The REBECCA Hardware/Software Edge AI platform

Iakovos Mavroidis[1], Darshak Sheladiya[2], Ioannis Papaefstathiou[3], Konstantinos Georgopoulos[1], Pavlos Malakonakis[3], Pablo Ghiglino[4]

[1]Technical University of Crete
[2] SYSGO GmbH
[3] Exascale Performance Systems, EXAPSYS
[4] Klepsydra Technologies AG

Abstract

The REBECCA project is pioneering advancements in edge AI systems using RISC-V technology, emphasizing power efficiency, scalability, and open-source accessibility. It integrates a multicore RISC-V-based architecture with AI-specific accelerators, neuromorphic computing, and security features to deliver a high-performance, cost-effective AI platform. The core of REBECCA is the CVA6 processor, leveraging a chiplet-based design and shared memory architecture to optimize real-time AI processing. The platform incorporates HyperRAM for efficient data access and a custom software stack to maximize efficiency and security. Initial prototypes using U55C development boards and FPGA-based have been used to validate the feasibility of RISC-V for AI-driven applications. Moreover, the REBECCA CVA6 processor has been benchmarked against the Microchip PolarFire ICICLE, yielding promising performance results. Future research will enhance neuromorphic computing, AI framework integration, and real-time performance optimization. With strong industry and academic collaboration, REBECCA is shaping the future of AI at the edge, positioning RISC-V as a compelling alternative to proprietary AI solutions.

Introduction

The REBECCA (Reconfigurable Heterogeneous Highly Parallel Processing Platform for safe and secure AI) KDT-JU project is an ambitious effort to advance edge AI systems using RISC-V technology. With a focus on open, power-efficient, and scalable computing, REBECCA is developing a multicore RISC-V-based system that integrates AI-specific accelerators, neuromorphic computing, and security. The goal is to provide a cost-effective, high-performance AI platform that meets the increasing demands of edge computing. By leveraging an open-source architecture, REBECCA ensures that AI development is more accessible and not restricted by proprietary constraints.

At the heart of REBECCA is the CVA6 processor core, built on the RV64GC RISC-V ISA, which supports a Unix/Linux operating system. The system uses a chiplet-based architecture, allowing modular expansion based on application needs. A global shared address space facilitates direct chiplet-to-chiplet memory access, ensuring efficient data movement for real-time AI applications. The HyperRAM-based memory structure optimizes bandwidth, helping AI models run smoothly. A custom OS and software stack have been designed to fully utilize the underlying hardware, ensuring maximum efficiency, security, and performance.

In summary, REBECCA is redefining what is possible with RISC-V-based AI hardware. By combining custom AI accelerators, shared memory architectures, and a fully optimized software stack, the project is setting new standards for power-efficient and high-performance edge AI systems. With strong collaboration between research institutions and industry leaders, REBECCA is fostering a thriving AI hardware ecosystem, making RISC-V a compelling alternative to proprietary AI solutions. The project's focus on efficiency, security, and scalability ensures that it will play a crucial role in shaping the future of AI at the edge.

Our contribution highlights key aspects of the project, including initial lessons learned from both the hardware and software perspectives.

Methodology

One of REBECCA's key features is its hardware-accelerated AI processing. The platform includes Intrusion Detection Systems (IDS) for security and anomaly detection, Neuromorphic Processing Units (NMP) for AI-driven pattern recognition, and specialized ML and CNN accelerators that enhance deep learning performance at the edge. These custom AI accelerators speed up inference tasks while maintaining low power consumption—an essential factor for edge deployments where efficiency is critical.

The REBECCA architecture is composed of the REBECCA ASIC, which integrates a RISC-V subsystem, three AI accelerators, and a chip-to-chip block that connects the ASIC to an external FPGA.

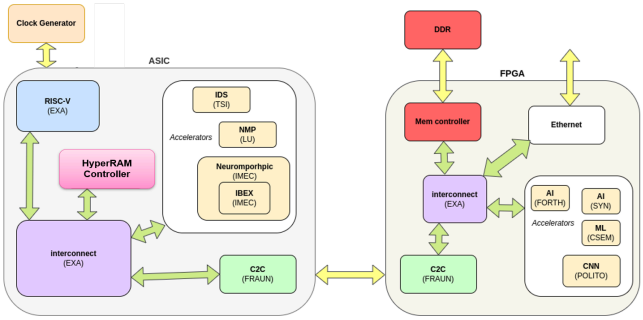

Figure 1. The REBECCA Architecture consisting of an ASIC and an accompanying FPGA.

The architecture supports a shared address space, allowing accelerators to directly access host memory. Additionally, the RISC-V NoC supports I/O coherency, enabling accelerators to access the processor's cache. This eliminates redundant memory transfers between host and accelerator memory, significantly enhancing performance and efficiency.

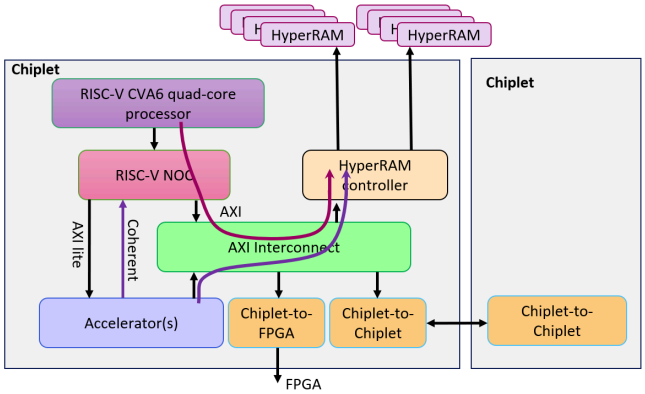

Figure 2. Shared address space between the host and the accelerators.

To ensure REBECCA's architecture is ready for real-world use, an initial emulation prototype has been built using U55C development boards. These boards are connected via 100-Gigabit Ethernet, providing a scalable testbed for evaluating AI performance and optimizing hardware-software integration. Additionally, the REBECCA platform has been tested on ALINX XKU15 FPGA-based platform, successfully booting Linux on a RISC-V system using HyperRAM as main memory, and accelerating tasks on the FPGA fabric. This marks a major milestone in demonstrating the feasibility of using RISC-V for AI-driven applications.

Another key innovation is the developed HyperRAM module, designed to improve memory management and data processing efficiency. It features two independent HyperRAM channels, each linked to four HyperRAM chips, supporting 512MB of total memory and allowing efficient data access for AI tasks. The module also includes an SD card and JTAG interface and an FMC connector in order to be connected to an FPGA-based emulator board, making it possible to fully emulate the REBECCA ASIC behavior and conduct extensive hardware testing. This improved memory subsystem plays a crucial role in enhancing overall system performance, allowing AI algorithms to handle large datasets with minimal delays. In addition to hardware innovation, REBECCA is also developing a rich RISC-V system software and application stack, including two hypervisors, Kubernetes orchestration and AI optimization and library support.

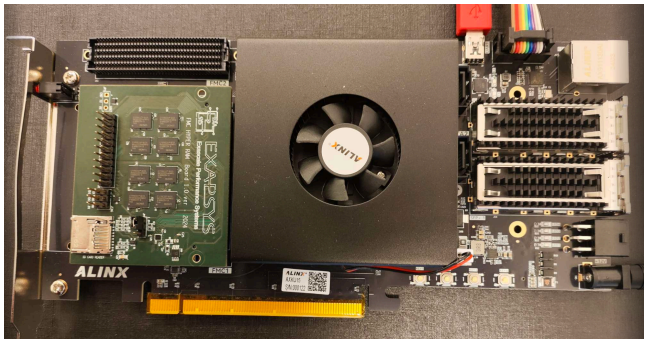

Figure 3. The REBECCA Emulation Environment

Looking ahead, REBECCA is pushing the boundaries of AI acceleration and secure edge computing. Future research will focus on expanding neuromorphic computing capabilities to improve AI learning and inference at the edge. There is also a strong emphasis on real-time AI performance optimization, ensuring low-latency AI execution with minimal energy use. The project aims to broaden software compatibility, integrating more AI frameworks and applications to make the system more versatile and widely adoptable. These advancements will position REBECCA as a leading-edge AI processing platform that is secure, high-performance, and highly scalable.

AI Performance Benchmarks

AI benchmarks were performed on the U55C soft-core using models from ESA's OBPMark-ML benchmarking framework, along with standard models such as AlexNet, MobileNetV1, and MobileNetV2. The performance results were then compared to those obtained from running the same models on the PolarFire ICICLE soft-core RISC-V. The comparative analysis shows highly promising outcomes, with the REBECCA CVA6 consistently delivering approximately 2× lower latency than the PolarFire.

Conclusion

In summary, REBECCA is redefining what is possible with RISC-V-based AI hardware. By combining custom AI accelerators, shared memory architectures, and a fully optimized software stack, the project is setting new

standards for power-efficient and high-performance edge AI systems. With strong collaboration between research institutions and industry leaders, REBECCA is fostering a thriving AI hardware ecosystem, making RISC-V a compelling alternative to proprietary AI solutions. The project's focus on efficiency, security, and scalability ensures that it will play a crucial role in shaping the future of AI at the edge.

Our contribution highlights key aspects of the project, including initial lessons learned from both the hardware and software perspectives.