Pixel Visual Core: Google's Fully Programmable Image, Vision, and AI Processor For Mobile Devices

Jason Redgrave, Albert Meixner, Nathan Goulding-Hotta, Artem Vasilyev, and Ofer Shacham
Motivation: The Vision

No HDR+  HDR+
Software Motivation

- Imaging, Vision, and AI are fast evolving fields
- Productizing new algorithms is difficult
  - ASIC design adds long delay into Research→Product cycle
  - CPU efficiency limits complexity in mobile space
  - Reliance on external vendors limits changes to the stack
- Software wants control, flexibility, and efficiency
Hardware Motivation

<table>
<thead>
<tr>
<th></th>
<th>CPU</th>
<th>GPU</th>
<th>ASIC</th>
<th>PVC’s IPU</th>
<th>Programmable Image Processing Unit (IPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (Op/sec)</td>
<td>10x</td>
<td>10-20x</td>
<td>&gt;1 Tera Ops/Sec</td>
<td>&gt;1 Tera Ops/Sec</td>
<td></td>
</tr>
</tbody>
</table>
Domain-specific Software

● High-level programming model is Halide
  ○ Domain-specific language for Image Processing
● IPU supports a subset of Halide language
  ○ No floating point
  ○ Limits on available memory access patterns
● Halide backend generates kernels and all API calls
  ○ Proprietary API for resource allocation and execution control
Compilation

- Halide backend generates high-level, virtual ISA (vISA)
  - RISC ISA with streaming-friendly memory model
  - Cross-generational & Architecture-independent
- Final compilation into physical ISA (pISA)
  - Compilation can be offline or on-device
  - Generation-specific VLIW ISA
  - All memory movements are explicit (no caches)
Conceptual View of Hardware: DAG of Kernels

- Many cores, each fully programmable
- Configurable DAG topology
Hardware resources view

- Line Buffer 0
- Line Buffer 1
- Line Buffer 2
- Line Buffer N

- Stencil processor 0
- Stencil processor M-1

From DRAM

To DRAM

Bus Ifc
Pixel Visual Core Architecture

- A53
- LPDDR4
- MIPI
- PCIe
- IPU

Chip Specs:
- TSMC 28nm
- 6.0 x 7.2 mm
- 426 MHz
- 512 MB DRAM
- Power <4500 mW
IPU Architecture

- **8x Cores**
  - STP
  - LBP
- **I/O**
  - DMA
  - MMU
  - SSP (Buffer)
- **Ring NoC**
Stencil Processor (STP)

- **Scalar lane**
  - Instruction RAM
- **SHG (Load/Store)**
- **2D Array**
  - 256 Compute lanes
  - 144 Halo lanes
  - Shared RAMs
  - Shift network
Compute Lane

- Single cycle!
- Dual 16b ALU
- 16b x 16b Multiply-Add
- Memory
  - Register File
  - Shared RAM (L0)
Read Neighbor Network

- Output pixel depends on neighboring inputs
- One vertical or horizontal shift per cycle
- Shifts are circular (toroidal)
- 1/2/3/4 hops in each direction
Line Buffer Pool (LBP)

- Data storage
- Synchronization
- 2D Line Buffer abstraction (2D FIFO)
Virtually Tall Line Buffer

- Memory saving over full buffer
- Elasticity set by SB width
- FB height set by kernel reuse
Results

Performance

- Align: 2.8x
- Merge: 2.8x
- Finish: 5.8x

Energy Efficiency

- Align: 6.9x
- Merge: 7.1x
- Finish: 16.7x

7-16x more energy-efficient despite 3-generation process gap!
Questions?