Vuong Nguyen,
Elochukwu Ifediora,
Farhad Modaresi

Ztachip

A MULTICORE, DATA-AWARE, EMBEDDED RISC-V AI ACCELERATOR FOR EDGE INFERENCING

Introduction

- Computing Limitations: Moore's Law, Dennard's scaling, Amdahl's Law etc.
- Efforts to improve: VLIW, SIMD, OoO, heterogeneous, paralallism, pipelining
- New effort: DSA to keep up with changes in AI models and algorithms while maintaining a very high compute performance at low power for embedded device

Method

Decouple data-oriented operations from computing operations, using

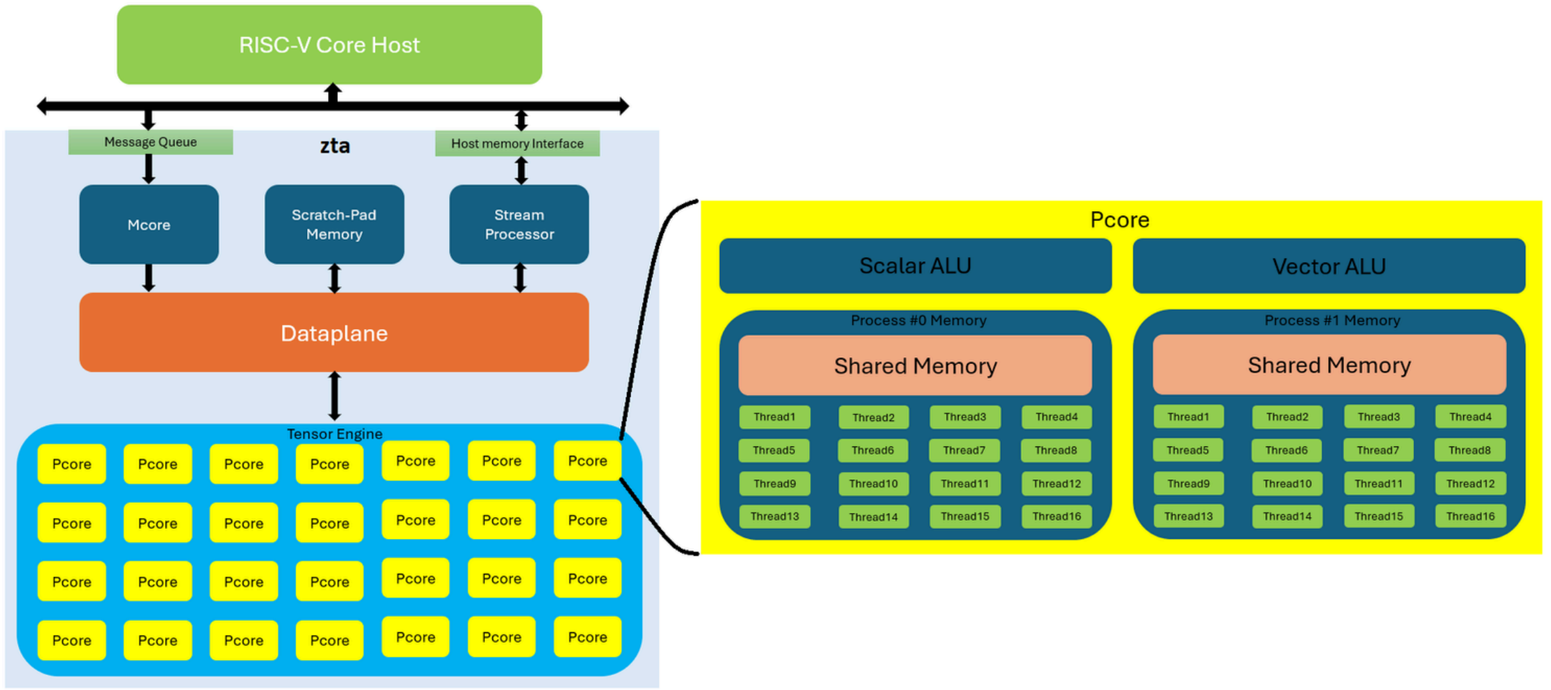
- An Mcore, a Scheduling Processor,
- A Dataplane, to stream the next data and instruction to the Tensor Engine .
- A Scratch-Pad Memory to temporarily hold data
- A Stream Processor to manage data IO
- Tensor Engine with 28x Pcores that can be configured to act like a systolic array to perform in memory compute each containing a Scalar and Vector ALU, with 16 Threads of execution on private memory.

Result

Mobinet v2.0 perf Ztachip > RPi 4

Table 2: Raspberry Pi 4 Vs Ztachip on Artix A7 FPGA.

Features	RaspberryPi4	Ztachip on A7 100T
Architecture	ARMv8	RISC-V
Core	4 Cortex A72	1 VexRiscv + Zta
Frequency	4 x 1.8 GHz	125MHz
Memory & Bandwidth	32-bit DDR4 @1600MHz	16-bit DDR3 @333MHz
Mobinetv2.0	483.5mS	220mS



It is evident that, by using a Domain Specific Architecture (DSA) to decouple data-oriented operations from computing operations, one can accelerate different AI workloads simultaneously at a fraction of the cost and power compared to an edge CPU and/or GPU

Discussion

The demonstration (in QR code below) features the Ztachip running four different AI workloads (Object Detection/Identification, Optical Flow/Motion Detection, Harris Corner and Edge Detection) concurrently at 30fps. With the given resource of Ztachip as a soft core running under 125MHz, it outperforms the Raspberry Pi 4 in Mobinetv2.0 and it is comparable to NV Jetson Nano

A software toolchain and SDK is also provided to support the portability to Python API, Micropython API and Arduino Library in an effort to democratize AI on edge devices.

The next step is to tape-out Ztachip as an ASIC on the open-source SKY130 or GF180 Process Development Kit (PDKs) on Efabless Inc. We expect to see a significant improvement in the performance of Ztachip in the benchmarks highlighted in this paper.

RISC-V Summit Europe, Munich, 24-28th June 2024