Ariane + NVDLA

Seamless Third-Party IP Integration with ESP

Davide Giri
Kuan-Lin Chiu
Guy Eichler
Paolo Mantovani
Nandhini Chandramoorthy (IBM Research)
Luca P. Carloni

CARRV 2020
Motivation

- SoCs are increasingly heterogeneous [1]
- Heterogeneity increases the engineering effort [2]
  → IP reuse enables the design of complex SoCs
- Thanks to open-source hardware (OSH) movement [3]
  → Proliferation of open-source IPs

Seamless third-party IP integration is key!

In this work

Enhance **ESP** with **support for third-party accelerators**

- ESP is our open-source platform for SoC design [4]

Demonstrate integration capabilities of **ESP**

- Rapid FPGA prototyping

Open-source release as part of **ESP**

- Hands-on tutorial: esp.cs.columbia.edu/docs/thirdparty_acc

ESP overview
ESP architecture
ESP methodology

Accelerator Flow
• Simplified design
• Automated integration

SoC Flow
• Mix&match floorplanning GUI
• Rapid FPGA prototyping
ESP methodology: SoC flow

Rapid Prototyping

SoC Integration

accelerator

third-party accelerator

NVDLA.org

Ariane

IP Library

** By lewing@isc.tamu.edu Larry Ewing and The GIMP
Third-party IP integration with ESP
ESP accelerator tile
ESP accelerator flow

ESP accelerator

Third-party accelerator

Accelerator skeleton

Accelerator specific functions

Test behavior

Generate RTL

Test RTL

Instantiate into SoC

Accelerator definition (xml)

RTL wrapper wiring

Makefile

Third-party RTL and SW files list

Accelerator definition (xml)

automated

manual

PyTorch

C

CHISEL

SystemC

OIPX

ESP accelerator flow
Ariane + NVDLA with ESP
Integration of Ariane

ESP processor tile
• RISC-V Ariane (new!) or Sparc-v8 Leon3
• Boot unmodified Linux
• AXI4 (new!) or AHB bus to access memory
• APB bus to access peripherals
• Optional L2 private cache
• Processor-specific interrupt controller placed in the I/O tile
NVDLA

NVIDIA Deep Learning Accelerator
• Open source
• Fixed function
• Highly configurable

NVDLA small
• 8-bit integer precision
• 64 MAC units
• 128 KB local memory
Evaluation: setup

SoCs evaluated on FPGA (Xilinx XCVU440)

- Ariane core
- 1-4 NVDLA tiles
- 1-4 memory channels

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Layers</th>
<th>Input</th>
<th>Model Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet</td>
<td>MNIST</td>
<td>9</td>
<td>1x28x28</td>
<td>1.7MB</td>
</tr>
<tr>
<td>Convnet</td>
<td>CIFAR10</td>
<td>13</td>
<td>3x32x32</td>
<td>572KB</td>
</tr>
<tr>
<td>SimpleNet</td>
<td>MNIST</td>
<td>44</td>
<td>1x28x28</td>
<td>21MB</td>
</tr>
<tr>
<td>ResNet50</td>
<td>ILSVRC2012</td>
<td>229</td>
<td>3x224x224</td>
<td>98MB</td>
</tr>
</tbody>
</table>
Evaluation: results

Performance of NVDLA small in ESP @ 50 MHz

- LeNet: 3.8 frames/second
- Convnet: 4.5 frames/second
- SimpleNet: 1.3 frames/second
- ResNet50: 0.4 frames/second

Scaling NVDLA instances and DDR channels @ 50 MHz

- 1 NVDLA: 1 frames/second
- 2 NVDLA: 2.1 frames/second
- 3 NVDLA: 3.1 frames/second
- 4 NVDLA: 3.9 frames/second

18x lower than NVIDIA’s results @ 1GHz
Performance preserved
Thank you from the ESP team!

Ariane + NVDLA
Seamless Third-Party IP Integration with ESP

Davide Giri
Kuan-lin Chiu
Guy Eichler
Paolo Mantovani
Nandhini Chandramoorthy (IBM)
Luca P. Carloni

CARRV 2020