Linux capable RISC-V CPU for IOb-SoC

Pedro Nuno de Melo Antunes
pedronmantunes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

January 2022

Abstract

The objective of this work is to accelerate an object detection application using the VersatCNN, a Coarse Grained Reconfigurable Array (CGRA) architecture developed for efficient inference of Convolutional Neural Network (CNN). The performance objective is to enable real-time execution - 30 frames per second - on low-end embedded systems. The object detection application accelerated is the Tiny YOLOv3, based on a CNN developed for constrained environments. The Tiny YOLOv3 network is composed of Convolutional, Maxpool, Route, Upsample and Yolo layers. The application also includes pre and post-processing routines. This thesis work uses the IOb-SoC, a RISC-V based system on a chip (SoC) from IObundle as a baseline hardware platform for the project development. Peripheral modules for measuring performance and enabling ethernet communication are integrated into the IOb-SoC. The embedded version of Tiny YOLOv3 is profilled for RISC-V only execution to identify the parts that require acceleration. The VersatCNN is integraded into the system as a peripheral. The work mainly focuses on the design of CGRA dataflow configurations, as the different parts of the application require distinct configuration strategies for the multiple layer types and the pre and post-processing routines. The final system achieves over 30 FPS for the complete Tiny YOLOv3, with a clock frequency of 143 MHz and a 832x parallelism factor for convolutional operations.

Keywords: Object Detection, Convolutional Neural Networks, Systems on Chip, Coarse Grained Reconfigurable Arrays.

The baseline hardware system used for this work is IOb-SoC [**?**], an open-source RISC-V SoC platform developed by IObundle. Figure **??** presents the block diagram of the system. A RISC-V soft CPU core controls the system.

The CPU used is the PicoRV32 RISC-V core [**?**], minimally modified to be integrated into IOb-SoC. The PicoRV32 CPU is designed to use minimal hardware resources and, as a consequence, has very low performance, taking 4 Cycles Per Instruction (CPI). The CPU can access the internal memory, the external DDR memory (via cache), and the four peripherals mentioned:

- the VersatCNN Accelerator Core is the main focus of this work and is used to accelerate the convolution operation.
- the Timer is used to measure performance.
- the UART is used for programming and basic user runtime messages.
- the Ethernet module provides a higher bandwidth communication facility used to transfer large datasets to IOb-SoC.

The author would like to thank ...

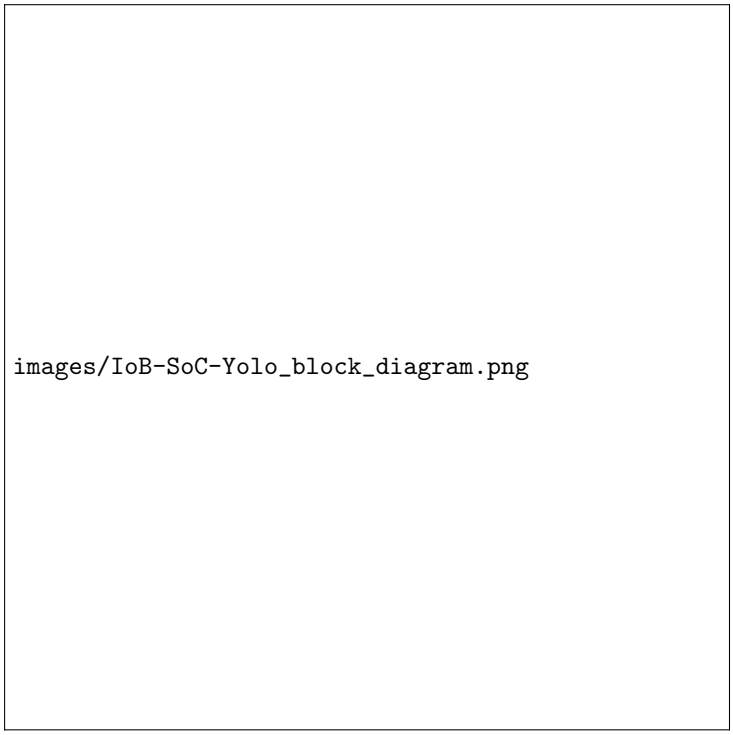

Figure 1: IOb-SoC-Yolo block diagram.