A 2.9 TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems

Giuseppe Desoli, Nitin Chawla, Thomas Boesch, Surinder-pal Singh, Elio Guidetti, Fabio De Ambroggi, Tommaso Majo, Paolo Zambotti, Manuj Ayodhyawasi, Harvinder Singh, Nalin Aggarwal

STMicroelectronics
Outline

Introduction
Chip architecture
DCNN mapping
Hardware accelerators
Physical implementation
Results
Outline

Introduction

Chip architecture
DCNN mapping
Hardware accelerators
Physical implementation
Results
Deep Convolutional Neural Networks a key enabler

- DCNNs excel in many computer vision applications
- Being applied to other tasks too
- Can become a key component of ‘intelligent’ IoT networks for edge computing
- BUT …
DCNN’s Complexity Evolution

Operations (GOPS)

Parameters (Millions)

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Layers</th>
<th>Operations (GOPS)</th>
<th>Parameters (Millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNs</td>
<td>1997-2007</td>
<td>3 layers</td>
<td>0.0002</td>
<td>0.01</td>
</tr>
<tr>
<td>AlexNET</td>
<td>2012</td>
<td>7 layers</td>
<td>1.0</td>
<td>60</td>
</tr>
<tr>
<td>GoogleLeNet</td>
<td>2014</td>
<td>22 layers</td>
<td>1.5</td>
<td>138</td>
</tr>
<tr>
<td>VGG19</td>
<td>2014</td>
<td>19 layers</td>
<td>19.6</td>
<td>150</td>
</tr>
<tr>
<td>ResNet</td>
<td>2015</td>
<td>152 layers</td>
<td>11.3</td>
<td></td>
</tr>
</tbody>
</table>

© 2017 IEEE
International Solid-State Circuits Conference
AlexNet basics

Tot. Operations: 832 M

Tot. Parameters: ~ 60M

Krizhevsky et al., NIPS 2012

© 2017 IEEE International Solid-State Circuits Conference
Introduction

Chip architecture

DCNN mapping

Hardware accelerators

Physical implementation

Results
A complete SoC for DCNN applications

Enabling use of DCNNs in embedded systems:
- Power efficiency, low cost
- Efficient memory hierarchy
- Flexibility to adapt to different DCNN topologies
- Input/output capability
- Integrated with state of the art Deep Learning tools
8 DSP clusters, each with 2 32-bit DSPs, 4-way 16KB I-Cache, 64KB Local RAM and a shared 64KB RAM

(6uW/Mhz@0.6V) up to 1GHz in ST FD-SOI 28nm technology

ISA extensions for DCNN execution

2D-DMA with independent channels, linked list, stride, padding, etc.
Memory Hierarchy

4 MB (4x16x64KB) of shared RAM grouped with a 64 bits bus port x bank to sustain peak DCNN throughput

Each 64KB cut has individual sleep line control to activate it on demand

L2 SW controlled cache for feature maps and parameters
Fits an AlexNet level of complexity (compressed)

Energy/power x word access

Local SRAM 1x
On-chip SRAM 10x
LPDDR 100x
HW Accelerators subsystem

- 8 Conv Accelerators
- 16 CDNN specialized streaming DMA engines

Configurable framework supports data-flow based processing

Additional IPs
- H264, MJPEG, 2 Census, 2 croppers, Corner detector, 4 color conv, 4 sensors input IFs, 1 DVI output IF, digital MIC array IF
Introduction
Chip architecture
DCNN mapping
Hardware accelerators
Physical implementation
Results
AlexNet HW/SW partitioning

Tot. Operations: 832 M

- CONV layers: 1 Conv Acc → 36 MACs
- Non conv layers to DSPs to accommodate DCNN future evolution (leaky RELU, etc.)
AlexNet memory footprint

- On-chip SRAM
  - 2318 KB for parameters (8 bits)
  - 1436 KB for feature maps (16 bits)
- ~10 MB of external RAM for FC layers
Logical to physical mapping

Feature maps and kernels are sliced into batches processed iteratively and results are accumulated.

Batch size set x layer
Matching features and kernels parameters to HW resources and ceilings.
Introduction
Chip architecture
DCNN mapping
Hardware accelerators
Physical implementation
Results
Reconfigurable Accelerator Framework

- Design time parametric
- Unidirectional stream links create ad-hoc processing chains.
- DMA engines run autonomously through linked lists and synch up with the DSP clusters.

Diagram:
- Color conv
- Crop
- JPEG
- Ctrl Regs
- Sensor IF
- Sensor IF
- Display IF
- BUS ARBITER & INTERFACE
- BATCH
- RGB IMAGE
- IMAGE
- BATCH^{-1}
- FMAP
- KERNEL
- CA 0
- CA 1
- CA 7
Reconfigurable Accelerator Framework

Virtual stream links

- Ferry data to/from acc, IFs and DMA engines
- Flow control supported
- Streams multicast to multiple destinations

More flexible than hardwired data paths
More power efficient than a bus
Virtual Stream Links

Simple chains

Chains with forks

Joins single IF

Joins with multiple IF

Forks and hops
DCNN Convolution Accelerator

Kernels: 1x1 to 12x12, 8bit => 16 bit up to 4 in parallel

Batch size: Up to 16

Feat. size: 512: ker > 6x6, 1024: ker > 3x3, 2048: others

H/V padding, H/V stride

MACs: 16 bits inputs, 40 bits accum. scaling, rounding, saturation
Exploit Parallelism and Locality

Parallel Batch Execution

Chained Batch Execution

Parallel/Chained Batch Execution

Chained and parallel batch execution on multiple CAs reduces bandwidth, power, and # of DMA channels
Parameter Compression

- Kernel weights can be quantized non linearly with 8 of fewer bits (e.g. with KNN),
- Convolution Accelerator supports decompression in HW
- AlexNet top-1 classification error rate increase of 0.3%
Ultra-Wide DVFS Range

- LVT design with heterogeneous Poly-Bias levels → perf vs leakage
- GALS and low insertion delay clock networks to minimize on chip variation margins;
- Mono Supply memories with fine grained power switches and sleep mode
- DVFS energy efficiency improvements via body bias

Wide DVFS Range

<table>
<thead>
<tr>
<th>Frequency</th>
<th>GOPS/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.575</td>
<td>2930</td>
</tr>
<tr>
<td>0.6</td>
<td>2691</td>
</tr>
<tr>
<td>0.7</td>
<td>1977</td>
</tr>
<tr>
<td>0.825</td>
<td>950</td>
</tr>
<tr>
<td>1</td>
<td>1175</td>
</tr>
<tr>
<td>1.1</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>266</td>
</tr>
<tr>
<td></td>
<td>450</td>
</tr>
<tr>
<td></td>
<td>1423</td>
</tr>
<tr>
<td></td>
<td>969</td>
</tr>
<tr>
<td></td>
<td>801</td>
</tr>
</tbody>
</table>

© 2017 IEEE International Solid-State Circuits Conference
High Performance Low Voltage Mono-supply SRAM

• Wide Voltage Range
  ▪ 0.12u² single p-well bitcell with reduced variability
  ▪ In-situ tracking of bitcell current and programmable read time for best speed and lowest dynamic power
  ▪ In-situ tracking of wordline delay and slope for robust low voltage read/write

• Energy conservation
  ▪ Independent array and periphery power switches
  ▪ In-built isolation for FSM stability in power-down modes
  ▪ Extensive internal signal and clock gating
  ▪ Programmable buffers to optimize performance and power across instances
### Chip Specs

<table>
<thead>
<tr>
<th>Feature</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>FD-SOI 28nm</td>
</tr>
<tr>
<td>Package</td>
<td>FBGA 15x15x1.83</td>
</tr>
<tr>
<td>Frequency</td>
<td>200MHz–1.175GHz</td>
</tr>
<tr>
<td>Supply voltages</td>
<td>0.575–1.1 V digital</td>
</tr>
<tr>
<td></td>
<td>1.8V I/O</td>
</tr>
<tr>
<td>Power (***)</td>
<td>41 mW</td>
</tr>
<tr>
<td>On-chip RAM</td>
<td>4x1MB 8x192KB 128KB</td>
</tr>
<tr>
<td>Host</td>
<td>ARM® Cortex®-M4</td>
</tr>
<tr>
<td>No of DSPs</td>
<td>16</td>
</tr>
<tr>
<td>Peak DSP perf</td>
<td>75 GOPS (dual 16b MAC(*) loop)</td>
</tr>
<tr>
<td>No of CAs</td>
<td>8</td>
</tr>
<tr>
<td>CAs perf</td>
<td>676 GOPS(*) peak</td>
</tr>
</tbody>
</table>

(*) 1 MAC defined as 2 OPS (ADD + MUL)  
(**) Only HW Acc avg power for AlexNet
Introduction
Chip architecture
DCNN mapping
Hardware accelerators
Physical implementation

Results
### AlexNet CAs Performance

<table>
<thead>
<tr>
<th>Layer</th>
<th>MOPS</th>
<th>Time [ms]</th>
<th>Load %</th>
<th>GOPs/W</th>
<th>Pwr [mW]</th>
<th>GOPs/W</th>
<th>Pwr [mW]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16(F)x16(W)</td>
<td>8(F)x8(W)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>max</td>
<td>avg</td>
<td>max</td>
<td>avg</td>
</tr>
<tr>
<td>1</td>
<td>210.8</td>
<td>2.5</td>
<td>80</td>
<td>1228</td>
<td>988</td>
<td>86</td>
<td>1810</td>
</tr>
<tr>
<td>2</td>
<td>447.8</td>
<td>6.5</td>
<td>86</td>
<td>1475</td>
<td>1262</td>
<td>54</td>
<td>2175</td>
</tr>
<tr>
<td>3</td>
<td>299</td>
<td>3.6</td>
<td>73</td>
<td>1987</td>
<td>1445</td>
<td>58</td>
<td>2930</td>
</tr>
<tr>
<td>4</td>
<td>224.2</td>
<td>2.7</td>
<td>73</td>
<td>1987</td>
<td>1445</td>
<td>58</td>
<td>2930</td>
</tr>
<tr>
<td>5</td>
<td>149.6</td>
<td>1.8</td>
<td>72</td>
<td>1987</td>
<td>1434</td>
<td>58</td>
<td>2930</td>
</tr>
<tr>
<td>Total</td>
<td>1331.6</td>
<td>17.1</td>
<td>77</td>
<td>1677</td>
<td>1287</td>
<td>61</td>
<td>2473</td>
</tr>
</tbody>
</table>

200MHz @ 0.575V 25C, 4 chains of 2 CAs, batch of 1 image (227x227)
Comparisons with prior work

<table>
<thead>
<tr>
<th></th>
<th>This work</th>
<th>[3]</th>
<th>[4]</th>
<th>[5]</th>
<th>[6]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process (nm)</td>
<td>28 FD-SOI</td>
<td>65</td>
<td>65</td>
<td>28</td>
<td>65</td>
</tr>
<tr>
<td>VDD (V)</td>
<td>0.575–1.1</td>
<td>1.2</td>
<td>NA</td>
<td>0.9</td>
<td>0.82–1.17</td>
</tr>
<tr>
<td>(1)Power (mW)</td>
<td>39</td>
<td>45</td>
<td>485</td>
<td>15970</td>
<td>278</td>
</tr>
<tr>
<td>Freq (MHz)</td>
<td>200-1175</td>
<td>125</td>
<td>980</td>
<td>606</td>
<td>100–250</td>
</tr>
<tr>
<td>Memory (kB)</td>
<td>4096</td>
<td>36</td>
<td>44</td>
<td>16x2048</td>
<td>108</td>
</tr>
<tr>
<td>8*192+128</td>
<td></td>
<td></td>
<td></td>
<td>4096</td>
<td></td>
</tr>
<tr>
<td>Die size (mm²)</td>
<td>2.2 CAs</td>
<td>16</td>
<td>3.02</td>
<td>67</td>
<td>12.25</td>
</tr>
<tr>
<td>34 whole chip</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak perf (GOPS)</td>
<td>676 (CAs)</td>
<td>64</td>
<td>452</td>
<td>5580</td>
<td>84</td>
</tr>
<tr>
<td>76 (DSP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak effic. (GOPS/W)</td>
<td>2930</td>
<td>1420</td>
<td>932</td>
<td>350</td>
<td>302</td>
</tr>
</tbody>
</table>

(1) Power for best efficiency (this work @ 200 MHz, 0.575V, 25C)

© 2017 IEEE
International Solid-State Circuits Conference

14.1: A 2.9 TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems
AlexNet Complete Application

Input image

Sensor IF

CROP 227x227

DMA

RGB → YUV

KER. MEM

IN FMAP

JPEG

DMA

DMA

CA

CA

HOST + SPI + DMA

MEM

DSPs

OUT FMAP

37.5 mW @ 200MHz, 0.6V

10 FPS (38 ms DSPs, 62 ms 2 chained CAs)

Dynamic: 10 mW CAs + 17 mW system

Static: 0.6 mW CAs + 9.9 mW system

© 2017 IEEE
International Solid-State Circuits Conference

14.1: A 2.9 TOPS/W Deep Convolutional Neural Network
SoC in FD-SOI 28nm for Intelligent Embedded Systems
Summary

• An ultra-low-power SoC for DCNN real-word embedded and IoT applications
  – Reconfigurable HW acceleration data-flow framework
  – Parametric HW accelerator for convolutional layers of large DCNNs
  – Exploits different kinds of parallelism to improve performance and reduce power
  – DSP array with an optimized ISA for other DCNN’s layers and additional processing needs
  – Designed in FD-SOI28, ultra wide DVFS capability
  – Peak efficiency of 2.9 TOPS/W on AlexNet