Lowest Power at the Edge: Keeping up with New Architectures Using HLS

Bryan Bowyer
Director of Engineering
Digital Design & Implementation Solutions Division
Agenda

- Design teams not able to optimize AI/ML systems in one iteration

- High-level Synthesis (HLS): A solution that has been waiting for this problem

- HLS already deployed for next generation ML hardware
DESIGN TEAMS NOT ABLE TO OPTIMIZE AI/ML SYSTEMS IN ONE ITERATION
Traditional Hardware Focuses on Simple Tasks We do Slowly or Poorly

- What are the next ten digits of Pi?
  3.1415926535
  8979323846

- Find all the edges in this picture
Machine Learning Hardware Needs Complexity: Targets Complex Tasks That We Do Well

- Identify what is in these pictures
Optimization Complexity: Evidence Power Could be Much Lower

**Bio-μCell**

<table>
<thead>
<tr>
<th>Memory:</th>
<th>~10⁴ bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic:</td>
<td>~300–150,000 bit</td>
</tr>
<tr>
<td>Power:</td>
<td>~10⁻² W</td>
</tr>
<tr>
<td>Heat:</td>
<td>~1 W/cm²</td>
</tr>
<tr>
<td><strong>Total energy/task</strong>:</td>
<td>~10⁻² J</td>
</tr>
<tr>
<td><strong>Task time</strong>:</td>
<td>510,000 s ~ 6 days</td>
</tr>
</tbody>
</table>

Source: Rahul Sarpeshkar, Analog Circuits and Biological Systems Group, Massachusetts Institute of Technology
More Complexity Coming Soon

- Analyzing one image is not enough
  - Context in Time
  - Context in Space

- New approaches to save energy

- Current Generation of HW will soon be obsolete
AI/ML Applications at the Edge
Challenges Moving from Idea to Implementation

- Edge ML Applications - critical requirements for performance and power
- CPU/GPU – too slow/too much power
- Even generic ML accelerator solutions will not be optimal for all networks especially for power
- Should you build your own?
- What architecture is best?
HLS: A SOLUTION THAT'S BEEN WAITING FOR THIS PROBLEM
Optimization Requires Multiple Passes Through Algorithm and Hardware Design

- Performance and Power are key
- Systems too complex to analyze before they are built
- Ongoing revolutionary changes in algorithm and hardware
- Many companies abandon first attempt at hardware

Algorithm Engineers work here.
Designers are already building NN HW using Catapult.
Catapult HLS
Most Practical Solution for Rapid Iterations

- Bring hardware and algorithm designers together
- Make late functional changes without impacting schedule
- New technology nodes are easy (or FPGA to ASIC)
- If you’re going to fail, fail fast

```c
void func (short a[N],
  for (int i=0; i<N; i++) {
    if (cond)
      z+=a[i]*b[i];
    else
      z+=a[i]*b[i];
  }
```
HLS ALREADY DEPLOYED FOR NEXT GENERATION ML HARDWARE
Catapult HLS Deployed in Production Designs Today

- **Video & Image Processing**
  - NVIDIA
  - Google
  - Facebook
  - Bosch
  - ST
  - Qualcomm

- **Machine Learning**
  - NVIDIA Research
  - Chips&Media
  - FotoNation
New Catapult HLS Toolkits
Jumpstart Building Low-Power AI/ML Accelerators

- Quality, working reference designs in vertical applications
- Four AI/Vision Toolkit designs available
  - Edge detection from HOG line-buffer architecture
  - 2-D convolution engine reconfigurable PE Array
  - 9 layer CNN full custom fused architecture
  - 9 layer CNN reconfigurable Eyeriss PE Array
- Includes FPGA demonstrator
- Platform includes CPU subsystem, HW/SW interface and HLS accelerator example for system integration
Summary

- HW and algorithm designers need to work together to optimize ML hardware
- Design teams using HLS can merge algorithm and hardware teams
- Novel ML architectures built using HLS already in production