COSMIC: A Multi-Vector-Core Heterogeneous RISC-V SoC for Intelligent Audio DSP in Intel 16

Ethan Gao, Jasmine Angle, Lucy Revina, Jacob Leigh, Wenda Zhang, Naichen Zhao, Tushar Goyal, Michael McCulloch, Jonathan Wang, John Lomax, Jessica Fan, Mihai Tudor, Rachel Lowe, Ted Kim, Kevin He, Nico Castaneda, Anto Kam, Rahul Kumar, Rohan Kumar, Felicity Aktan, Connor Dang, Shichen Qiao, Joshua Lin, Andy Chen, Minh Nguyen, Vedang Joshi, Bryan Ngo, Ella Schwarz, Ken Ho, Viansa Schmulbach, Nikhil Jha, Yufeng Chi, Jerry Zhao, Borivoje Nikolić

University of California, Berkeley, CA, USA, eygao@eecs.berkeley.edu

***Abstract*—The growing demand for intelligent audio and speech signal processing on mobile terminals requires energy efficient programmable support for digital signal processing (DSP) and machine learning (ML) applications. This paper presents COSMIC, a heterogeneous RISC-V SoC designed for DSP/ML workloads. Leveraging an agile open-source design flow, the 2mm x 2mm SoC, fabricated in Intel 16 FinFET, integrates four RISC-V vector cores, a single-path delay feedback FFT accelerator, convolution accelerator, and programmable direct memory access engine running at up to 1.25GHz, enabling efficient execution of spectral analysis, filtering, and machine learning workloads. This work accelerates convolution compute by up to 12x and FFT transforms by 2x. The fully featured SoC runs a 260k parameter Llama model at up to 85 tok/sec and 1.11 mJ/tok and real-time audio beamforming applications.**

***Index Terms*—System-on-a-chip, heterogeneous computing, VLSI, digital signal processing, RISC-V, domain-specific accelerator.**

I. Introduction

Edge devices implement a diverse set of audio, voice and image processing applications by using a combination of traditional digital signal processing (DSP) and machine learning (ML) algorithms. Some example applications include speech recognition, voice synthesis, and audio beamforming, which require efficient processing of convolution, filtering, and domain transform kernels. Typically, the front-end of the signal processing pipeline implements traditional signal processing functions, while the back-end increasingly relies on machine learning models requiring efficient and flexible implementation of vector and matrix linear algebra. Recently, DSP applications have been implemented on very-long instruction word (VLIW) [1] or packed-SIMD machines to support flexibility while maintaining energy efficiency by utilizing varying application vector lengths.

This work presents COSMIC, a heterogeneous system-on-chip (SoC) for digital signal applications that is based on four short vector length open-source RISC-V cores that can attain high utilization with both short and long vector lengths [2]. The SoC integrates fixed-function accelerators for fast-fourier transform (FFT), convolution, and direct-memory access to speed up common application kernels. All programmable and fixed-function cores are connected via a network-on-chip (NoC) to a banked L2 cache and standard peripherals. COSMIC has been generated from a design template by using the Chipyard framework [3], from concept to tapeout in 15 weeks by a class of mostly undergraduate students. The effectiveness of the 16nm FinFET test chip is demonstrated by analyzing convolution, FFT, and vector benchmarks at performance and efficiency corners as compared to scalar CPU workloads as seen in Table 2. The flexibility of the architecture to support the increased demand for DSP in AI-IoT systems is demonstrated by running a TinyStories Llama model at up to 85 tokens/second and down to 1.11 mJ/token. The test chip also runs live beamforming and I^{2}S audio recording and playback demos.

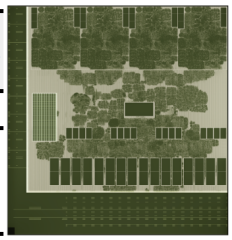
II. Architecture and Implementation

A. Chip Architecture

Fig. 1 and Table 1 outline the SoC's architecture. COSMIC contains four Saturn tiles, each having a 5-stage in-order scalar RISC-V Rocket (RV64GC) core [4] and a Saturn vector unit [2] for vector operations. Chip interconnects are instantiated using the Constellation network-on-chip (NoC) generator [5], with one primary 128-bit unidirectional 1-D torus system NoC, a crossbar that attaches all peripherals, and a memory bus crossbar connecting general memory.

TABLE 1. CHIP SPECIFICATIONS AND DIE MICROGRAPH.

Chip Specifications		**IP Blocks**
Technology	Intel 16 FinFET	Vector Cores, DMA
Area	4 mm^2	FFT, Convolution Accel
Scratchpad	32 kB	I2S Audio
Operation	**Nominal**	**Functional**
Voltage	0.85 V	0.55–1.10 V
Frequency	900 MHz	50-1250 MHz
Power	340 mW*	10.9–875 mW*
Energy Efficiency	254 GOPS*/W	137-440 GOPS*/W



*power and MAC operations on 4 fully saturated vector cores (2 ops / MAC)

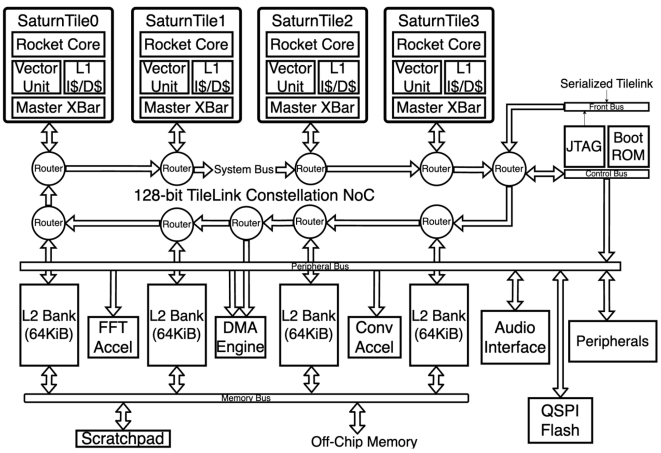


Fig. 1 COSMIC SoC architecture.

TABLE 2. EVALUATION SUMMARY ACROSS PERFORMANCE (HIGH VOLT AND FREQ) AND EFFICIENCY (LOWEST ENERGY) CORNERS.

PERFORMANCE BENCHMARKS (HIGH PERF = 1100 MHz @ 1.1V; HIGH EFF = 150 MHz @ 0.55V)

Benchmark	**IGEMM**		**16384-Point 1D Convolution Transform**						**128-Point Fast Fourier Transform**					
Block	One Saturn Core		FP32 CPU		FP16 Conv Engine				CPU-FFT		DMA-FFT			
Corner	High Perf.	High Eff.	High Perf.	High Eff.	High Perf.	*CPU/x*	High Eff.	*CPU/x*	High Perf.	High Eff.	High Perf.	*CPU/x*	High Eff.	*CPU/x*
Avg Latency (ms)	1.839**E-4**	1.349**E-3**	3.812	27.260	0.309	*12.32*	2.271	*12.00*	0.029	0.214	0.014	*2.14*	0.100	*2.14*
Avg Energy (uJ)	0.06	0.0182	111.299	31.987	9.065	*12.28*	2.676	*11.96*	0.881	0.257	0.403	*2.18*	0.119	*2.17*
Avg Power (mW)	343.20	13.53	292.00	11.73	293.11	*1.00*	11.78	*1.00*	302.21	12.02	296.11	*1.02*	11.87	*1.01*
(averages per iteration)	(64 x 64 8-bit int matmul kernels)		(hand written floating point kernels)						(port of open source KISS-FFT)		(hand written 16-bit fixed point kernels)			

The L2 cache comprises four 64KiB banks arranged as 256KiB 4-way set associative cache. The chip also contains a 32KiB on-chip scratchpad and a 10MHz-2.2GHz PLL provided by Intel.

The SoC contains five clock domains, one for each of the Saturn tiles and one shared by the remainder of the SoC. The chip implements I^{2}S, UART, JTAG, I^{2}C, PWM, SPI, QSPI, serial TileLink, and GPIO peripheral interfaces. The two QSPI modules support execute in place (XIP) for Flash and PSRAM memory. For access to bulk off-chip memory and debugging, the chip exposes two serialized TileLink interfaces, one 8-bit bidirectional cached interface for memory and one 1-bit coherent interface for connecting peripherals.

B. Saturn RISC-V Cores With Embedded Vector Units

The Saturn RISC-V [2] cores in COSMIC combine scalar RV64GC pipelines with a datapath that supports RISC-V vector extension 1.0 (RVV1.0) optimized for DSP workloads. Unlike general-purpose cores, VPUs are capable of applying the same operation across wide vectors of data in a single instruction, making them highly efficient for operations such as convolution, filtering, and spectral analysis. This single-instruction, multiple-data capability reduces control overhead and improves throughput, aligning well with the performance and energy constraints of real-time DSP applications. Saturn is an extensible and highly parameterized generator of scalable vector units that can generate a wide variety of vector datapath widths (DLEN). It adopts a decoupled access-execute architecture, comprising a vector frontend unit for instruction dispatch and fault checking, a vector load-store unit for memory operations, and a vector datapath for executing vector instructions arranged as shown in Fig. 3. COSMIC's four Saturn instances are configured as VLEN=256, DLEN=64 vector units, offering a short 256-bit vector length in a compact 64-bit datapath width for 64-bit operations per cycle, suitable for audio processing, offering a different evaluation point from big/little cores in [6].

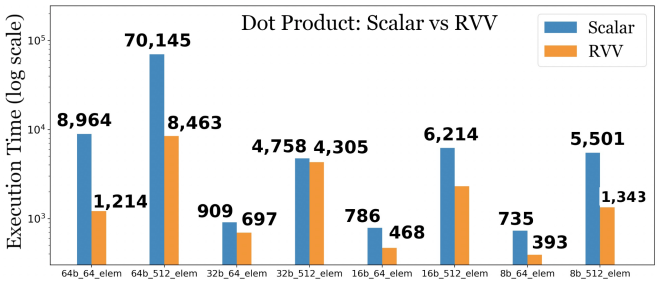


Fig. 2 Integer dot-product scalar versus vector comparison.

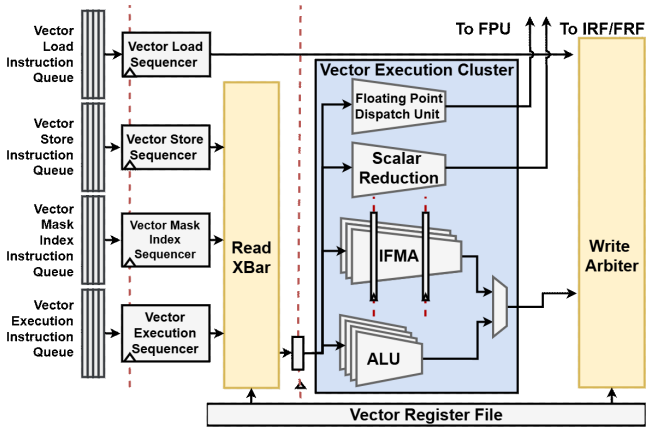


Fig. 3 Saturn vector core architecture.

4 Core Saturn POWER DRAW RANGE

Corner	High Perf.	High Eff.
Time (ms)	2,000	2,000
Avg E (mJ)	1,681.68	21.46
Avg P (mW)	840.84	10.73
@Freq (MHz)	1,250.00	50
@Volt(V)	1.1	0.55

*MACs fully saturating cores

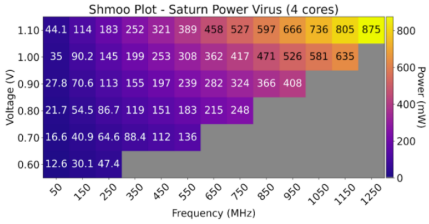


Fig. 4 Saturn power draw at near-maximum utilization.

COSMIC's Saturn is benchmarked on representative DSP workloads, including vector dot-products, integer general matrix multiply (IGEMM), and floating-point 16-bit 1D convolutions which are summarized in Fig. 2.

C. Direct Memory Access Engine

In addition to the vector cores, COSMIC contains a streaming convolution accelerator and memory-mapped I/O (MMIO) FFT accelerator and must contend with large amounts of audio data coming from multiple I^{2}S ports and main memory. To facilitate the efficient transfer of data between the peripherals and accelerators without burdening the CPU cores, a high performance fully programmable DMA engine has been designed. The DMA engine features seven independent channels split into two main types. Three C-channels are designed to accelerate bulk data transfer between memory locations and four P-channels perform a handshake before each transfer to enable flow control when talking to lower bandwidth peripherals like I^{2}S. Each channel is independently programmable and supports transactions of up to 4 MiB long and 64 bytes wide, with programmable stride. Each channel has a transaction queue to allow

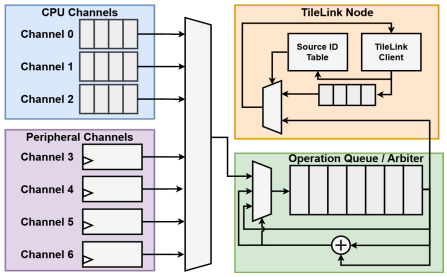


Fig. 5 DMA engine architecture.

TABLE 3. MEMORY BANDWIDTH BENCHMARKS.

Bandwidth Tests	GLIBC Memcpy		RVV Memcpy		DMA Memcpy
Block	Scalar CPU	DMA/CPU	Vector Core	DMA/CPU	DMA Engine
L2$ @ 900 MHz	180.475	*11.08*	447.711	*4.47*	2,000.168
L2$ @ 1250 MHz	250.66	*12.30*	621.82	*4.96*	3,084.04
Serial TileLink*	4.293	*2.40*	6.160	*1.67*	10.301
*frequency irrelevant	MiB/sec		MiB/sec		MiB/sec

enforcing transaction ordering. DMA memory accesses are fully pipelined and can support up to 8 in-flight requests. This architecture is summarized in Fig. 5. As seen in Table 3, the block dramatically speeds up bulk memory copies compared to the scalar and vector cores, both within the L2 cache and when communicating to off chip memory over the SerialTL bus thanks to its ability to aggressively pipeline accesses. It has a peak bandwidth of 3.1 GiB/s when running at 1.25 GHz, a 12.3× improvement over *glibc*'s standard *memcpy* routines.

D. Convolution Accelerators

The convolution accelerator implements a low-latency, streaming datapath for 1D convolutions over FP16 or INT8 inputs. Input and kernel data are packed into 64-bit words and transferred via DMA into a 256 entry queue. The pipelined datapath feeds shift registers into parallel multiplication units and a binary tree adder structure, with pipelining at every other adder level as described in Fig. 6. A five-state FSM (idle, head, body, tail, halt) orchestrates input/output handshaking, pipeline delay masking, and status handling. Finally, outputs are packed and enqueued into the output queue. The integration of a 1D convolution accelerator targets high-throughput, low-latency execution of core signal processing workloads such as filtering, cross-correlation, and feature extraction. These are convolution-dominant, compute-intensive, and highly regular in DSP applications.

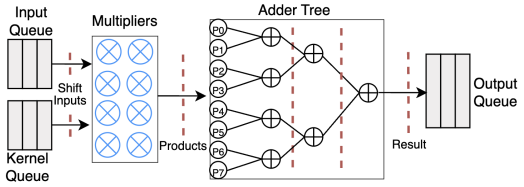


Fig. 6 Convolution accelerator architecture.

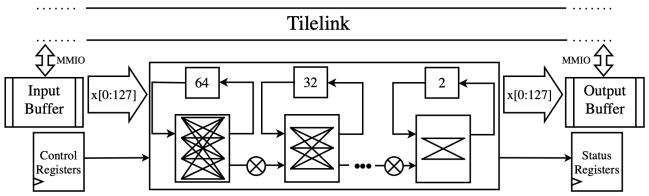


Fig. 7 FFT processor design.

As seen in Table 2, convolution accelerates the compute by 12× with a 12× average energy savings and no average power consumption increase compared to software.

E. Single-Path Delay Feedback FFT Processor

While the Saturn cores can execute FFTs, their prevalence in DSP applications justifies inclusion of a fixed-function FFT accelerator. The integrated FFT design is pipelined such that data flowing into each stage is time-multiplexed. Each stage of the FFT combines pairs of points and utilizes buffers to align the first-half of the input points with the second-half. The same hardware is used to combine N/2 pairs of points at each stage, as shown in Fig. 7. The integrated SDF-FFT [7] is configured as a 128-point Radix-2 FFT with one pipeline stage and is interfaced over MMIO registers. The FFT utilizes a 32 bit fixed point data format split into two 16 bit fixed point numbers for the real and imaginary components. The SDF-FFT accelerator was evaluated by comparing its performance and energy efficiency to KISS-FFT, a popular C FFT library, as well as with an end-to-end application to recognize notes from an audio file. Table 2 demonstrates an improvement of 2.14× in performance and 2.17× in energy consumption over software.

F. Audio Subsystem

COSMIC features an 8-channel audio subsystem designed to allow end-to-end intelligent audio processing applications. The audio interface features 4 independent stereo I^2S channels to interface with commercial high quality audio codecs, enabling multi-microphone input and spatial audio output, alongside two built-in 16-bit $\Delta\Sigma$ DACs for direct analog playback without external components. The $\Delta\Sigma$ DACs are 1-bit first order $\Delta\Sigma$ modulators attached to I^2S channel 0. The I^2S transceiver is fully programmable and spec compliant, supporting 8, 16, 24, and 32-bit audio samples and I^2S clock generation with a configurable $2\text{-}2^{16}$ clock divider for 240 Hz - 8 MHz sample rates. The ability to act as an I^2S peripheral with externally driven clocks also enables arbitrary sample rates. The high level architecture is described in Fig. 8. Each I^2S channel presents 4 independent queues for transferring packed stereo audio samples and contains onboard data conversion between FP16 and integer audio for integration with the convolution engine. It provides full DMA support including DMA requests to allow audio playback and processing without CPU intervention. The audio subsystem end-to-end demonstration with an I2S DAC is shown in Fig. 9.

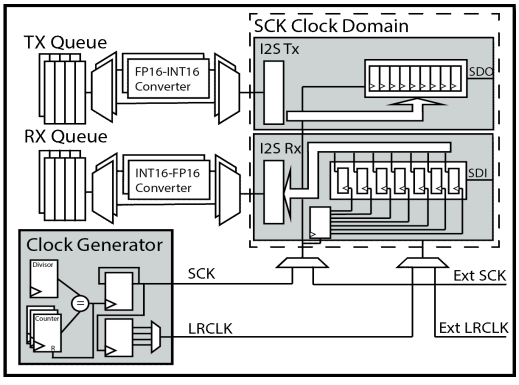


Fig. 8 I^2S audio subsystem architecture.

III. Measurement and Results

A. Evaluation Setup

The 2mm x 2mm test chip was fabricated in Intel 16 technology as one of four tiles in a 16mm^2 array, then mounted onto an interposer package to integrate easily with a custom PCB and FPGA setup for bring-up and evaluation. In the main setup, an Intel Agilex FPGA connected to a host through a USB interface sends the commands to the chip across a 560-pin FMC+ connector on the FPGA, allowing access to external DDR4 memory over Serialized TileLink. A second setup utilizes similar configuration with DataStorm FPGAs through custom breakout boards as seen in Fig. 9. All benchmarks were designed using a custom UART-based protocol to automate power and performance data collection. At the nominal voltage of 850 mV the maximum clock frequency is 900 MHz. The maximum clock frequency of 1.25 GHz is at the maximum tested voltage of 1.1V. Realistic benchmark performance is summarized in Table 2 while the energy consumption at core saturation is shown in Fig. 4.

B. End-to-End Applications

The audio subsystem is capable of receiving audio data and replaying it through the I^{2}S channels with an external DAC as well as through the ΔΣ DAC. The I^{2}C peripheral interface on the chip assists in configuring the external DAC with the appropriate sampling rates and bit depth. The DMA engine enables efficient intermediate CPU computation, such as applying a reverb effect, placing data into a double buffer and playing the new audio. We then run a dual microphone beamforming application on the test chip. Two microphones are set up in a front-facing broadside array and their responses are summed and normalized to be more responsive to sound from the front and back but attenuate sound from the sides. COSMIC is also well suited for edge inference tasks thanks to its efficient Saturn vector engines. The TinyStories [8] LLM and Llama model architecture were ported to our chips with the attention dot products vectorized with RVV. The 260K parameter model shows our system can run full end-to-end ML workloads, not just individual kernels. While our performance is largely limited by the relatively low bandwidth Serial TileLink interface, we achieve peak performance of 85.5 tokens/second, an average speed of 23.8 tokens/second, and maximum efficiency of 1.1 mJ/token as seen in Table 4.

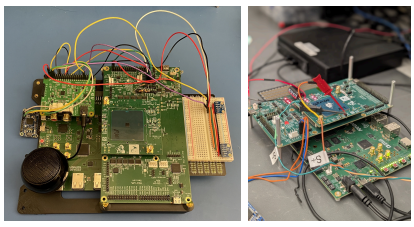


Fig. 9 Demo and evaluation setup.

TABLE 4. LLAMA MODEL AT VARIOUS CORNERS.

Llama 260k-Parameter TinyStories Quantized Int 8 Model

Operating Corner	Peak Tokens/Sec	Avg Tokens/Sec	Energy/Token
Nominal (.85V 900 MHz)	50.77	22.77	5.58 mJ
High Eff. (.55V 150 MHz)	20.62	10.44	1.11 mJ
High Perf. (1.05V 1100 MHz)	85.48	23.78	10.44 mJ

TABLE 5. COMPARISON TABLE TO RECENT WORK.

Paper	COSMIC [this work]	NeCTAr [9]	ECHOES [10]	Marsellus [11]
Technology	Intel 16	Intel 16	TSMC 65nm	GlobalFoundries 22nm
Year	2025	2024	2023	2023
Die Area	4mm^2	4mm^2	4mm^2	18.7mm^2
Int Precision	INT 8	INT 8	INT 8 [?]	INT 8 [?]
Voltage	0.55-1.1V	0.55-0.85V	0.9-1.2V	0.5-0.8V
Power	10.9-875mW	7.3-171mW	20-133mW	123mW
Max Freq	1250MHz	400MHz	350MHz	420MHz
Peak Eff.	558 GOPs/W (MACs on 4 cores)	109 GOPs/W (near mem MACs)	199.8 GOPs/W (FFT frequency-domain)	3.32 GOPs/W (MAC/convolution)
Peak GOPs	120 GOPs	6.02 GOPs	0.16 GOPs	180 GOPs

IV. Conclusion

COSMIC demonstrates the viability of agile, open-source methodologies in designing high-performance SoCs for audio DSP and edge ML workloads that compare favorably with the state of the art in Table 5. By integrating 4 RV64GCV vector cores with domain-specific accelerators (FFT, convolution), a custom DMA engine, and an I^{2}S audio subsystem, the SoC achieves 12× speedup in convolution and 2× faster FFT transforms over software implementations while improving energy efficiency. The architecture's versatility is validated through end-to-end applications including mic beamforming, reverb effects, and inferencing a 260k-parameter Llama model. The design leverages the meta-programming and parameter systems made possible by open-source module generators and tools within the Chipyard framework, integrating open-source libraries with commercial EDA tools.

Acknowledgment

This project is the work of many in UC Berkeley's special topics chip tapeout and bring-up classes. We thank Apple's New Silicon Initiative for supporting the tapeout class through guest lectures, design reviews, and funding for course staff; and Intel University Shuttle Program for donating the chip fabrication and packaging, and NSF CCRI ENS #2016662 and NSF POSE 2303735 Awards. We acknowledge Jared Zerbe, Ajith Amerasekera, Eric Smith and Ramesh Abhari from Apple, and Bryan Casper, Matt Rebsom, and Nancy Robinson from Intel for support, SLICE and BWRC students, staff, and, and member companies.

References

[1] E. Mahurin, "Qualcomm® Hexagon™ NPU," *2023 IEEE Hot Chips 35 Symposium (HCS)*, Palo Alto, CA, USA, 2023, pp. 1-19, doi: 10.1109/HCS59251.2023.10254715.

[2] J. Zhao *et al.*, 'The Saturn microarchitecture manual', Dec. 2024.

[3] A. Amid *et al.*, "Chipyard: Integrated design, simulation, and implementation framework for custom SoCs," *IEEE Micro*, vol. 40, no. 4, pp. 10–21, 2020.

[4] K. Asanović *et al.*, 'The Rocket Chip generator', *University of California, Berkeley, Tech. Rep.*, Apr. 2016.

[5] J. Zhao, A. Agrawal, B. Nikolic, and K. Asanovic, "Constellation: An open-source SoC-capable NoC generator," in *15th IEEE/ACM Int.Workshop on Network on Chip Architectures*, 2022, pp. 1–7.

[6] V. Jain *et al.*, "Cygnus: A 1 GHz heterogeneous octa-core RISC-V vector processor for DSP," in IEEE Symp. VLSI Technol. and Circuits, 2025.

[7] V. M. Milovanović and M. L. Petrović, "A highly parametrizable Chisel HCL generator of single-path delay feedback FFT processors," *2019 IEEE 31st International Conference on Microelectronics (MIEL)*, Nis, Serbia, 2019, pp. 247-250, doi: 10.1109/MIEL.2019.8889581.

[8] R. Eldan and Y. Li, "TinyStories: How small can language models be and still speak coherent English?", *arXiv e-prints*, Art. no. arXiv:2305.07759, 2023. doi:10.48550/arXiv.2305.0.

[9] V. Schmulbach, "NeCTAr: A heterogeneous RISC-V SoC for language model inference in Intel 16", *arXiv e-prints*, Art. no. arXiv:2503.14708, 2025. doi:10.48550/arXiv.2503.14708.

[10] M. Sinigaglia et al., "ECHOES: a 200 GOPS/W frequency domain SoC with FFT processor and I^{2}S DSP for flexible data acquisition from microphone arrays," *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, Monterey, CA, USA, 2023, pp. 1-5, doi: 10.1109/ISCAS46773.2023.10181862.

[11] F. Conti *et al.*, "Marsellus: A heterogeneous RISC-V AI-IoT end-node SoC With 2–8 b DNN acceleration and 30%-boost adaptive body biasing," in *IEEE Journal of Solid-State Circuits*, vol. 59, no. 1, pp. 128-142, Jan. 2024, doi: 10.1109/JSSC.2023.3318301.