DECADES: A 67mm^2, 1.46TOPS, 55 Giga Cache-Coherent 64-bit RISC-V Instructions per second, Heterogeneous Manycore SoC with 109 Tiles including Accelerators, Intelligent Storage, and eFPGA in 12nm FinFET

Fei Gao[1], Ting-Jung Chang[1], Ang Li[1], Marcelo Orenes-Vera[1], Davide Giri[2]*, Paul J. Jackson[1], August Ning[1], Georgios Tziantzioulis[1], Joseph Zuckerman[2], Jinzheng Tu[1], Kaifeng Xu[1], Grigory Chirkov[1], Gabriele Tombesi[2], Jonathan Balkind[3], Margaret Martonosi[1], Luca Carloni[2], David Wentzlaff[1]

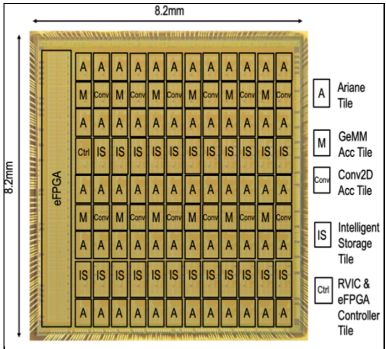
[1]Princeton University, [2]Columbia University, [3]University of California, Santa Barbara

As Moore's Law is coming to an end, heterogeneous SoCs have become ubiquitous, improving performance and efficiency with specialized hardware. However, the addition of hardware accelerators makes data supply more challenging. Feeding data to accelerators becomes a bottleneck, especially for data-intensive workloads such as graph analytics, sparse linear algebra, and machine learning applications. DECADES addresses this issue with a combination of accelerators, embedded FPGA (eFPGA), and its unique "intelligent storage" (IS) tile. DECADES is one of the largest chips ever built in academia and has the highest core count of cache-coherent, OS-capable, 64-bit RISC-V processors.

The SoC was manufactured in GlobalFoundries 12nm FinFET, measuring 8.2mmx8.2mm. It consists of 108 heterogeneous tiles arranged in a 12x9 2D mesh topology, connected by three full-duplex 64-bit wide Network-on-Chip (NoC) channels, and an eFPGA tile. Tiles in the mesh have a similar structure and can be divided into a logic Core and a Socket (Fig. 1). The Socket contains three NoC routers (one for each channel), an L2 cache, and a slice of a shared Last-Level-Cache (LLC). The manycore cache hierarchy is built based on BYOC[1]. The L2 cache is an 8KB write-back private cache, and the distributed and shared LLC is 64KB per tile, totaling 6.9MB on chip. A directory-based MESI cache-coherence protocol is maintained in the cache hierarchy, which supports RISC-V atomic operations. 60 cache-coherent Ariane tiles are the host processors. Each implements a 6-stage, in-order, single-issue, OS-capable CPU, supporting the RV64GC ISA. It includes an IEEE-compliant Floating-Point Unit (FPU), a 16KB L1 instruction cache, and an 8KB write-through L1 data cache. Each Ariane core is connected to the L2 cache in the Socket. For other tiles, like accelerators, the Core may directly talk to the NoC interface. Although bypassing the L2 cache, they can still access the LLC coherently.

Inspired and informed by our group's research on GraphAttack[2] and additional compiler and simulator research, DECADES is designed to provide both hardware specialization and efficient data supply. Specialization is achieved with accelerators and the eFPGA. The DECADES chip embeds two types of fixed-function hardware accelerators (12 tiles each) -- general matrix multiply (GeMM) and 2D convolution (Conv2D). Both accelerators are designed using high-level synthesis, leveraging the accelerator design methodology from Embedded Scalable Platforms (ESP)[3]. The logic for loading data into the accelerator, computing the results, and writing the data back to main memory are decoupled, thereby allowing the three processes to operate in a pipelined manner over datasets that do not fit into the local scratchpad of the accelerator. DECADES also contains an eFPGA which allows users to generate "soft" accelerators post-fabrication. The eFPGA has 7040 multi-functional, 6-input LUTs, 32 40-bit hard multipliers, and 32 16Kbit, dual-port, block RAMs. It is synthesized from an RTL description and generated by PRGA[4] down to standard cells using off-the-shelf digital EDA tools. The eFPGA is tightly integrated with the rest of the system and fully coherent with the cache system.

Efficient data supply is achieved through two different approaches with 23 IS tiles. The first one is by increasing on-chip storage. Besides the cache hierarchy, each accelerator tile has a 48KB scratchpad for local access, and each IS tile (Fig. 2) contains a 64KB global accessible scratchpad memory, providing a total of 2.6MB extra on-chip SRAM. The IS core is capable of prefetching data into the scratchpad for accelerators. Additionally, the scratchpad serves as a communication buffer when accelerators are activated together and working in a pipelined fashion. The second approach is using prefetching and decoupling of data supply and compute to mitigate memory latency. The Memory Access Parallel-Load Engine (MAPLE)[5] in IS tile can fetch loops of indirect memory accesses (A[B[i]]) in hardware. Each MAPLE unit can fetch up to 128 indirect memory requests in parallel. The fetched data is placed into MAPLE's queues, which are configured at runtime as circular FIFOs.



Die micrograph.

Each IS tile also contains a near-memory compute engine Nibbler, which is composed of 64 SIMD lanes capable of running arithmetic 32-bit RISC-V Vector instructions and a scalar execution lane. It breaks down each instruction into multiple subword(8 bits)-serial micro-operations and processes a 32-bit input with 4 cycles. Nibbler is fully programmable, which provides support for efficient, arbitrary, computation on data stored in the IS core scratchpad.

Fig. 3 shows the maximum frequency under different supply voltages. A single Ariane tile reaches the max frequency of 1.25GHz under 1.2V, ambient temperature. Due to IR drop, the tiles in the middle of the chip can not run as fast as those in the corners; thus, the max frequency of all 60 Ariane tiles simultaneously running is 911MHz. The eFPGA runs under a different user clock domain, and its max frequency depends on the user application. We evaluate two user applications: a pseudo-random number generator using a 64-bit Linear-Feedback Shift Register (LFSR), and an INT8-precision, complex, 64-point FFT. LFSR uses 4% of LUTs and runs up to 224MHz; FFT uses 36% of LUTs and runs up to 96MHz.

The abundant on-chip resources provide a significant amount of parallelism. The 60 Ariane tiles can issue 55 giga cache-coherent RV64 instructions per second (GCCRV64IS). **It is the highest coherent RV64 throughput on a single chip to date.** Together with 1495 Nibbler lanes and 1172 hard multipliers/adders in both the accelerators and the eFPGA, **the entire SoC achieves a peak performance of 1.46TOPS**. Fig. 6 compares DECADES with other state-of-the-art designs. We also illustrate how different components on DECADES provide efficient computation and data supply with real-world applications. Fig. 4 shows the speedup of the hardware accelerators compared with the software implementation running on Ariane. We run inference tasks which utilize both accelerators with two neural networks LeNet and Dwarf6. It shows up to 25x speedup and 133x energy saving. Fig. 5 demonstrates the benefit achieved from three programmable acceleration units: the eFPGA, MAPLE and Nibbler. Running 64-point FFT, the eFPGA is 15x faster and 7x efficient than a CPU implementation. We characterize MAPLE with a sparse linear algebra task and compare it with DOALL parallelization using the same amount of Ariane cores. Prefetching mode scales better and achieves up to 7.4x speedup compared to the 2-core baseline. Nibbler is energy efficient in running RISCV vector instructions, and its EPI is 3.6-6.2pJ under 0.7V supply voltage.

Acknowledgements: This material is based on research sponsored by the Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement No. FA8650-18-2-7862§.

References: [1] J. Balkind *et al*., ASPLOS, Mar. 2020. [2] A. Manocha *et al*., TACO, 18(4), 2021. [3] P. Mantovani *et al*., ICCAD, Nov. 2020. [4] A. Li *et al*., FPGA, Feb./Mar. 2021. [5] M. Orenes-Vera *et al*., ISCA, Jun. 2022. [6] A. Gonzalez *et al*., ESSCIRC, Sep. 2021. [7] C. Schmidt *et al*., ISSCC, Feb. 2021. [8] D. Rossi *et al*., ISSCC, Feb. 2021. [9] A. Rovinski *et al*., VLSI, Jun. 2019.

* This paper is dedicated to the memory of Davide Giri.

§ The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) or the U.S. Government.

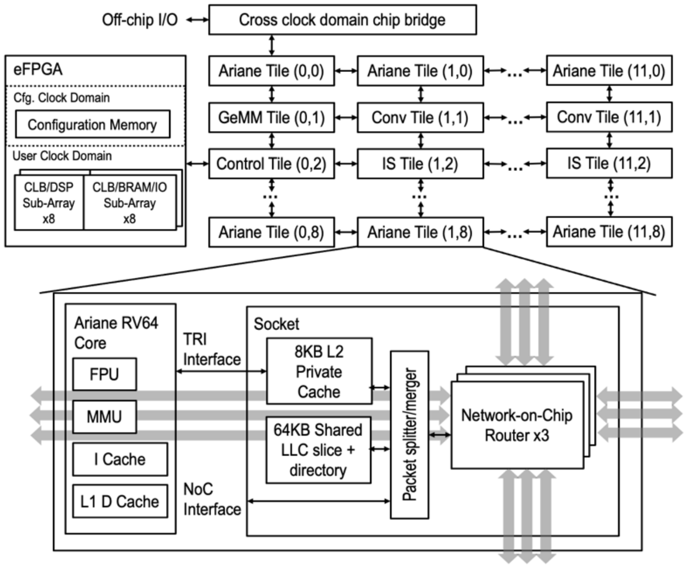


Fig. 1. Architecture block diagram of the SoC and an individual Ariane Tile

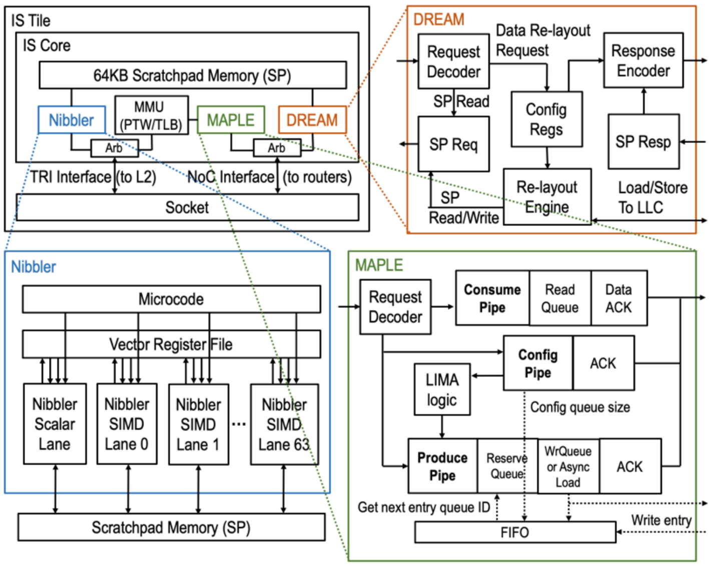


Fig. 2. Architecture block diagram of the Intelligent Storage (IS) Tile and its components

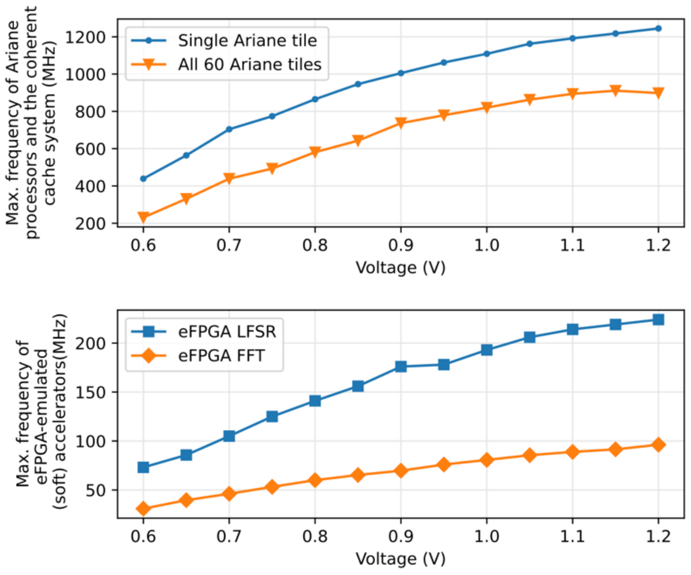
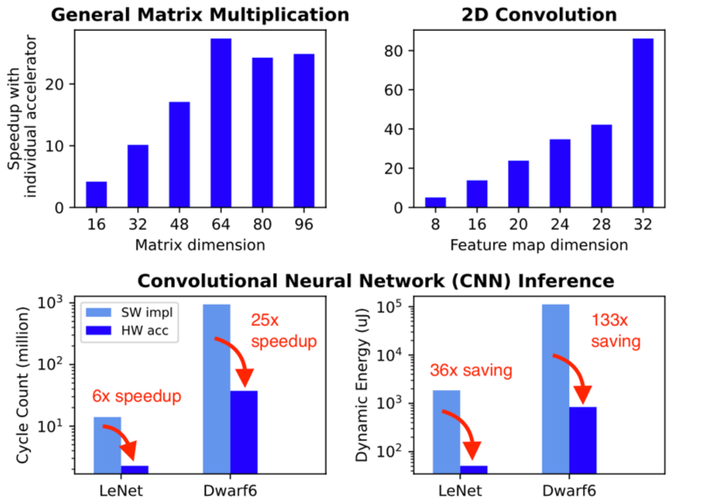
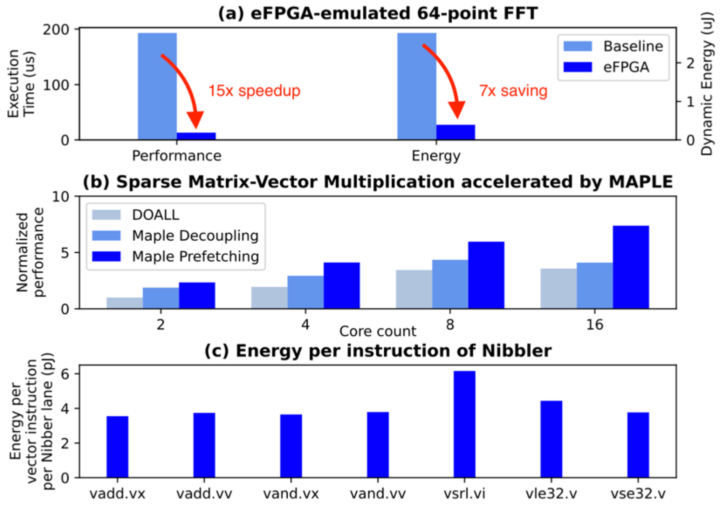


Fig. 3. Maximum operating frequency vs. supply voltage



Notes: We run with two networks. The 5-layer LeNet performs digit recognition on the MNIST dataset, and the 6-layer Dwarf6 performs image classification of the CIFAR-10 dataset. Chip measured at 0.9V 800MHz

Fig. 4. Performance of individual GeMM and Conv2D accelerators and the evaluation with CNN inference tasks using both of them



Notes: (a) The soft accelerator in the eFPGA is measured at 0.9V 70MHz. The baseline is a software implementation in an Ariane core, measured at 0.9V 700MHz.
(b) Chip measured at 0.9V 700MHz.
(c) Nibbler measured at 0.7V 300MHz.

Fig. 5. Evaluation of programmable acceleration units

	This work	ESSCIRC '21 [6]	ISSCC '21 [7]	ISSCC '21 [8]	VLSI '19 [9]
Technology	12nmFinFET	Intel 22FFL	16nm FinFET	22nm FDSOI	16nm FinFET
Die Area (mm^2)	**67**	16	24.01	12	15.25
Transistor count	**2.2 Billion**	-	125 Million	-	-
LLC (MB)	**6.9**	1	3	1.56	3.9
ISA	RV64GC+RV32IV	RV64GC	RV64GC	RV32I	RV32IM
Cache coherent	**Yes**	No	No	No	No
Total RV64 Cores	**60**	2	9	-	-
Total RV32 cores	**1495** PNM Lanes **3.6-6.2** pJ/inst.	-	-	10	496
Accelerators	**12 GeMM,** **12 Conv2D**	DNN, Vector	Vector	2 ML acc	-
eFPGA LUTs	7040	-	-	-	-
Intelligent Storage	23 Tiles 1.5MB scratchpad	-	-	-	-
Voltage	0.6V-1.2V	0.75V-0.9V	0.5V-1V	0.6V-0.8V	0.6V-0.98V
Fmax (MHz)	1245 for single core 911 for 60 cores	961 for ML 210 for GP	1440	450	1400
Instruction throughput	**55GCCRV64IS[1] +** **5.2GRVV32IS[2]**	-	-	-	695GRV32IS
Operation throughput	**1.46TOPS**	-	368GHFLOPS	2GFLOPS	-
CoreMark/MHz	169 (Only Ariane)	2.37	-	-	580

[1] Giga cache-coherent 64-bit RISC-V instructions per second
[2] Giga 32-bit RISC-V vector instructions per second

Fig. 6. Comparison to the state-of-the-art