KODIAK: RVV Vector Processor Chiplet Featuring Outer-Product Matrix Engine, Chip-to-Chip Link, and UCIe Interface in Intel16

Kevin Anderson*, Di Wang*, Rahul Kumar, Rohan Kumar, Ella Schwarz, Mahsa Sadeghi, Miles Rusch,
Lux Zhang, Yufeng Chi, Dima Nikiforov, Daniel Lovell, Jerry Zhao,
Vikram Jain, Yakun Sophia Shao, Krste Asanović, and Borivoje Nikolić
University of California, Berkeley, Berkeley, CA, USA
Email: {kevinand, di_wang, rahulkumar, rohankumar, schwarzem, mahsa.sadeghi, miles.rusch,
iansseijelly, chiyufeng, vnikiforov, dlovell98, jerryz123, vikramj, ysshao, krste, bora}@berkeley.edu
*Equal contribution.

***Abstract*—KODIAK is a 16 mm^2 dual-core RISC-V Vector 1.0-compliant vector processor chiplet fabricated in Intel16 and operating up to 810 MHz. It features a directly integrated outer-product matrix unit, a Chip-to-Chip Link, and a UCIe-compatible interface. Two KODIAK chiplets are co-packaged on an organic 2D interposer, with one chiplet rotated by 180° to align the UCIe interfaces. A single matrix unit achieves 90% utilization with a peak energy efficiency of 2061 GOPS/W while saturating the vector backend functional units. To demonstrate chiplet interconnect capability, a UCIe-compatible PHY enables 256 Gbps bidirectional inter-chiplet communication; in addition, an open-source, TileLink-based, digital Chip-to-Chip interface supports low-bandwidth communication.**

***Index Terms*—SoC, Vector Processor, Matrix Extension, Chiplet, RISC-V, Machine Learning**

I. Introduction

Modern computing workloads increasingly demand flexible, high-utilization platforms capable of handling dense linear algebra with mixed vector and matrix operations. To meet this demand, outer-product-based matrix engines have emerged as efficient accelerators [6], [8], [11]. These engines decompose matrix multiplication into a sequence of rank-1 updates that are accumulated into a destination matrix C. Simultaneously, chiplet-based architectures have become essential for scaling performance beyond the limits of monolithic SoCs, particularly for partitioned inference systems that require high interconnect bandwidth. KODIAK is a prototype chiplet-based system that integrates a high-throughput outer-product matrix unit as a tightly coupled functional unit within a RISC-V vector processor. This integration enables efficient operand sharing via the vector register file (VRF), leverages the existing load-store unit (LSU), ensures high vector reuse, and maintains seamless compatibility with the RISC-V software ecosystem.

II. System Architecture

Fig. 1 shows the system architecture of a single KODIAK chiplet. A chiplet contains one in-order RV64GC RISC-V core and two vector-core-based compute tiles. Each compute tile consists of one dual-issue, superscalar, in-order RV64GCBV_ZFH_ZVFH RISC-V core with a tightly coupled RVV 1.0-compliant vector coprocessor (VLEN = 512 bits, DLEN = 256 bits; DLEN denotes datapath). An outer product matrix unit is directly integrated into the vector backend as a specialized functional unit. The compute tiles contain a globally addressable 128 KB of high-bandwidth local SRAM (termed the TCM) capable of serving the vector backend through the vector registers. The memory subsystem consists of a 512 KB LLC, a 64 KB scratchpad for firmware, and a 128 KB cacheable scratchpad for applications. A 256-bit TileLink (TL) ring-topology NoC facilitates communication between all digital blocks and on-chip SRAMs. To simplify physical design, the NoC separates coherent and non-coherent channels into distinct physical paths. Off-chip communication is performed over the low-bandwidth Chip-to-Chip Link, or the high-bandwidth 256 Gbps UCIe module.

III. Outer Product Unit

The Outer-Product Unit (OPU) is a custom matrix unit integrated into the vector backend of *Saturn* [1] as a specialized functional unit. Data movement occurs via the vector register file (VRF). Vector operands from the VRF are sent to the array, and resulting matrix rows are written from the matrix register file (MRF) to the VRF. The OPU is composed of: (1) a control sequencer, (2) the matrix register file, and (3) the processing array. The implementation operates on signed integers (input vectors are SEW=8 and the output rows are SEW=32 bits) supporting the four instructions shown in Table I.

Sequencer: Control logic is implemented within a specialized execution sequencer within the *Saturn* backend. The sequencer manages data transfer between the VRF and MRF, orchestrates control signals for OPU instructions, and features a simplistic scoreboard to support hazard checking. Through the sequencer, the OPU can stall hazardous instructions or back-pressure dispatch from issue queues.

TABLE I
OPU Instructions

Instruction	Description
`VMV_RV(md, rs1, vs2)`	Vector register move to matrix register row
`VMV_VR(vd, rs1, ms2)`	Matrix register row move to vector register
`BCAST(md, vs2)`	Broadcast vector register column-wise across matrix register
`VOPACC(md, vs2, vs1)`	Compute outer product and accumulate in matrix register

979-8-3195-0631-3/26/$31.00 ©2026 IEEE

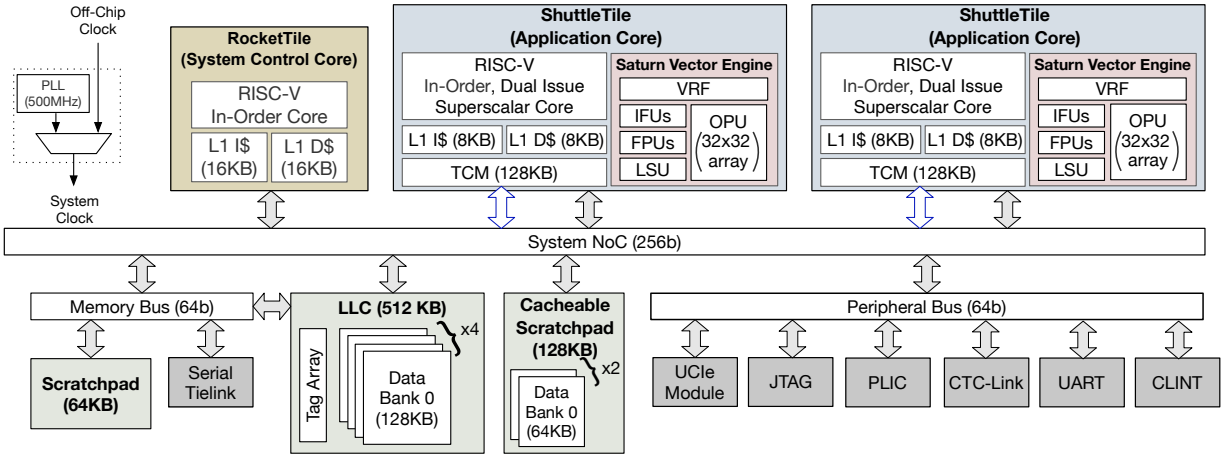


Fig. 1. Block diagram of the KODIAK system architecture, including the memory system and key peripherals; bold labels denote major blocks.

Matrix Register File: The matrix register file is an architectural state exclusively accessible to the OPU exposed in the extended ISA. Each register is two-dimensional (VLEN-by-VLEN) and is row-accessible at the DLEN granularity via systolic readout. The MRF resides in the OPU, is physically distributed, and is inaccessible to any other functional unit or memory subsystem.

Processing Element Array: Logically, the processing element array is 64×64. However, for area efficiency and to match DLEN, it is implemented as a single quadrant or a 32×32 array. Vector operands are broadcast both vertically and horizontally, allowing for a single-cycle operation. The architecture follows a three-level hierarchy: processing element, cluster, and full array. This hierarchy provides efficiency for readout and simplicity for the physical design of the array.

Processing Element: A processing element (PE) is the lowest level of the hierarchy. Fig. 2 shows the microarchitecture of a single PE. It contains a signed 8-bit multiplier, a signed 32-bit full-adder, and two sets of register banks. A register bank has four registers, and each register in the bank holds one value per quadrant (as a consequence of the microarchitectural ratio VLEN/DLEN=2 and the quadratic nature of scaling).

Cluster: Clustering is critical to the overall efficiency of MRF accesses. The OPU architecture is directly influenced by the physical design; clustering is performed at both the RTL and physical design levels. A cluster is a square sub-array of sixteen PEs. The column ordering does not follow the intuitive, logical pattern (i.e., every four consecutive columns). Columns are grouped by modulo-4, due to bit growth, which allows only one column per cluster to be read at a given time.

Array: The array is an 8×8 square grid of clusters. The control signals and LHS vector are broadcast from the left, and the RHS vector elements are broadcast from the top. Rows are read systolically in four beats from the bottom of the array in DLEN segments. Only row readouts are supported to reduce logical complexity, given that operand order can be switched and the vector backend is capable of supporting transpositions.

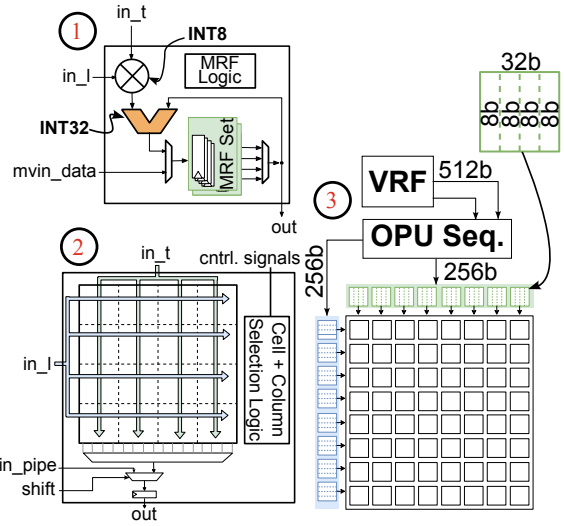


Fig. 2. Three-level hierarchical RTL implementation of the outer product unit inside the vector co-processor: (1) Processing Element, (2) Cluster, and (3) Array

Physical Design: The OPU physical design uses a three-level hierarchy: cluster, processing element array, and the outer-product unit. Control signals are interspersed with data lines to minimize routing, while inputs are broadcast on higher metal layers and routed down through vias to the PEs. As shown in Fig. 4, the array is laid out in quadrants with spacing for global clock and reset routing. Abutted clusters and cell-level clock gating enable scalability.

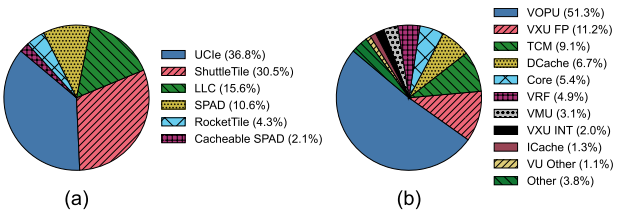


Fig. 3. Area breakdowns of (a) the full SoC and (b) one ShuttleTile instance.

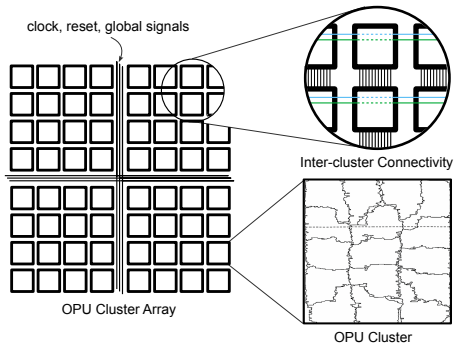


Fig. 4. Conceptual diagram showing the physical design approach to the OPU. Quadrant-based layout with channels for global signals, an outline of cluster place-and-route exhibiting the flat floorplanning of cells, and the intercluster connections for systolic readout (vertical black lines) and control signals (horizontal colored lines) buffered through each cluster.

IV. Inter-Chiplet Communication Links

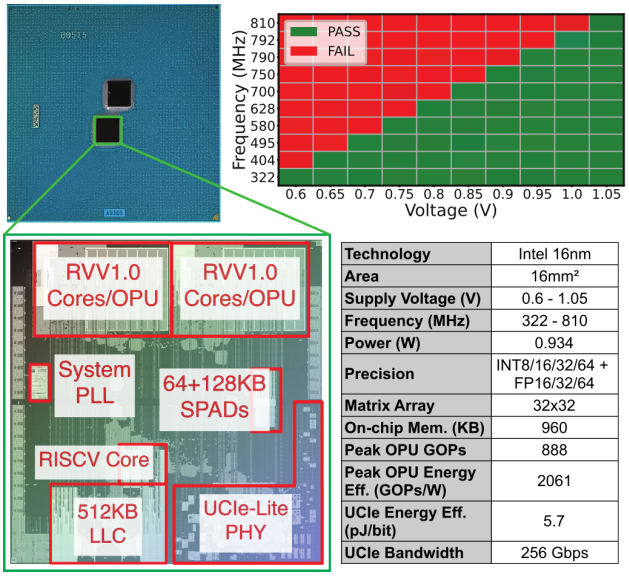
Chip-to-Chip: KODIAK includes a Chip-to-Chip (CTC) translator and a digital transceiver for low-speed inter-chiplet communication of data and control signals. The implemented link is an 8-bit-wide digital PHY with a ready-valid protocol and a source-synchronous forwarded clock. Packets have a minimum size of 32 bits and use a protocol-agnostic format, while the link operates in a non-coherent, single-in-flight transaction mode.

UCIe Module: KODIAK also features a UCIe-compatible standard package PHY that supports bidirectional data transmission at 8 Gbps per lane. Utilizing 16 lanes in each direction over an organic interposer, the interface achieves a shoreline bandwidth of 0.232 Tbps/mm. The PHY employs a clock-forwarded architecture, incorporating a per-lane delay-locked loop (DLL) and phase interpolator (PI) for precise deskewing. Data exchange between the PHY and the digital datapath is managed by 32:1 binary-tree serializers and deserializers. Operating at full capacity (8 Gbps $\times$ 16), the PHY consumes a total of 1.371 W across both chiplets.

V. Evaluation

KODIAK was evaluated on microbenchmarks and realistic applications. Fig. 6 shows the experimental setup. The FPGA independently controls both dies via an FMC connection. The DDR3 banks (512 MB) on the evaluation board serve as backing memory for the chiplets. The FPGA-to-SoC link is a 4-bit serialized variant of the TileLink protocol running at 25 MHz. This limits potential performance gains, as memory accesses can cost up to 40 SoC cycles. Performance was measured in clock cycles.

GEMM chained with Trigonometry: A representative use case is chaining a GEMM with transcendental function evaluation. The OPU performs the GEMM and retains intermediates in the MRF, avoiding costly round-trips through the VRF that increase register pressure, while the vector functional units compute the transcendental. This yields a 4.39$\times$ speedup over a VFU-only execution path.



Technology	Intel 16nm
Area	16mm²
Supply Voltage (V)	0.6 - 1.05
Frequency (MHz)	322 - 810
Power (W)	0.934
Precision	INT8/16/32/64 + FP16/32/64
Matrix Array	32x32
On-chip Mem. (KB)	960
Peak OPU GOPs	888
Peak OPU Energy Eff. (GOPs/W)	2061
UCIe Energy Eff. (pJ/bit)	5.7
UCIe Bandwidth	256 Gbps

Fig. 5. Dual-chiplet package, labeled die photo, single-chiplet specifications, and GEMM shmoo plot measured on the OPU from TCM.

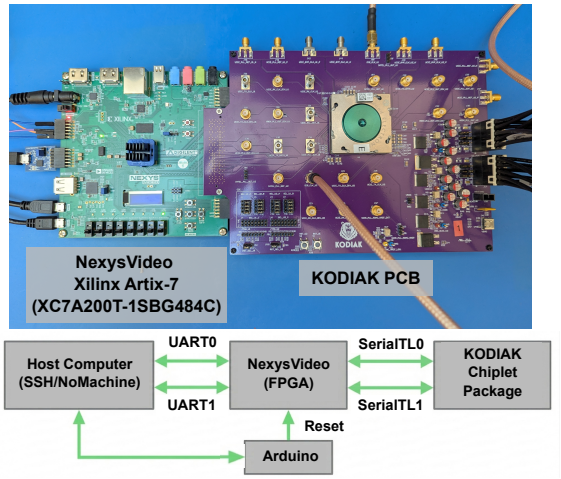


Fig. 6. Benchtop chip bring-up setup with a custom KODIAK PCB connected to an Artix-7 FPGA evaluation board via FMC to control both chiplets.

Feedforward Network (FNN): As an example application, a fully connected layer is run on a single chiplet using the Zephyr RTOS and ExecuTorch. Fig. 7 shows speedups for a range of input and output channel sizes, achieving a maximum of 22.91$\times$ improvement over the scalar baseline without memory-system optimizations. This approach can extend to transformer self-attention blocks.

Convolutional Neural Networks (CNN): CNNs are more memory-efficient than transformers and remain relevant evaluation targets. A 64×64×3×3 ResNet50 convolution layer with a 64×56×56 CHW input is evaluated as a bare-metal application using two variants: vector-only and vector + OPU. Compared to the vector-only implementation, the vector + OPU implementation achieves a 1.69$\times$ speedup while consuming only 81.5% of the power.

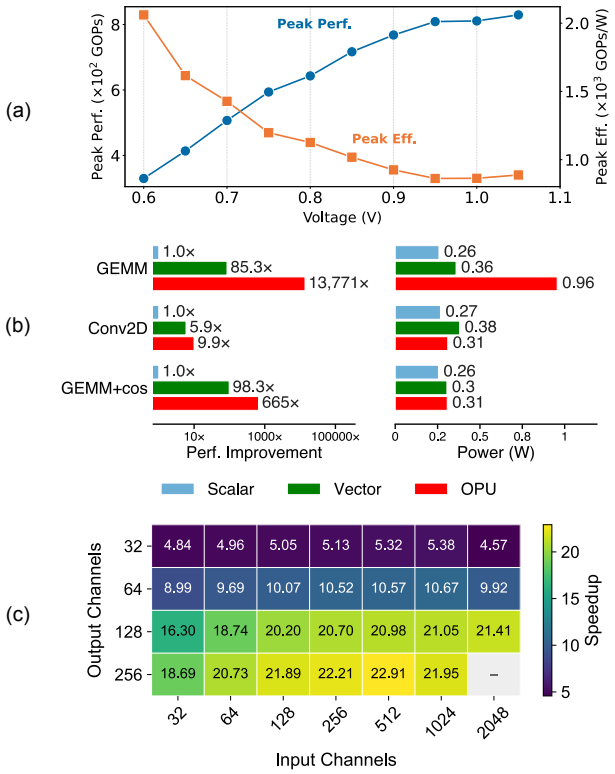


Fig. 7. Measured performance: (a) peak performance and energy efficiency across supply voltages, (b) performance improvement (log scale) and power measurements across benchmarks, (c) FNN speedup relative to scalar execution across input and output channel sizes.

Table II compares KODIAK with recent fabricated vector processors and matrix accelerators. KODIAK achieves the highest INT8 peak performance and peak energy efficiency among the listed works. It achieves 4.98× higher peak performance and 3.24× higher peak energy efficiency than the next-best chip, while using one quarter as many cores.

VI. Conclusion

KODIAK was fabricated in the Intel16 FinFET process to demonstrate the scalability of chiplet systems while optimizing for flexible compute capabilities. Two chiplets are co-packaged on an organic 2D substrate. The SoC can operate at a maximum frequency of 810 MHz with a 1.05 V supply voltage. The chiplet achieves a peak performance of 888 GOPS with a corresponding peak energy efficiency of 2061 GOPS/W. It consumes 0.311 W at idle with the highest-performance voltage and frequency pairing, and 0.767 W without concurrent off-chip communication. Relative to a vector baseline, GEMM execution on the OPU achieves a speedup of 161.4×.

Acknowledgement

This work was supported in part by the Microelectronics Commons Program, a DoW initiative, and Intel silicon and packaging donation through USP and HIP. We also acknowledge funding from NSF POSE 2303735. ChatGPT 5.4 was used for plotting-script assistance and manuscript concision.

TABLE II
Comparison to similar works with fabricated chips

	[9]	[10]	[11]	This work
Technology	40nm	16nm	Intel16	Intel16
Area (mm2)	65.6	10.15	16	16
Supply (V)	0.9	0.45-0.85	0.6-1.0	0.6-1.05
Max Clk Freq	—	60-450 MHz	1.01 GHz	810 MHz
Total Cores	1	4xDense MM 4xSparse MM	8 (4Big,4Little)	2
VLEN (bits)	—	—	512, 128	512
Matrix Unit	Systolic Array 16x16	Outer-Product 16x16x4	—	Outer-Product 32x32
ISA	Custom	Custom	RV64GCV RVV1.0	RV64GCV, RVV1.0 w/ Custom OPU
On-chip Memory	>12MB	1320 KB	512 KB (LLC) 256 KB (TCM)	512 KB (LLC) 256 KB (TCM)
INT8 Peak En. Eff. (GOPS/W)	0.43-0.50 (Posit8)	15.20-40.30	414	2061
INT8 Peak Perf. (GOPS)	0.0256-0.0512 (Posit8)	0.81-2.15	274	888
FP32 Peak En. Eff. (GFLOPS/W)	—	—	109	1.34
FP32 Peak Perf. (GFLOPS)	—	—	69.60	3.33

References

[1] J. Zhao, D. Grubb, M. Rusch, T. Wei, K. Anderson, B. Nikolic, and K. Asanovic, "Instruction scheduling in the saturn vector unit," 2024. [Online]. Available: https://arxiv.org/abs/2412.00997

[2] UC Berkeley Architecture Research, "Shuttle: A Rocket-based RISC-V superscalar in-order core," https://github.com/ucb-bar/shuttle, 2024.

[3] UCIe Consortium, "Universal Chiplet Interconnect Express (UCIe) specification 2.0," https://www.uciexpress.org/specifications, 2024.

[4] A. Amid, D. Biancolin, A. Gonzalez, D. Gruber, S. Karandikar, H. Liew, A. Magyar, H. Mao, A. Ou, N. Pemberton, P. Rigge, C. Schmidt, J. Wright, J. Zhao, Y. S. Shao, K. Asanović, and B. Nikolić, "Chipyard: Integrated design, simulation, and implementation framework for custom SoCs," in *IEEE Micro*, 2020.

[5] H. Liew, D. Grubb, J. Wright, C. Schmidt, N. Krzysztofowicz, A. Izraelevitz, E. Wang, K. Asanović, J. Bachrach, and B. Nikolić, "Hammer: A modular and reusable physical design flow tool," in *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*, San Francisco, CA, USA, 2022.

[6] N. Stephens, S. Biles, M. Boettcher, J. Eapen, M. Eyole, G. Gabrielli, M. Horsnell, G. Magklis, A. Martinez, N. Premillieu, A. Reid, A. Rico, and P. Walker, "The ARM Scalable Vector Extension," *IEEE Micro*, vol. 37, no. 2, pp. 26–39, 2017.

[7] C. Schmidt, J. Wright, Z. Wang, E. Chang, A. Ou, W. Bae, S. Huang, A. Flynn, B. Richards, K. Asanović, E. Alon, and B. Nikolić, "An eight-core 1.44GHz RISC-V vector machine in 16nm FinFET," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2021, pp. 58–60.

[8] D.-h. Park, S. Pal, S. Prabhakar, M. Thottethodi, T. N. Vijaykumar, M. Beaumont, and R. Dreslinski, "A 7.3 m output non-zeros/j, 11.7 m output non-zeros/gb reconfigurable SpMM accelerator," *IEEE Journal of Solid-State Circuits (JSSC)*, 2020.

[9] J. Liu, Q. Li, Y. Luo, H. Zhang, J. Lu, S. Fan, J. Ye, Y. Liu, X. Liu, Y. Yang, Z. Ye, Y. Zeng, A. Shen, R. Huang, W. Cong, X. Zou, and M. Gao, "Titan-i: An open-source, high performance risc-v vector core," in *Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 675–690. [Online]. Available: https://doi.org/10.1145/3725843.3756059

[10] K. Prabhu, R. M. Radway, J. Yu, K. Bartolone, M. Giordano, F. Pedding-haus, Y. Urman, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, S. Mitra, and P. Raina, "Minotaur: A posit-based 0.42–0.50-tops/w edge transformer inference and training accelerator," *IEEE Journal of Solid-State Circuits*, vol. 60, no. 4, pp. 1311–1323, 2025.

[11] S. Moon, M. Li, G. K. Chen, P. C. Knag, R. K. Krishnamurthy, and M. Seok, "T-REX: A 68-to-567μs/token 0.41-to-3.95μJ/token transformer accelerator with reduced external memory access and enhanced hardware utilization in 16nm FinFET," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, 2025, pp. 406–408.

[12] V. Jain, D. Grubb, J. Zhao, K. Anderson, K. T. Ho, Y. Chi, E. Schwarz, K. Asanović, Y. S. Shao, and B. Nikolić, "Cygnus: A 1 ghz heterogeneous octa-core risc-v vector processor for dsp," in *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2025, pp. 1–3.