The Next Generation of GAP
An IoT Application Processor for Inference at the Very Edge

Presented by:
Martin Croome
VP Marketing
GreenWaves Technologies
GreenWaves

• French fabless semiconductor startup
• Based in Grenoble, France
• Founded in November 2014
• Focused on designing and selling chips for AI and signal processing on battery operated IoT and wearable devices
Machine learning is becoming an attribute of all connected things

Sense

Interpret & Analyse

Act & communicate

IoT & wearable Devices

require

Signal processing and inference capabilities within a highly constrained power budget
Machine learning is becoming an attribute of all connected things
The fundamentals of GAP – Intelligence at the very edge

• MCU class energy consumption
  • Highly efficient parallelization
  • Sophisticated architecture (including instruction set architecture extensions)
  • Explicit memory movement

• Agility
  • Fine grained compute / energy scaling
  • Ultra fast state transitions

• Programmability
  • Applicable to many real world problems – not just CNNs
  • Exploits fast evolution of state-of-the-art
  • Single code model across architecture

Single chip solution for an intelligent sensor
GAP8

- First generation GAP processor
- Based on 5 years of research in PULP program
- TSMC 55nm process
- 9 Extended ISA RISC-V RV32 IMC cores
- 8 core cluster
- 1 core ‘fabric controller’
- HDKs since May 2018
- Production qualified
- First shipping products Q1 2020
GAP8 has already achieved industry leading performance

• QVGA Face ID
  • Face detection: ~25ms ~1mW / frame / second
  • Face Reidentification: 400ms 22mW / frame / second
  • 93% accuracy on Labelled Faces in the Wild dataset
  • Embeddable owner detection on battery operated devices

• IR people detection
  • 80 x 80 IR Image - LynRED ThermEye
  • Image preprocessing + human detection
  • 62ms ~4.4mW / frame / second
  • 99% accuracy on internally collected training set.
  • A full solution for people counting / occupancy detection on a battery for > 5 years
Introducing GAP9

- GAP8
  - Combined market leading architecture...
  - ...with mature semiconductor process TSMC 55nm LP

- GAP9
  - Tunes GAP8 architecture with experience gained from GAP8
  - Exploits market leading semiconductor process Global Foundries 22nm FDX

- GAP9 establishes a new capability / power consumption point in the industry
  - 10 times larger problems than GAP8
  - 5 times less power than GAP8
  - Increases agility and programmability
GAP9 – Examples of architectural evolutions

• Increased Capability
  • Larger problems
    • 1.6MB internal RAM
    • Peak cluster L1 bandwidth of 41.6 GB/sec
    • Peak L2 bandwidth of 7.2 GB/s
    • Hardware compression
  • More compute states
    • 400MHz cluster top frequency
    • New power states
  • More flexibility
    • 32 / 16 / 8-bit floating point support
    • New bi-directional multi-channel digital audio interfaces
    • Additional CSI2 camera interface
• Increased security
  • HW AES 256/128 bit
  • HW Programmable Unclonable Function (PUF)
## GAP9 vs. Arm M7 on MobileNet v1

<table>
<thead>
<tr>
<th>Target</th>
<th>Clock (MHz)</th>
<th>Time (ms)</th>
<th>Cycles (M)</th>
<th>FPS</th>
<th>Active Power (mW)</th>
<th>Image Size</th>
<th>Channel Scaling</th>
<th>Top 1 ImageNet Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>STM32 H7</td>
<td>400</td>
<td>162.5</td>
<td>65</td>
<td>6.2</td>
<td>170</td>
<td>160x160</td>
<td>0.25</td>
<td>43%</td>
</tr>
<tr>
<td>GAP9</td>
<td>29</td>
<td>162.5</td>
<td>4.77</td>
<td>6.2</td>
<td>5</td>
<td>160x160</td>
<td>0.25</td>
<td>43%</td>
</tr>
<tr>
<td>GAP9</td>
<td>400</td>
<td>11.925</td>
<td>4.77</td>
<td>83.9</td>
<td>50</td>
<td>160x160</td>
<td>0.25</td>
<td>43%</td>
</tr>
<tr>
<td>GAP9</td>
<td>400</td>
<td>167.5</td>
<td>67</td>
<td>6.0</td>
<td>50</td>
<td>192x192</td>
<td>1</td>
<td>70%</td>
</tr>
</tbody>
</table>

STM H7 figures - Running MobileNet on STM32 MCUs at the edge, Manuele Rucci
Accuracy estimates from TensorFlow model library
ImageNet performance in 1000 image classes
But architecture is only 50% of the story – tools is the rest

• What are customers expecting (Different things)?
  - Expecting a packaged solution
  - Expecting a known network
  - Expecting to revolutionize the world

• Each of these customers requires a different tool set
GAPFlow

- A series of build system agnostic, modular tools that convert Graphs to GAP code
- Build examples based on Makefiles but usable with any build system
- NN focus but by no means limited to NN
- Use one, use all, use none
- Extendable
- Clear points of failure
- Enhanced with examples networks and full applications that use it
GAP AutoTiler – Explicit memory movement

- Data caches are not good for streamed data – low cache efficiency
- ML / Signal Processing data traffic sizing is known at compile time
- Generate code for automatic data tiling and pipelined memory transfer interleaved with parallel call to compute kernel

<table>
<thead>
<tr>
<th>AutoTiler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph Description</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Generators</td>
</tr>
<tr>
<td>Kernel</td>
</tr>
</tbody>
</table>
GAP family is enabling ground breaking applications at the very edge

- GAP9 development boards available in early 2020 for lead customers
- GAP9 simulator has been in customer hands since May 2019
- GAP8 shipping now in production
- GAP8 development boards and evaluation boards for vision and IR vision shipping now
- GAP SDK available on our GitHub repository
- Come and see our demonstrations on the Open HW Group booth

Real ... Now ...
Thank you!

Questions?