Cadence Tensilica Edge AI Processor IP Solutions for Broad Market Use Cases

Pulin Desai
Group Director, Vision & AI Product Marketing
September 2020
Cadence Tensilica Processor and DSP IP Business

**TENSILICA® CUSTOMERS**

7B+ Processors Shipping Annually

**DSP LICENSING REVENUE**

#1 DSP IP LICENSING REVENUE

**Processor LICENSING REVENUE**

#2 Processor IP LICENSING REVENUE

**SEMICONDUCTORS**

19 of the Top 20 Semiconductor Vendors Use Tensilica

**TENSILICA LICENSEErs**


300+

GLOBAL ECOSYSTEM

200+ Ecosystem Partners

© 2020 Cadence Design Systems
Examples of Edge AI: Voice & Vision

Voice AI
- Smart Watch
- Mobile
- Headphones/Hearables
- Smart Speaker

Vision AI
- Mobile
- AR/VR
- Smart Surveillance
- Autonomous Vehicles
Processing Flow for Inference in the Embedded System

- Any sensing application always has a pre & post processing
- Need to solve the whole data path
Machine Learning Innovations Using Microphone as a Sensor

Voice UI
- Voice detection
- Voice recognition
- Speech enhancement
- Voice trigger/voice command
- Noise reduction
- Speaker ID

Sound Analytics
- Environment
- Sound classification
- Audio scene detection
- Baby cry
- Machine failure detection
- Song ID

Privacy, latency, power, and availability of network drives voice UI and other ML applications to edge devices
System Integration – Smart Speaker Example

AI Workload
Keyword/command recognizer
Voice enhancement & Noise suppression

Non-AI Workload but needs a DSP
EQ, MP3 Decoder, Speech decoder are Classical audio algorithms

© 2020 Cadence Design Systems
Computer Vision and Image Processing Pipeline

- Any Vision application has a pre & post processing
## Vision Applications: Mix of Vision and AI

<table>
<thead>
<tr>
<th>Face Detection</th>
<th>Vision Operation</th>
<th>AI Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Scale Pyramid</td>
<td>Face Detector</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HDR</th>
<th>AI Operation</th>
<th>Vision Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene Detection</td>
<td>HDR Processing</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bokeh with Single Camera</th>
<th>AI Operation</th>
<th>Vision Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation</td>
<td>Blurring and De-Blurring</td>
<td></td>
</tr>
</tbody>
</table>

- All use cases still have mix of vision and AI operations
- Need for both vision and AI processing in the camera pipeline
In MTCNN use case with 3 separate nets (P/O/R), glue processing is inserted between the nets to perform end to end face detection.

Even if entire networks are sped up the application speedup will be bottlenecked by glue processing which does not need to be “standardized”

A high performance DSP targeted at Vision applications can keep up with the evolving glue processing requirements.
Hypothetical Speedup when only Convolution is Accelerated

- Convolutions are common target for acceleration with hardware
- As convolutions speed up, different networks will achieve varying amounts of overall speedup
  - Some networks achieve only modest speedups bottlenecked by “other” layers
  - “Other” layers need to keep up with convolutions acceleration
Significant Cycle Consuming Non-Convolution Layers

- The set of “other” layers is large and continuously evolving
  - Non-convolution layers can be 20% to as high as >80% of the cycles
- A high performance DSP targeted at NN applications can keep up with the evolving requirements
# Neural Network Deployment Options in Embedded Systems

<table>
<thead>
<tr>
<th>Option</th>
<th>Metric:</th>
<th>Flexibility</th>
<th>Power</th>
<th>Performance</th>
<th>Time to Market</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td></td>
<td>✓ Highly flexible</td>
<td>Very high</td>
<td>Depends</td>
<td>Fast</td>
</tr>
<tr>
<td>CPU+GPU</td>
<td></td>
<td>✓ Highly flexible</td>
<td>Very high</td>
<td>Depends but could be high</td>
<td>Fast</td>
</tr>
<tr>
<td>Hardware Development</td>
<td></td>
<td>Not flexible</td>
<td>Low</td>
<td>High</td>
<td>Long development cycle and verification</td>
</tr>
<tr>
<td>Programmable DSP</td>
<td>✓ Highly flexible</td>
<td>✓ Low</td>
<td>Depends</td>
<td>✓ Fast</td>
<td></td>
</tr>
<tr>
<td>Programmable DSP+HW accelerator</td>
<td>✓ Highly flexible</td>
<td>✓ Low</td>
<td>High</td>
<td>✓ Fast</td>
<td></td>
</tr>
</tbody>
</table>
Tensilica AI Portfolio: AI Inference at the Edge

- **DNA** Scalable AI Processor

### Processor Range

- **2K MAC**
- **1K MAC**
- **512MAC**
- **256MAC**
- **<32MAC**

- **HiFi**
- **Voice+AI**
- **Vision Q7**
- **Vision P6**

**Key Features**

- ** Highly Flexible/Multi Use**
- **Best in class Power, Performance**

© 2020 Cadence Design Systems
HiFi DSP Is Ideal for ML Solutions on the Edge

Tensilica® HiFi DSP is the leading IP solution for audio, speech, and machine learning applications

Cadence enables developers to train and seamlessly deploy their machine learning applications on HiFi DSPs using TensorFlow Lite for Microcontrollers

HiFi DSP has extensive production-ready software solutions from Cadence and the largest ecosystem with 160+ software partners

Tensilica® Xtensa Audio Framework (XAF) accelerates integration of ML and traditional audio workloads, reducing time to market
Vision P6/Q7: Ideal DSP for Vision + AI

<table>
<thead>
<tr>
<th>Feature</th>
<th>Details</th>
</tr>
</thead>
</table>
| **Vision Acceleration**     | • 512-bit SIMD  
• 5 VLIW Slots  
• 8bit/16bit/32bit int, SP and HP FP Support  
• 0.7 to 1.7 TOPS |
| **AI Acceleration**         | • 256/512 8-bit MAC  
• FP16 support |
| **High Data Throughput**    | • 1024 bit memory I/F  
• Scatter Gather Technology  
• Multi-channel iDMA  
• 128/256 bit Axi |
| **Lib & Tools**             | • OpenCV Base imaging/vision Lib, SLAM Lib  
• OpenCL, Halide, OpenVX Support  
• NN Compiler (ONNX & GLOW) & Android NN API Support |
Tensilica DNA Processor: Single Processor for Vision & AI Sparse Compute Engine & Scalable

DNA = XNNE + Vision P6
- XNNE (Xtensa Neural Network Engine)
  - Scalable Sparse Engine
  - Scales from 256MAC to 2048MAC
  - Unified Buffer
- Scalable Sparse Compute Engine
- Quantization
- Pooling/Vector Processing Unit
- Unified Buffer
- AXI Interconnect

Vision P6:
- VLIW SIMD DSP
- For future proofing & expanding
- Pre & Post Processing
- CV workload
Typical Vision + AI Work: On Tensilica DNA Processor

<table>
<thead>
<tr>
<th>Pre-Processing</th>
<th>AI Workload (Convolution)</th>
<th>Post-Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runs on Vision P6</td>
<td>Runs XNNE HW of DNA Processor</td>
<td>Runs on Vision P6</td>
</tr>
</tbody>
</table>

- **Pre-Processing**
  - Multi-scale Pyramid
  - Color-space
  - Noise Reduction

- **AI Workload (Convolution)**
  - NN Workload
  - Mostly various convolution
  - Non-Convolution Layer

- **Post-Processing**
  - Non-Max Suppression (NMS)
  - ROI Align: Bilinear Interpolation
  - De-blurring
  - HDR

Example Networks:
- Yolo, SSD, MaskRCNN, CRNN
- Could be >10% cycles
- Depends on # of detection
- Depends on accuracy
XNNC-Link
NN Code Generation for customer’s AI Accelerator

- ONNX Model
  - Float to fixed-point conversion
  - NN graph optimizations
  - Multiple target backend support

Tensilica XNNC
- Vision P6 DSP
- Vision Q7 DSP
- DNA 150 Processor

Customer AI Accelerator

Vision DSP + Customer’s accelerator for your AI Solution

© 2020 Cadence Design Systems
“MediaTek has confirmed that the P60 integrates a Cadence Vision P6 core for its AI accelerator.” Source (1)

Mediatek i500: “AI acceleration for APU based on Tensilica Vision P6 (i500)” Source: (2)

“Toshiba Selects Cadence Tensilica Vision P6 DSP as Image Recognition Processor for its Next-Generation ADAS Chip” Source (3)

Kneron, the San Diego and Taipei-based low-power edge AI startup, “KL720 NPU IP integrated with Cadence Tensilica Vision P6 DSP IP” Source (4)

“NXP’s enablement for Glow is tightly coupled with the Neural Network Library (NNLib) that Cadence provides for its Tensilica HiFi 4 DSP delivering 4.8GMACs of performance.” Source: (5)

“Baidu today released its new chipset Honghu at the annual AI Developer Conference in Beijing. It features HiFi4 custom instruction set, dual-core DSP, and only 100mV power dissipation on average.” Source: (6)
• Voice+AI & Vision+AI for edge requires pre & post processing in addition to neural network processing

• In addition to convolution layers (which are MAC heavy) other layers take considerable cycles
  • MAC acceleration only does not accelerate the AI performance

• Combination of hardware acceleration for convolution and programmable DSP provides the best solution for Edge AI workload
  • Single hardware can be used for both concurrent Vision+AI and Voice+AI workload

• Cadence Tensilica provides Vision+AI and Voice+AI HW IP with NN Compiler SW
  • Cadence Tensilica IP is shipping in large number of edge-ai products
Resource Slide

Cadence Resources
https://ip.cadence.com/vision
https://ip.cadence.com/ai

External Resources
https://onnx.ai/
https://onnx.ai/supported-tools.html