New Frontiers in On-Chip Image Processing

Richard Younger

CPAD Instrumentation Frontier Workshop 2021

March 18-21, 2021

©2021 Massachusetts Institute of Technology.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.
Evolution of High Performance Imagers

- **Useful Information Rate**
  - **PERCEIVE**
  - **RECOGNIZE**
  - **OBSERVE**

- Early Digital Imagers
- Consumer Electronics
- High Performance Imagers
- Adaptive Imagers
- Imagers with intelligence to adapt measurement approach in real time

- Film Cameras
- Next Gen Consumer Electronics
ROIC Fabrication Technology Landscape

As density increases, so does cost

- Mistakes are expensive
- Only the biggest experiments can afford to fabricate

More Transistors
High Capability

Transistors per Pixel

Circuit Density (Transistors per cm²)

Technology Node

- 100 µm
- 1.5 µm
- 800 nm
- 90 nm
- 65 nm
- 14 nm
- 7 nm

More Transistors
High Capability

Analog pixel
CMOS imagers
High Performance Imager Availability

- How can we bring advanced processing ROICs to more applications?

  ➢ Increase hardware *modularity* and *reconfigurability* to reuse ROICs and amortize costs
Hardware Modularity: Split the ROIC

Direct Bond Interconnect (Ziptronix)  Cu-Cu bond (Tezzeron)

Through-oxide-vias (Lincoln Lab)

IO pitch < 10 µm  Short interconnect lengths

Tier 3  Waveband Specific Detector Array
Tier 2  Detector Specific Analog Interface
Tier 1  Reconfigurable Digital Circuit with Integrated FPGA Processing
Demonstrating Modularity and Reconfigurability

Two tier 1 chips fabricated under the DARPA ReImagine program

- Advanced CMOS
- $1280 \times 1024$
- $12\mu m$ pitch
Demonstrating Modularity and Reconfigurability

**Griffin**
FPGA-based architecture

- Programmable Pixel Logic
- Reconfigurable Processing
- FPGA resource count in class with Kintex-7

**Manticore**
SoC-based architecture

- Programmable pixel array processor
- SIMD vector processor
- FPGA
- Microprocessor control core

- Advanced CMOS
- 1280 × 1024
- 12μm pitch

Each 16x16 macropixel has comparable MOPS to a Pentium I
Using Smart Imagers

Multimodal Sensing
Self-trigger a change in sensing based on events

Machine Learning
CNNs are feasible to implement on-chip

Conventional Processing
Many types of on-chip processing at rates in excess of 100,000 FPS

- In-pixel NUC
- In-array 2D FFT (slower)

More sophisticated adaptive processing is possible in-pixel and in-array
ReImagine ROIC Status

**Griffin**
- Multiple partners with Griffin tier-2 chips
- Griffin T1-only hardware development kit shipped January 2021

**Manticore**
- Multiple partners in process of tier 2 design
- Anticipate at least one bonded assembly to begin late 2021
- Demo camera in 2022
What’s next?

Near Future

• Demonstration cameras
• Design refinement
• Open to collaborations on Griffin & Manticore
• Open to studying needs for science & space applications

Looking Ahead

• Is on-chip ML the path for perceptive imagers?
• Create deep stacking architectures for full modularity
• Abuttability
Thank You

younger@ll.mit.edu
## MIT/LL Relimagine Team

<table>
<thead>
<tr>
<th>Rich Younger</th>
<th>Maria Blood</th>
<th>Renee Lambert</th>
<th>Tim Gagnon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valerie Finnemeyer</td>
<td>Glenn Garvey</td>
<td>Donna Yost</td>
<td>Erik Duerr</td>
</tr>
<tr>
<td>Jon Frechette</td>
<td>Kate Gillis</td>
<td>Sandra Lee</td>
<td>Eric Dauler</td>
</tr>
<tr>
<td>Sue Burzyk</td>
<td>Philemon Chose</td>
<td>Hannah Whisnant</td>
<td>Dan Ripin</td>
</tr>
<tr>
<td>Brian Tyrrell</td>
<td>Kate Thurmer</td>
<td>Praneeth Vepakomma (MIT)</td>
<td>Kaitlyn Dixon</td>
</tr>
<tr>
<td>Matt Stamplis</td>
<td>Jordan Lahanas</td>
<td>Tom Karolynshyn</td>
<td>Stephan Chase</td>
</tr>
<tr>
<td>Jonathan Leu</td>
<td>Tom Ross</td>
<td>Greg Rowe</td>
<td>Kenny Sims</td>
</tr>
<tr>
<td>Peter Grossmann</td>
<td>Ian Brown</td>
<td></td>
<td>Elaine Swenson</td>
</tr>
<tr>
<td>Matt Gregory</td>
<td>Tony Kryzak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>George Jordy</td>
<td>Austin Holloway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tom Cheng</td>
<td>Tony Soares</td>
<td>Hernan Castro</td>
<td></td>
</tr>
<tr>
<td>Jim Wey</td>
<td>David Volfson</td>
<td>Dan Santiago</td>
<td></td>
</tr>
<tr>
<td>Alice Lee</td>
<td>Jerry Lipson</td>
<td>Domenic Terranova</td>
<td></td>
</tr>
<tr>
<td>Ana-Maria Mandrila Vacca</td>
<td>Mike Cooper</td>
<td>Tim Smith</td>
<td></td>
</tr>
<tr>
<td>Brian Yu</td>
<td></td>
<td>Dmitriy Kaplan</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bob D'Ambra</td>
<td></td>
</tr>
</tbody>
</table>

**Designing and producing these chips is the product of a large team of professionals**
Machine Learning Feasibility

- Character recognition example using client/server split learning architecture
  - Split learning separates input data from full body of inference network
- Used hls4ml to translate character recognition split learning model in Keras to Vivado HLS
- Results: This model should, with further optimizations, fit on Griffin
  - Need to investigate network compression (pruning) & DSP re-use
  - 2D convolutional layers were a challenge for hls4ml (as of v0.1.6)

**Character Recognition Model FPGA Resource Utilization**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Split Learning Model</th>
<th>Kintex UltraScale 115</th>
<th>Griffin FPIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUTs</td>
<td>92,395</td>
<td>663,390</td>
<td>275,032</td>
</tr>
<tr>
<td>I/O</td>
<td>634*</td>
<td>702</td>
<td>275</td>
</tr>
<tr>
<td>DSP tiles</td>
<td>3,286</td>
<td>5,530</td>
<td>1,736</td>
</tr>
<tr>
<td>BRAM</td>
<td>60.5</td>
<td>2,160</td>
<td>820</td>
</tr>
</tbody>
</table>

* Includes image