Tensilica - What’s New?

LUH Hanover Feb 7th 2018
Marcus Binning
Agenda

• Trends

• General DSP

• Audio / Voice

• Computer Vision / AI
Trends in the SoC World

• MORE
  – FEATURES!
  – Embedded Processing required for computer vision, CNN classifiers, far field voice processing, high speed cellular modems …
  – Power efficiency in low power IoT devices
  – Activity in the automotive ADAS / VR / AR markets – established companies and new ones
  – (Bigger) M&A .. Well almost!

• LESS
  – Power, less Energy per workload, time to develop, time to deploy

• MORE Innovation
  – Custom/Novel architectures still prevalent, still differentiators. Not everything can be “done in software on vanilla platforms”
Trends in the SoC World

• 7nm is here …
General DSP
The recently announced Cadence Tensilica Fusion G3 DSP IP core is a high-performance licensable programmable digital signal processor core targeting diverse signal processing applications such as communications, audio and industrial applications. BDTI, a technology analysis firm, benchmarked the Fusion G3 core on several typical digital signal processing functions, 

Finally, BDTI implemented and optimized a custom DSP function from scratch on the Fusion G3 

This report presents BDTI's independent evaluation of the Fusion G3 core's performance and ease of software development. The Fusion G3 DSP core’s wide SIMD (single-instruction, multiple-data) operations and VLIW (very long instruction word) instruction set provide excellent cycle efficiency on many DSP tasks, and yield performance that surpasses that of <snip!>.

Fusion G3 is also noteworthy for its doubleprecision floating-point support for precision-critical tasks. Cadence provides robust software development tools and DSP function libraries to help users effectively realize the core’s performance potential.
Fusion G6 Introduced

• Bigger brother to Fusion G3

• Same ISA, double the width, double the performance (in loop bodies!)

• Same scalable programming model

• Same richness of libraries
Fusion G DSP Block Diagram

- Instruction Memories
- Instruction Cache
- Data Cache
- Data Memory 0
  - Bank 0
  - Bank 1
- Data Memory 1
  - Bank 0
  - Bank 1

**Features**

<table>
<thead>
<tr>
<th></th>
<th>Fusion G3</th>
<th>Fusion G6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VLIW</strong></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>4-slot, 128-bits wide</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>SIMD Vector Width</strong></td>
<td></td>
<td>Vector Elements</td>
</tr>
<tr>
<td>16-bit fixed point</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>32-bit fixed and floating point</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td><strong>Data Path Width</strong></td>
<td></td>
<td>Bits</td>
</tr>
<tr>
<td>LD/ST, vector register files</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>Data cache/memories (4 banks)</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>Instruction cache/memories</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td><strong>Fixed-Point Compute</strong></td>
<td></td>
<td>MACs</td>
</tr>
<tr>
<td>16*16-bit</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>32*32-bit</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td><strong>Floating-Point Compute</strong></td>
<td></td>
<td>FMA/MACs, ADDSUB</td>
</tr>
<tr>
<td>Single-precision VFPUs</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Double-precision VFPUs</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

**Instruction Memory Interface (ECC/Parity)**

- Load/Store Unit 0
  - Vector Register File (32)
  - Align Register File (8)
  - Wide Vector Register File (4)
  - Vector Boolean Register File (16)
- Load Unit 1
  - 32-bit Scalar Register File (16/32/64)

**Bus Interface (ECC)**

- AXI4, AHB-Lite, ACE-Lite or PIF

**Cache Controller**

- Memory Protection Unit
- 4-Slot VLIW Instruction Decoder
- Timers, Interrupts, Performance Counters
- Power Management
- Debug Module
- iDMA

**Data Memory Interface (ECC/Parity)**

- Bank 0
- Bank 1

**32-bit Scalar Processing Unit**

- Custom Instructions

**Vector Processing Units**

- Vector FPU Single and Double Precision
Easy DSP Software Development with Xtensa Xplorer

- High-performance optimizing C/C++ Compiler
- Tools “know” your Fusion G DSP configuration
- Cleanly map C/C++ to SIMD & VLIW with no assembly
- Launch on ISS, SystemC, RTL, FPGA, or Silicon
- Extensive software DSP library & examples
- Code coverage, profiling, PC trace, multi-core support
- Familiar Eclipse-based GUI
- 3rd-party JTAG debug and real-time trace
Fusion G DSP for Radar, IoT, Audio & Multi-Purpose Applications

**Scalable DSP Solution**
- 128-Bit VLIW & up to 256-Bit SIMD
- Fixed and Floating Point
- 8/18/32/64-bit Data Types
- Single and Double Precision

**Multi-Purpose**
- Sensor Fusion, Radar, Voice
- Trigger, 802.11ah, Audio, and Low-End Vision

**Software & Ecosystem**
- 550+ DSP Functions
- Easy to Program in C/C++
- Auto-Vectorization

**Algorithm Performance**
- FFT, FIR, Matrix Multiply
- Dot Product, Biquad Filters
- Complex FFT…
Digital Audio / Voice
Audio Algorithms Increasing in Complexity
In the Home... and Soon in Automobiles

Surround-Sound Audio
Limited by speaker placement, # of channels

Object-Based Audio
(Dolby Atmos, DTS:X, MPEG-H)
Sounds can be moved around the listening space regardless of # of speakers
Other Audio Applications

- Far-field multi-mic scenarios – “The Alexa World”
  - Digital Assistants / smart speakers becoming ever more sophisticated

- More intelligent language interfaces
  - Neural networks moving to voice / language parsing
  - RNNs vs CNNs

- All needs to be handled in low power, at the edge
  - Not in the cloud
  - In low power
HiFi 3z - Enhancements to HiFi 3

• HiFi 3z highlights
  – Load Store support in 2 slots – HiFi 3 has L/S support in 1 slot
  – Advanced FLIX bundling (multiple base ISA ops per cycle)
  – Double the MACs for 16x16 (octal MAC)
  – Enhanced ISA for accelerating FFTs, FIRs and IIRs
  – New Instruction extensions to improve codecs (especially EVS) performance - for Mobile
  – 4 way 8-bit load for improved voice trigger performance
  – Support for ITU-T STL 2017 (pending approval)

• Performance improvement on some DSP functions
  – 16x16 FIR: +49%
  – 16x16 FFT: +38%
  – 32x32 FFT: +26%
HiFi 3z Target Applications

• Mobile – Smartphones, Tablets, Laptops
  – Audio and voice

• Home Entertainment – DTV, STB, Soundbars Gaming
  – Audio codecs such as Dolby, DTS, MPEG-H
  – Audio post processing
  – Immersive audio
  – Interactive audio and voice codecs for real time gaming

• Automotive – Digital Radio, Head Unit Infotainment
  – Audio codecs such as Dolby, DTS
  – Audio post processing, active noise control, in cabin communications

© 2018 Cadence Design Systems, Inc. All rights reserved.
Computer Vision
Solving the complete camera processing pipeline: Vision DSP

Cadence Imaging/Vision DSP Focus

Sensor

ISP

Bayer Noise Reduction
Different Sensor support
RGIR
RGBW

2D/3D Noise Reduction
Image Stabilization
HDR/WDR
Super Resolution

Vision P5/P6 Post processing (YUV)

Face Detection
People Detection
Object Detection
Gesture Detection
Neural Networks

Image/Video Analysis and AI

Vision P6 Vision C5

Image Processing

Vision & NN Processing
CNN Algorithm Development Trends

Increasing Computational Requirements
(~16X in <4 years)

- AlexNet (2012)
- Inception (2015)

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>MACS/IMAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALEXNET</td>
<td>724,406,816</td>
</tr>
<tr>
<td>INCEPTION V3</td>
<td>5,713,232,480</td>
</tr>
<tr>
<td>RESNET-101</td>
<td>7,570,194,432</td>
</tr>
<tr>
<td>RESNET-152</td>
<td>11,282,415,616</td>
</tr>
</tbody>
</table>

Network Architectures Changing Regularly

- AlexNet (bigger convolution); Inception V3 and ResNet (smaller convolution)
- Linear network vs. branch

New Applications and Markets

- Automotive, server, home (voice-activated digital assistants), mobile, surveillance

How do you pick an inference hardware platform today (2017) for a product shipping in 2019-2020+? How do you achieve low-power efficiency yet be flexible?
Vision P5/P6 Architecture (recap)
**Tensilica® Vision C5 DSP for Neural Networks**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete, standalone DSP that runs all layers of NN</td>
<td>convolution, fully connected, normalization, pooling…</td>
</tr>
<tr>
<td>Building a DSP for changing NN field – general purpose and programmable</td>
<td></td>
</tr>
<tr>
<td>Not a “hardware accelerator” paired with a vision DSP, rather a dedicated, NN-optimized DSP</td>
<td></td>
</tr>
<tr>
<td>Architected for multi-processor design – scales to multi-TMAC/sec solution</td>
<td></td>
</tr>
<tr>
<td>Same proven software tool set as Vision P5/P6 DSP</td>
<td></td>
</tr>
</tbody>
</table>
Vision C5: Architecture

Tensilica Vision C5 DSP for Neural Networks

- 1024 MACs
- 128-Way SIMD VLIW Architecture
- On-the-Fly Decompression
- Flexible ISA for Quantization

Instruction Memory 1
128

Instruction Memory 2
128

Instruction Cache
128

Timers, Interrupts, Performance Counters

Power Management

Pipeline Management

Instruction Decoder

Scalar Register File

Cache Controller

Scalar Processing Units

Data Memory 0

512

Bank 1

Bank 0

Data Memory 1

512

Bank 1

Bank 0

Memory Mux

Load

Load/Store

Custom Instructions

Vector Register File (32)

Accumulator Register File (16)

Vector Processing Units

AXI4 Interface

128

AXI4

128

AXI4 Interface
Vision C5: DSP Architecture

- Fixed point DSP with 8-bit and 16-bit data type support
- 1024 8x8 MAC or 512 16x16 MAC throughput per cycle
  - Emphasis on high utilization of MACs across range of layer dimensions
- SIMD architecture for high performance vector computing
  - 512-bit vector register file that can work as 1024-bit register (pairing 2 512-bit registers)
    - 128-way SIMD for 8-bit data type, 64-way SIMD for 16-bit data type
    - 3072-bit accumulator registers
- VLIW architecture to exploit instruction level parallelism
  - 88-bit wide VLIW instructions, supports 3 and 4 slot instruction formats
  - Ability to perform load/store, MAC/ALU/SELECT, PACK, decompress operations in parallel
- Dual load/store architecture, capable of two 512-bit loads or one load and one store in parallel
  - Including support for loading unaligned data from memory
  - Special addressing mode for efficient access of 3-D data
# Vision P6 vs Vision C5

<table>
<thead>
<tr>
<th></th>
<th>Vision P6</th>
<th>Vision C5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus</strong></td>
<td>Imaging and NN</td>
<td>NN</td>
</tr>
<tr>
<td><strong>MAC (8x8)</strong></td>
<td>256</td>
<td>1024</td>
</tr>
<tr>
<td><strong>MAC (16x16)</strong></td>
<td>64</td>
<td>512</td>
</tr>
<tr>
<td><strong>Single and Half Precision VFPU</strong> (optional)</td>
<td>32 way FP16, 16 way FP32</td>
<td>Not Required for Inference</td>
</tr>
<tr>
<td><strong>Accumulators</strong></td>
<td>4 x 1536b</td>
<td>8 x 3072b</td>
</tr>
<tr>
<td><strong>MAX SIMD Width</strong></td>
<td>64 way SIMD</td>
<td>128 way SIMD</td>
</tr>
</tbody>
</table>
| **Special Features**| Scatter Gather (needed by Imaging Applications) | • On the Fly Decompression Support  
• Special addressing modes  
• Richer set of convolution multipliers (signed and unsigned)  
• Extensive data rearrangement and selection |
Xtensa Neural Network Compiler (XNNC) (Starting From Vision P6)

- Connects to existing industry CNN frameworks by using their Trained Model descriptions
- and auto-generates Trained Model optimized code for Cadence CNN DSPs

Three Major components to XNNC:
- CNN Parser: Float to Fixed Point conversion
- CNN Code Generation and Optimization
- CNN Library for Vision DSP

First CNN Framework support: Caffe, followed by Tensorflow
- For both Vision P6 and Vision C5