the two temperature points using industry-standard thermal analysis tools (Cadence® Celsius ™) (Fig. 5) showing the spread of junction temperatures as well as the peak temperature on the die. Results exhibit 4X reduction in maximum temperature increase (ΔT_{max}) on a CPU at 100K owing to the improved silicon-bulk thermal conductivity. This study presents potential to use even higher processor counts at a system-level than RT systems while adhering to the same Thermal-Design-Power limit.

In this paper, for the first time, we showcase the benefit of low-temperature CMOS at a design-level by simulating and optimizing the design implementation of an Arm Cortex-A53 core for cryogenic computing and capturing the impact of low *T* from materials-to-systems. RT calibrated device-/wire- models are extended to 100K temperature range and production-grade SC libraries are re-characterized at every design point to obtain maximum benefits in terms of either performance (1.56X improvement at 0.8V/100K) or performance-per-watt (4X improvement at 0.4V/150K) at the chip level. These, alongside the improved self-heating effects at low *T*, could provide multiple generations of Moore's Law's advancements.

References:

[1] S. Rumley, et. al, Parallel Computing 2017 [2] W. Chakraborty, et.al, IEDM 2019 [3] W. Chakraborty, S. Datta., Discussions on measured I-V for Gen-I FinFET down to 6K [4] A. Beckers et al., arXiv 2019 [5] H.L. Chiang et al, VLSI-T Symp 2020 [6] L Wei, et al, IEEE TED 2012 [7] K. Fuchs, Math. Proc. Of Cambridge Phil. Soc. 1938 [8] A. F. Mayadas et al, Phys Rev 1970, [9] Cryocomp, Cryodata Inc. 1999 [10] G.E. Childs et al, Natl. Bureau of Stds, 1973 [11] B. Cline et al, VLSI-T Symp 2019 [12] D. Prasad, et al IEDM 2019 [13] I. Cutress et al, AnandTech May 8th, 2019.

978-1-7281-7581-2/21/$31.00 ©2021 IEEE

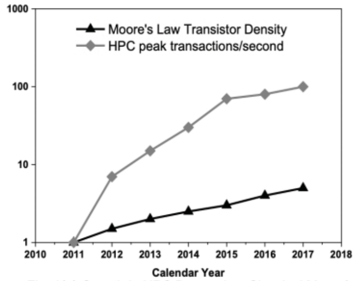


Fig. 1(a) Growth in HPC Demand vs Classical Moore's Law Transistor Density

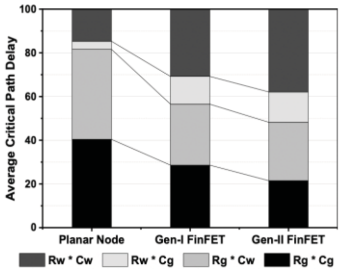


Fig. 1(b) Elmore Delay Breakdown of top critical paths in High Perf CPU across Foundry Nodes (Rg, Cg: Gate, Rw, Cw: Wire)

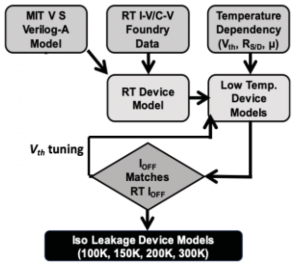

MIT V S Verilog-A Model
RT I-V/C-V Foundry Data
Temperature Dependency (V_{th}, $R_{S/D}$, μ)
RT Device Model
Low Temp. Device Models
V_{th} tuning
I_{OFF} Matches RT I_{OFF}
Iso Leakage Device Models (100K, 150K, 200K, 300K)

Fig. 1(c) Device Modelling & Calibration Methodology

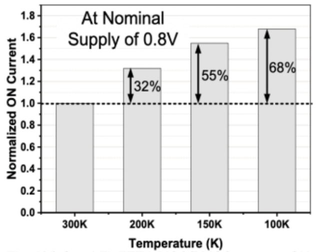


Fig. 1(d) Gen-I FinFET NMOS/PMOS average ON Current with V_{th} Tuning for iso-leakage

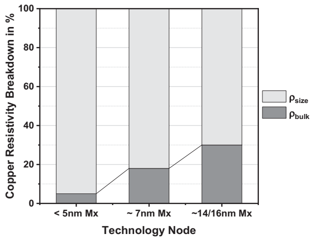


Fig. 2(a) Cu resistivity breakdown for Local Metals (Mx) across nodes; Dimensions/Models derived from [11]-[12]

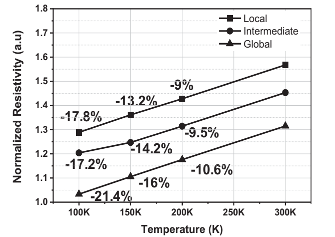
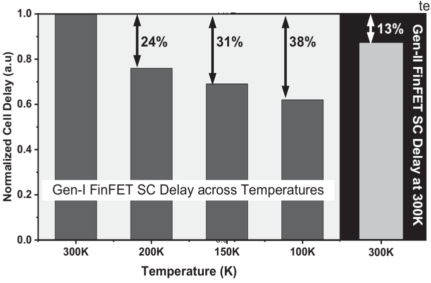


Fig. 2(b) BEOL Stack Cu resistivity across temperatures for foundry Gen-I FinFET and projected for lower temperatures using models from [7]-[8]



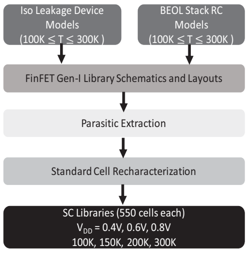

Iso Leakage Device Models (100K ≤ T ≤ 300K)
BEOL Stack RC Models (100K ≤ T ≤ 300K)
FinFET Gen-I Library Schematics and Layouts
Parasitic Extraction
Standard Cell Recharacterization
SC Libraries (550 cells each)
V_{DD} = 0.4V, 0.6V, 0.8V
100K, 150K, 200K, 300K

Fig. 2(c) Improvement in standard cell delay (avg. rise/fall times of all cells in the library) across Temperature compared with improvements obtained from technology scaling

Fig. 2(d) SC Library Design Flow to recharacterize libraries at Low Temperatures

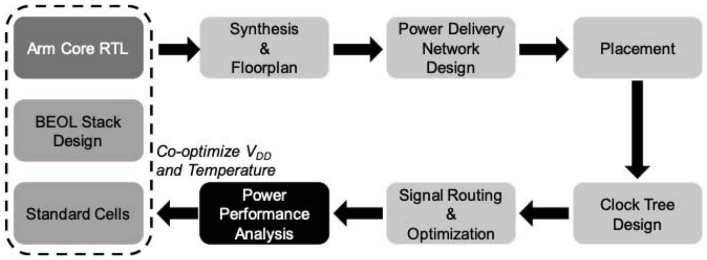

Arm Core RTL
BEOL Stack Design
Standard Cells
Synthesis & Floorplan
Power Delivery Network Design
Placement
Clock Tree Design
Signal Routing & Optimization
Power Performance Analysis
Co-optimize V_{DD} and Temperature

Fig. 3(b) VLSI design Flow utilized in this work to benchmark Arm microprocessor and perform design technology co-optimization

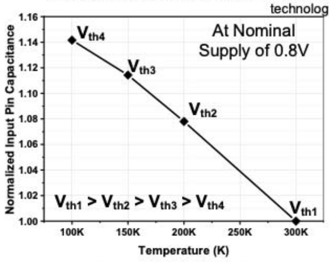


Fig. 3(a) Standard Cell average input-pin capacitance across temperatures and dependency on V_{th}

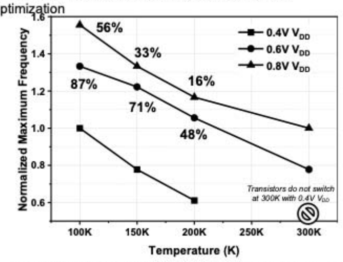


Fig. 3(c) Maximum Frequency across temperatures and supply voltages (% indicated are normalized to maximum RT frequency at each V_{DD})

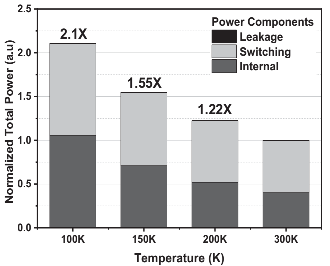


Fig 4 (a) Power Breakdown at maximum frequency across Temperatures at nominal supply of 0.8V

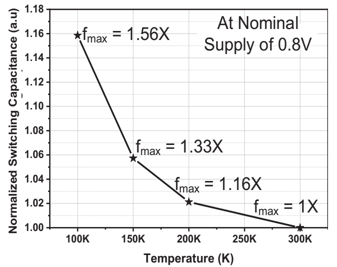


Fig 4 (b) Variation of Extracted Switching/Output Capacitance of all gates in the design at f_{max}

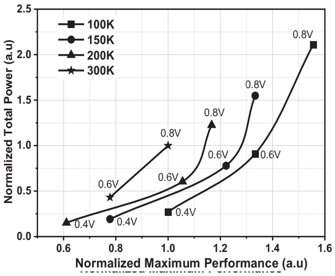


Fig 4 (c) Total Power Dissipation at multiple supply voltages across different temperatures

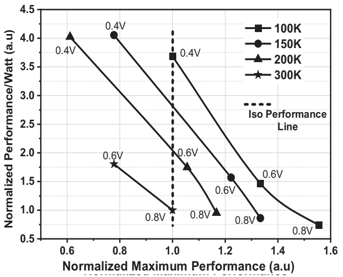
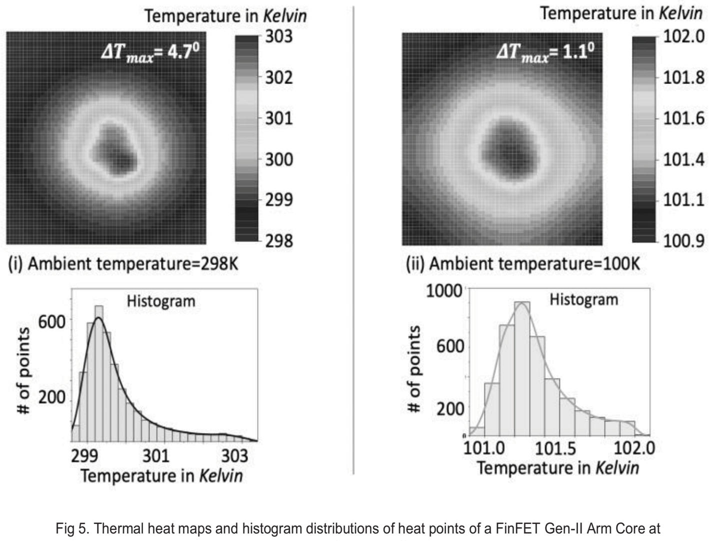


Fig 4 (d) Performance per Watt Improvements at low temperatures



(i) Ambient temperature=298K

(ii) Ambient temperature=100K

Fig 5. Thermal heat maps and histogram distributions of heat points of a FinFET Gen-II Arm Core at T = 298K and T = 100K depicting lower self-heating effects at lower temperatures owed to the improved thermal conductivity of bulk silicon, showing potential to pack more cores on die to achieve higher performance at 100K; typically designs are throttled due to junction -temperature rise (Δt_{max} = maximum rise in temperature from ambient)

Table I: Specifications and Process assumptions for PPA studies	
Process	Foundry Gen-I FF (Low V_{th})
Supply Voltage (V_{DD})	0.8V (nominal), 0.6V, 0.4V
Temperature Points	300*K*, 200*K*, 150*K*, 100*K*
BEOL stack and RC models	At 300*K*: Foundry calibrated; T<300*K*: Models and measured low-temperature resistivity from literature [3],[4],[7]
Device Models	At 300*K*: Foundry calibrated (LVT); T<300*K*: Models and measured parameter temperature dependency from [7, 8]
Design (Design I)	High-efficiency Arm Cortex-A53 (excluding L1/L2 caches)
Standard Cells (SC)	Production-grade 9-track library with ~550 cells

Table II: Thermal Analysis Setup	
Process:	Foundry Gen-II FF (Multi V_{th})
Silicon bulk thermal conductivity	300*K*: 0.13 mW/(μm·C); 100*K*: ~1 mW/(μm·C)
Copper thermal conductivity	300*K*-100*K*: 0.4 mW/(μm·C)
Design (Design II)	High-performance Arm Core (with L1/L2 caches)
CPU Workload	Max Power
System assumption	Single core, Single workload

Acknowledgements: The authors are grateful for the useful discussions with Dr. Suman Datta and Wriddhi Chakraborty from Univ. of Notre Dame via ASCENT, and would like to thank product teams at Arm Inc. for design support.