Broadening the Exploration of the Accelerator Design Space in Embedded Scalable Platforms

Luca Piccolboni, Paolo Mantovani, Giuseppe Di Guglielmo, Luca Carloni
Columbia University, New York, NY, USA
Why Hardware Accelerators?

- High-performance embedded systems are heterogeneous:
  - they include multiple general-purpose processor cores
  - they include special-function hardware accelerators

---

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Embedded Scalable Platforms (ESP)

• To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: Embedded Scalable Platforms

Embedded Scalable Platforms (ESP)

• To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: Embedded Scalable Platforms

ESP instance for **WAMI (Wide-Area Motion Imagery)**

- Memory Controller
- Proc. core LEON3 CPU
- I/O Misc. Channels, etc.
- Accelerator WARP
- Accelerator GRAYSCALE
- Accelerator MATRIX-SUB
- Accelerator MATRIX-RES
- Accelerator SD-UPDATE
- Accelerator GRADIENT
- Accelerator MATRIX-ADD
- Accelerator CHANGE-DET
- Accelerator HESSIAN
- Accelerator DEBAYER
- Accelerator MATRIX-MUL
- Accelerator STEEP-DESC.
- Memory Controller
To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: Embedded Scalable Platforms

System-Level Design with **High-Level Synthesis (HLS)**

- Memory Controller
- Proc. core LEON3 CPU
- I/O Misc. Channels, etc.
- Accelerator WARP
- Accelerator GRAYSCALE
- Accelerator MATRIX-SUB
- Accelerator MATRIX-RES
- Accelerator SD-UPDATE
- Accelerator GRADIENT
- Accelerator MATRIX-ADD
- Accelerator CHANGE-DET
- Accelerator HESSIAN
- Accelerator DEBAYER
- Accelerator MATRIX-MUL
- Accelerator STEEP-DESC.
- Memory Controller
Embedded Scalable Platforms (ESP)

- To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity, we proposed: Embedded Scalable Platforms

  System-Level Design with High-Level Synthesis (HLS)

---

ESP instance for WAMI (Wide-Area Motion Imagery)
Hardware Accelerators with HLS

SystemC Specification

GRAYSCALE Interface

GRAYSCALE Logic

- load
- compute
- store

bank  bank
bank  bank
bank  bank

Input PLM

bank  bank

Output PLM

Private Local Memories (PLMs)

- Memory Controller
- Proc. core LEON3 CPU
- I/O Misc. Channels, etc.
- Accelerator WARP
- Accelerator DEBAYER
- Accelerator MATRIX-MUL
- Accelerator GRADIENT
- Accelerator MATRIX-ADD
- Accelerator CHANGE-DET
- Accelerator MATRIX-MUL
- Accelerator DEBAYER
- Accelerator MATRIX-ADD

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Hardware Accelerators with HLS

SystemC Specification

GRAYSCALE Interface

GRAYSCALE Logic

- Load
- Compute
- Store

Bank
Bank
Bank
Bank
Bank
Bank

Input PLM
Output PLM

Private Local Memories (PLMs)

High-Level Synthesis (HLS)

knob conf. #1

Cost (Area)

RTL

Performance (Latency)
Hardware Accelerators with HLS

SystemC Specification

GRAYSCALE Interface

GRAYSCALE Logic

load compute store

bank bank
bank bank
bank bank

Input PLM

Output PLM

Private Local Memories (PLMs)

High-Level Synthesis (HLS)

knob conf. #2

Cost (Area)

Performance (Latency)

RTL

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Hardware Accelerators with HLS

SystemC Specification

GRAYSCALE Interface

GRAYSCALE Logic

- load
- compute
- store

Input PLM

Output PLM

Private Local Memories (PLMs)

High-Level Synthesis (HLS)

knob conf. #3

Cost (Area)

Performance (Latency)

RTL
Hardware Accelerators with HLS

SystemC Specification

GRAYSCALE Interface

GRAYSCALE Logic

load
compute
store

bank
bank
bank
bank
bank
bank

Input PLM
Output PLM
Private Local Memories (PLMs)

High-Level Synthesis (HLS)

knob conf. #4

Cost (Area)

Performance (Latency)

RTL

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Hardware Accelerators with HLS

SystemC Specification

GRAYSCALE Interface

GRAYSCALE Logic

load | compute | store

bank | bank | bank | bank | bank | bank
Input PLM | Output PLM

Private Local Memories (PLMs)

High-Level Synthesis (HLS)

Pareto Optimal
Pareto Dominated

Cost (Area) vs. Performance (Latency)

Private Local Memories (PLMs)
Standard HLS Knobs

**Standard** knobs provided by the current HLS tools

<table>
<thead>
<tr>
<th>Knob</th>
<th>Settings and Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop manipulations</td>
<td>Unrolls, pipelines or breaks the body of loops</td>
</tr>
<tr>
<td>Array mappings</td>
<td>Maps arrays to registers or on-chip memories</td>
</tr>
<tr>
<td>Clock period</td>
<td>Sets the target clock period for synthesis</td>
</tr>
</tbody>
</table>

• These knobs enable already a rich design-space exploration
• However, they are not sufficient for exploring accelerators

→ We need other knobs to broaden the exploration
Motivational Example #1

- **Limiting factor**: limited bandwidth to the on-chip memory
- **We need** knobs to tailor the PLM to the accelerator needs

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Motivational Example #2

- **Limiting factor**: limited bandwidth to the off-chip memory
- **We need** knobs to operate on the communication interfaces
Contributions: Xknobs

**eXtended Knobs** for High-Level Synthesis

<table>
<thead>
<tr>
<th>XKnob</th>
<th>Settings and Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLM PORTS</td>
<td>Sets the on-chip memory bandwidth</td>
</tr>
<tr>
<td>DMA WIDTH</td>
<td>Sets the off-chip memory bandwidth</td>
</tr>
<tr>
<td>DMA CHUNK</td>
<td>Sets the size of the input and output PLM</td>
</tr>
</tbody>
</table>
Xknob #1: PLM PORTS

- Sets the number of read/write ports of input/output PLMs
- Higher values of PLM PORTS $\rightarrow$ more read/write accesses
- Higher values of PLM PORTS $\rightarrow$ higher area (more banks)
Xknob #2: DMA WIDTH

- Set the size in bits of the DMA communication channels
- Higher values of DMA WIDTH $\rightarrow$ higher mem. throughput
- Higher values of DMA WIDTH $\rightarrow$ higher area (more banks) (higher number of write/read ports of input/output PLMs)

### Graph

**DMA WIDTH**
- DMA WIDTH = 64
- DMA WIDTH = 128
- DMA WIDTH = 256
- DMA WIDTH = 512

**Normalized Area** vs. **Normalized Effective Latency**

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Xknob #3: DMA CHUNK

- Set the size of the PLM in multiple of the stored data type
- Higher values of DMA CHUNK → optimized communication
- Higher values of DMA CHUNK → higher area (for the PLM)

DMA CHUNK = 256  ■  DMA CHUNK = 512  ●  DMA CHUNK = 1024  ▲  DMA CHUNK = 2048  ▼

without contention

with contention

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Experimental Results

• We evaluate the combined effects of the **XKnobs** by using:
  • **GRAYSCALE** → accelerator limited by communication
  • **DEBAYER** → accelerator limited by computation

• The other WAMI accelerators behave similarly to either the GRAYSCALE accelerator or the DEBAYER accelerator
Experiment #1

- We consider two **XKnobs**: PLM PORTS and DMA WIDTH

<table>
<thead>
<tr>
<th>DMA WIDTH</th>
<th>PLM PORTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>64</td>
<td>4</td>
</tr>
<tr>
<td>128</td>
<td>2</td>
</tr>
<tr>
<td>256</td>
<td>1</td>
</tr>
</tbody>
</table>

- **GRAYSCALE** → accelerator limited by communication
Experiment #1

- We consider two **XKnobs**: PLM PORTS and DMA WIDTH

- **DEBAYER** → accelerator limited by computation

---

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Experiment #2

- We consider two **XKnobs**: PLM PORTS and DMA CHUNK

DMA CHUNK = 256  DMA CHUNK = 512  DMA CHUNK = 1024  DMA CHUNK = 2048

- **GRAYSCALE** \(\rightarrow\) accelerator limited by communication
Experiment #2

- We consider two **XKnobs**: PLM PORTS and DMA CHUNK

**DMA CHUNK** = 256 ▼ **DMA CHUNK** = 512 ● **DMA CHUNK** = 1024 ▲ **DMA CHUNK** = 2048 ▼

- **DEBAYER** → accelerator limited by computation

IEEE High Performance Extreme Computing Conference (HPEC), 2017
Concluding Remarks

• We presented the **XKnobs**
  • a set of knobs that aims at extending the standard knobs used in current HLS tools

• For WAMI, the **XKnobs** broaden the design space by up to **8.5x** for performance and **3.5x** for cost

• The **XKnobs** can be integrated in any HLS tools and design-space exploration methodologies to enrich the set of Pareto-optimal implementations of hardware accelerators
Broadening the Exploration of the Accelerator Design Space in Embedded Scalable Platforms

Thank you for the attention!

Speaker: Luca Piccolboni
Columbia University, NY