DECADES: A 67mm², 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 Gao1, Ting-Jung Chang1, Ang Li1, Marcelo Orenes-Vera1, Davide Giri1,*, Paul J. Jackson1, August Ning1, Georgios Tziantzioulis1, Joseph Zuckerman2, Jinzheng Tu1, Kaifeng Xu1, Grigory Chirkov1, Gabriele Tombesi1, Jonathan Balkind1, Margaret Martinsoni1, Luca Carloni2, David Wentzlaff1

1Princeton University, 2Columbia University, 3University 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 on-chip shared LLC is 64KB per tile, totalling 8.5MB 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 RUV64GC 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 16bit, dual-port, block RAMs. It is synthesized from an RTL description and generated 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[ Bij][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. 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 2244MHz; FFT uses 36% of LUTs and runs up to 966MHz.

The abundant on-chip resources provide a significant amount of parallelism. The 60 Ariane tiles can issue 512 cache-coherent requests per second (CGRV64AS). 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.

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.


* 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.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on September 25,2023 at 11:58:33 UTC from IEEE Xplore. Restrictions apply.
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

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

Fig. 5. Evaluation of programmable acceleration units

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