Tensilica - What's New ?

LUH Hanover Feb 7th 2018
Marcus Binning

cadence®

Agenda

- Trends
- General DSP
- Audio / Voice
- Computer Vision / AI

Trends

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

Fusion G3 BDTI Report (excerpt) from 2016 …

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

…

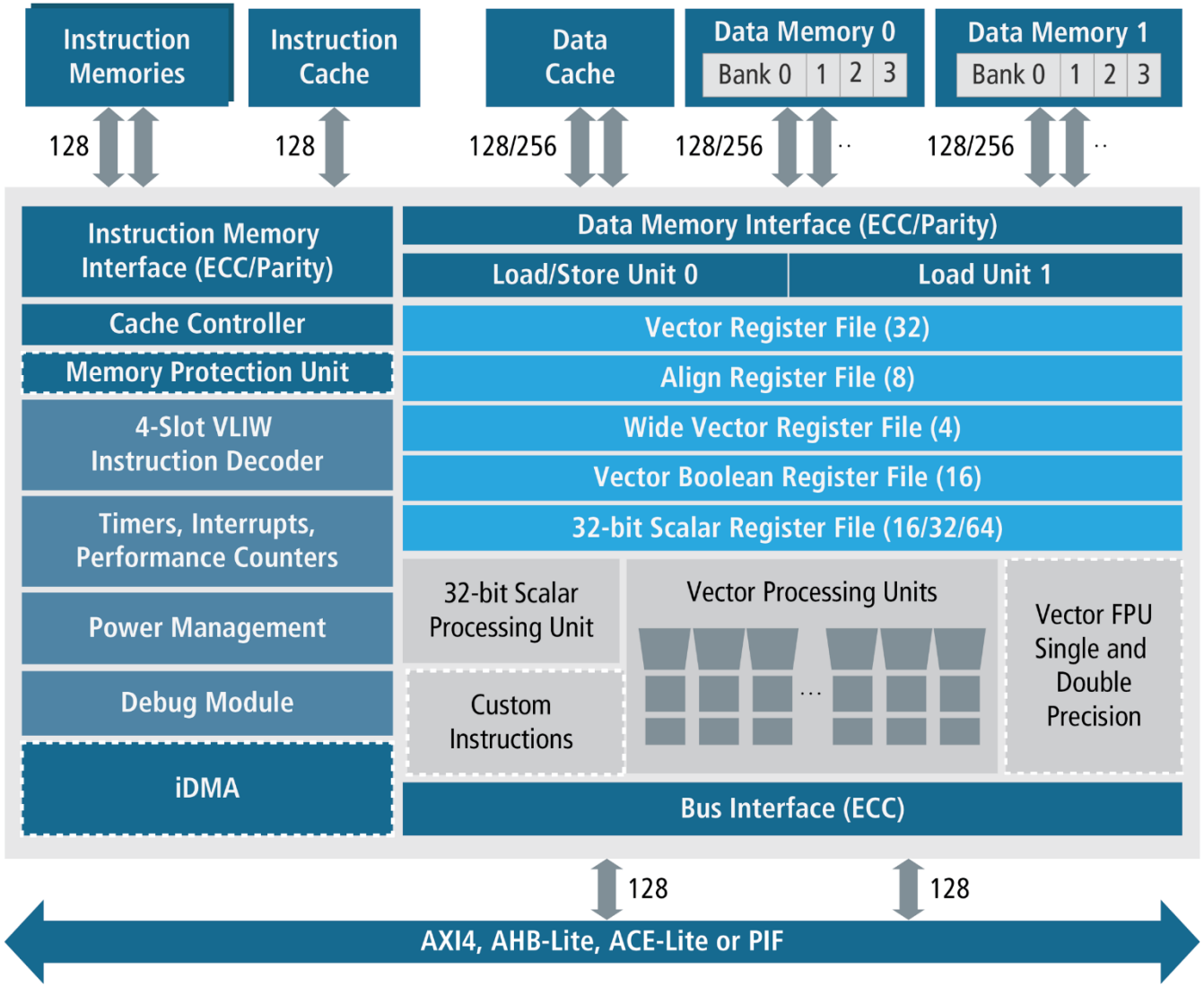
This report presents BDTI's independent evaluation of the Fusion G3 core's performance and ease of software development. The Fusion G3 DSP core's wide SIMD (single-instruction, multiple-data) operations and VLIW (very long instruction word) instruction set provide **excellent cycle efficiency** on many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



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

cadence

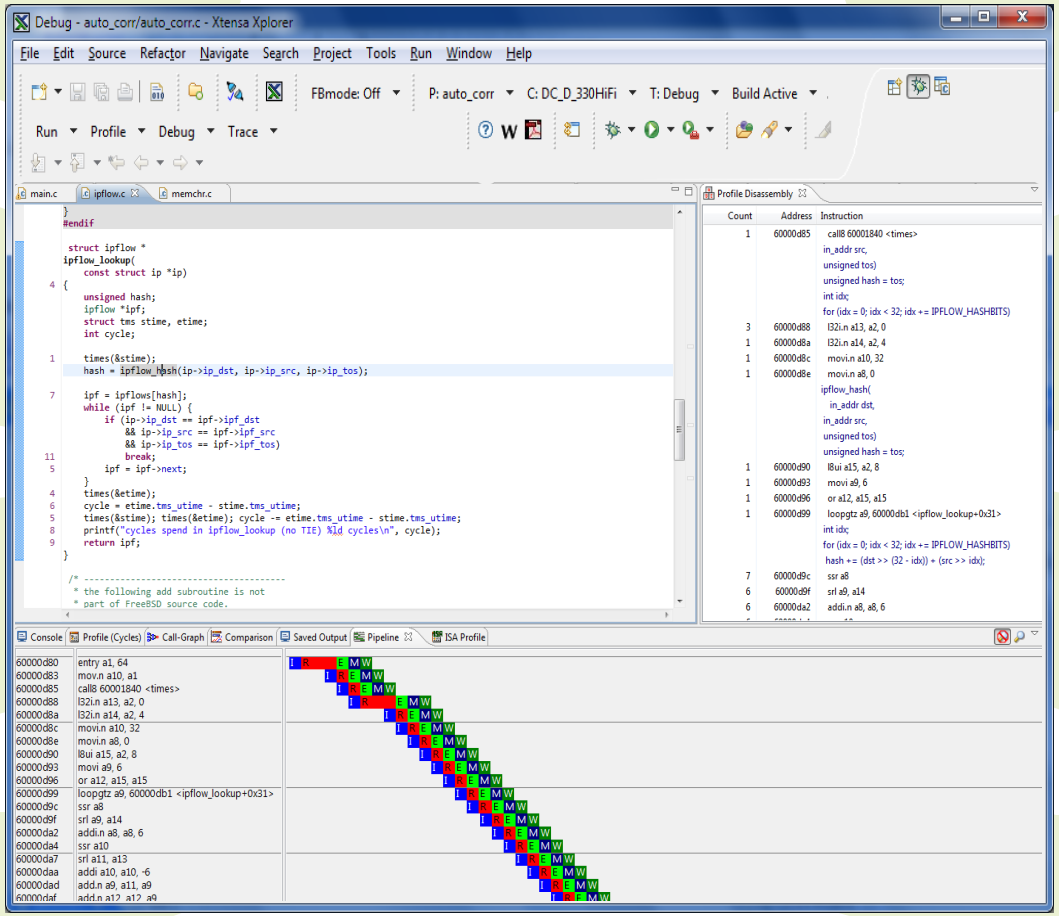
Easy DSP Software Development with Xtensa Xplorer

High-performance optimizing C/C++ Compiler

Cleanly map C/C++ to SIMD & VLIW with no assembly

Extensive software DSP library & examples

Familiar Eclipse-based GUI



Tools “know” your Fusion G DSP configuration

Launch on ISS, SystemC, RTL, FPGA, or Silicon

Code coverage, profiling, PC trace, multi-core support

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

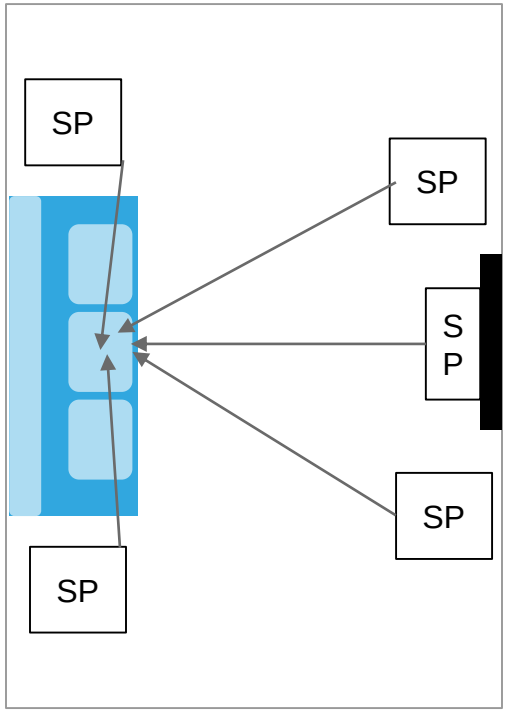
550+ DSP Functions
Easy to Program in C/C++
Auto-Vectorization

Algorithm Performance

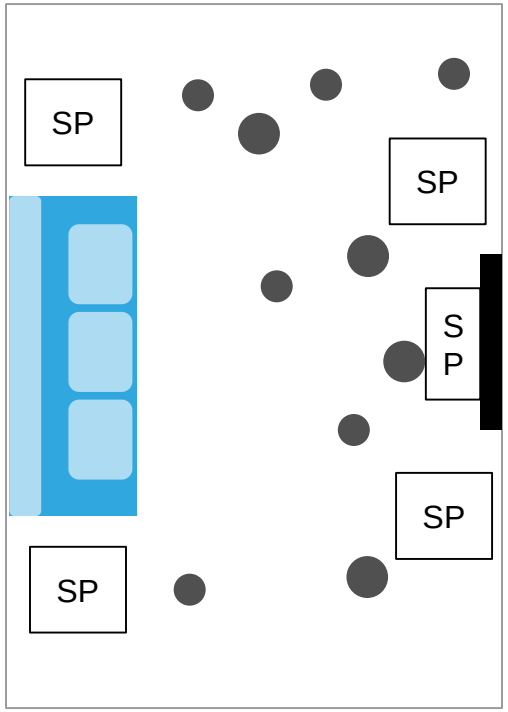
FFT, FIR, **Matrix Multiply**
Dot Product, Biquad Filters
Complex FFT…

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

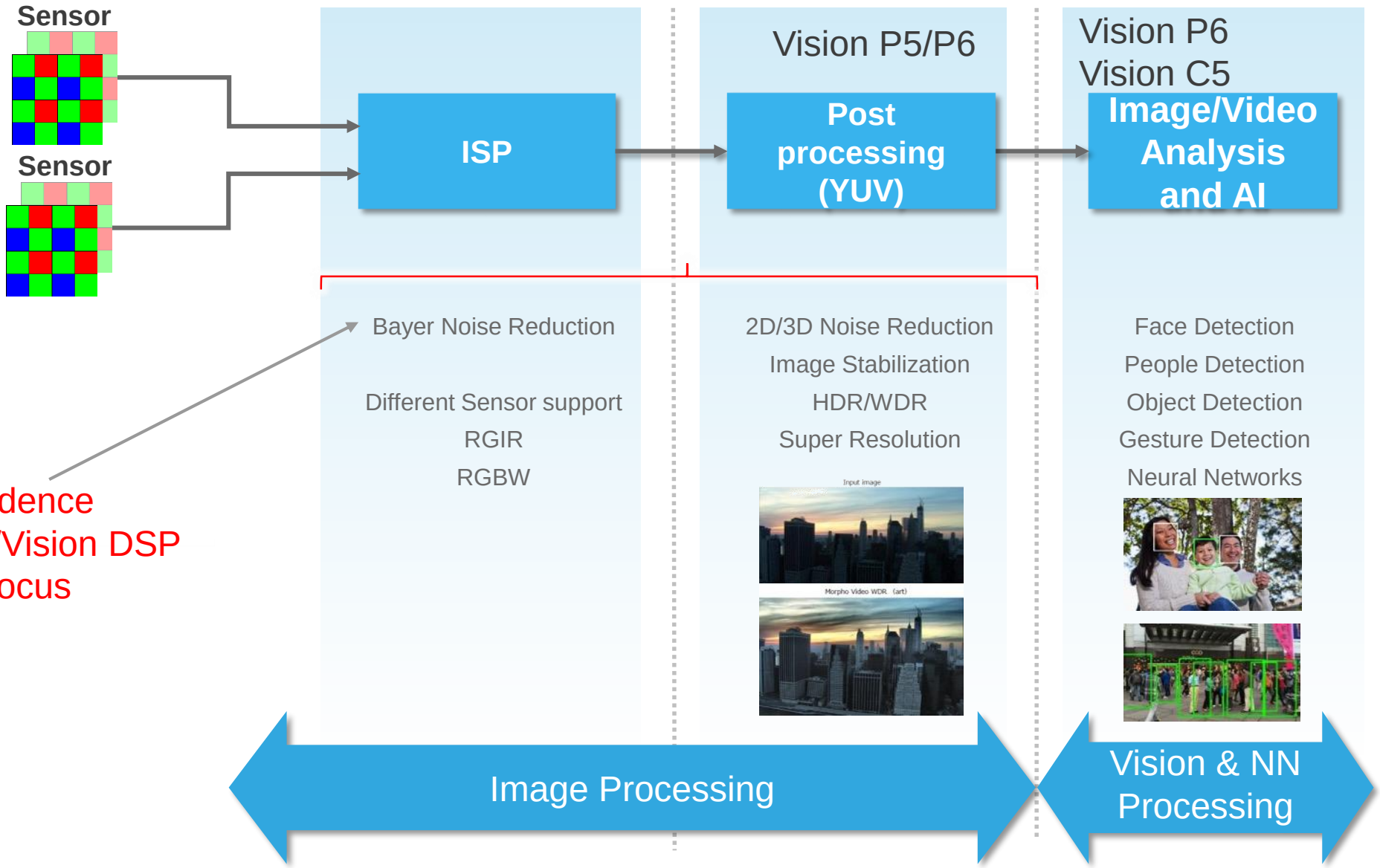
- Mobile – Smartphones, Tablets, Laptops
  - Audio and voice

- 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

Computer Vision

Solving the complete camera processing pipeline: Vision DSP



cadence

CNN Algorithm Development Trends

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

- AlexNet (2012)
- Inception (2015)
- ResNet (2015)

NETWORK	MACS/IMAGE
ALEXNET	724,406,816
INCEPTION V3	5,713,232,480
RESNET-101	7,570,194,432
RESNET-152	11,282,415,616

Network Architectures Changing Regularly

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

New Applications and Markets

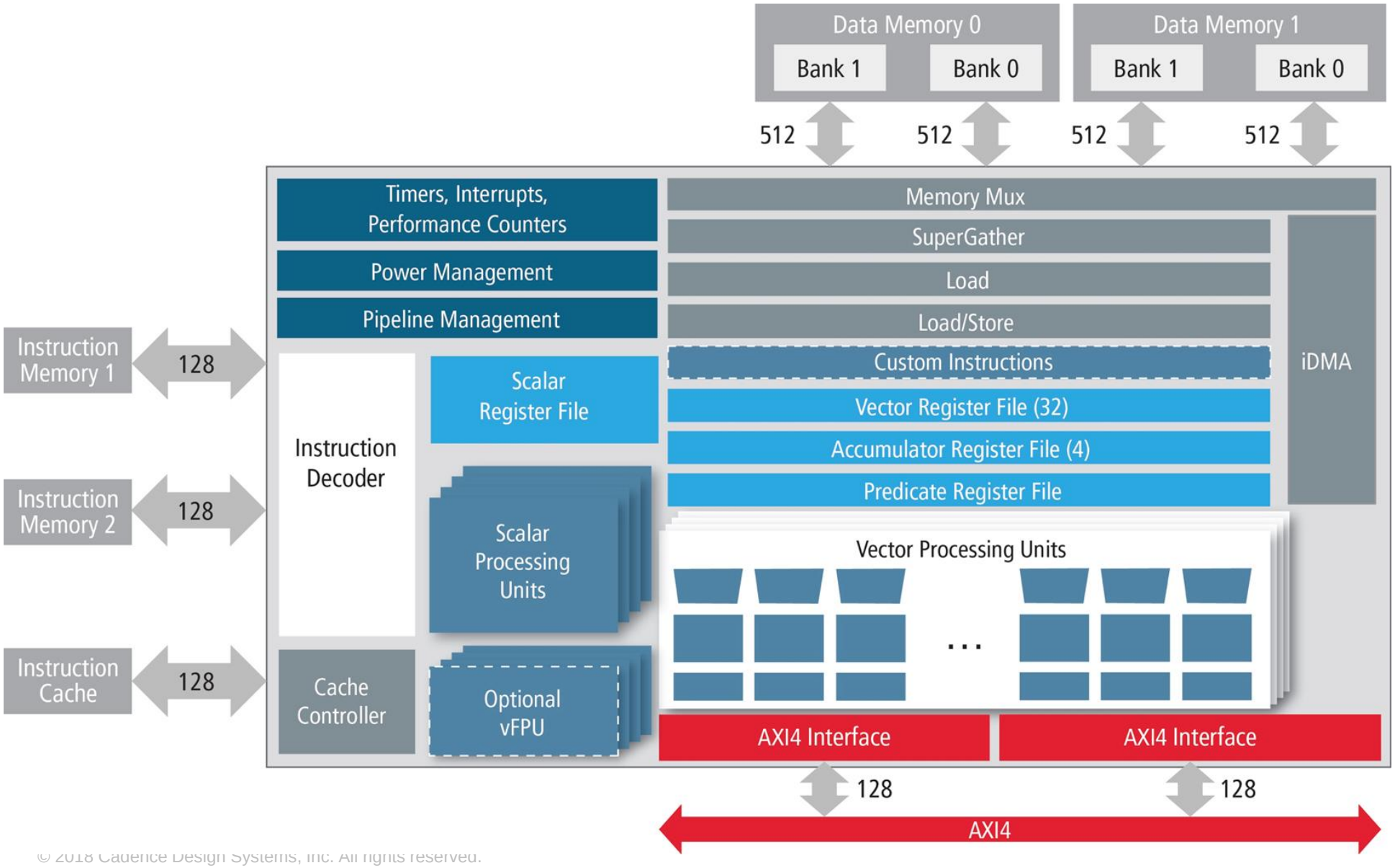
- Automotive, server, home (voice-activated digital assistants), mobile, surveillance

Lower

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?

cadence

Vision P5/P6 Architecture (recap)



Tensilica® Vision C5 DSP for Neural Networks

Complete, standalone DSP that runs all layers of NN (convolution, fully connected, normalization, pooling…)

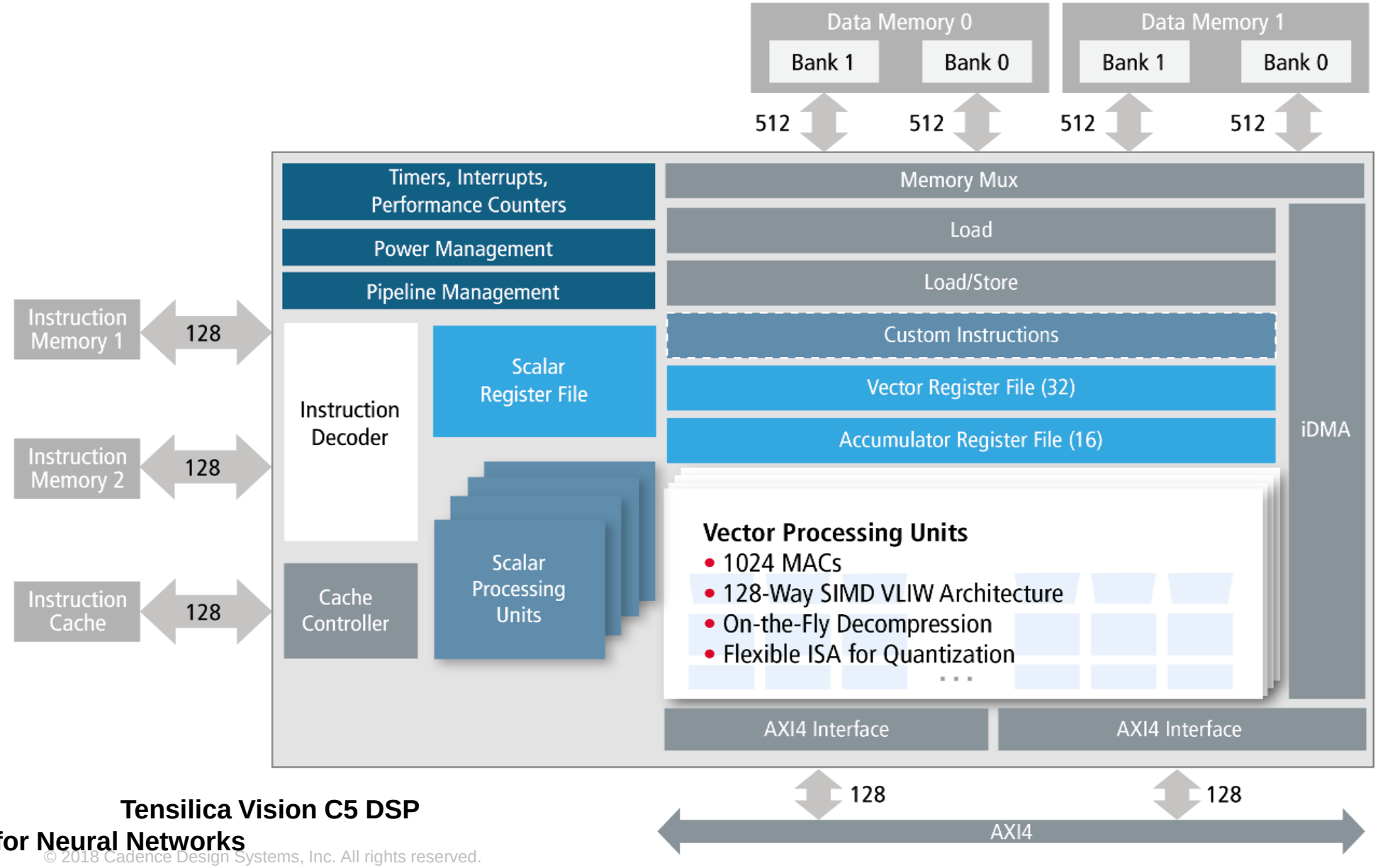
Building a DSP for changing NN field – general purpose and programmable

Not a “hardware accelerator” paired with a vision DSP, rather a dedicated, NN-optimized DSP

Architected for multi-processor design – scales to multi-TMAC/sec solution

Same proven software tool set as Vision P5/P6 DSP

Vision C5: Architecture



Tensilica Vision C5 DSP for Neural Networks

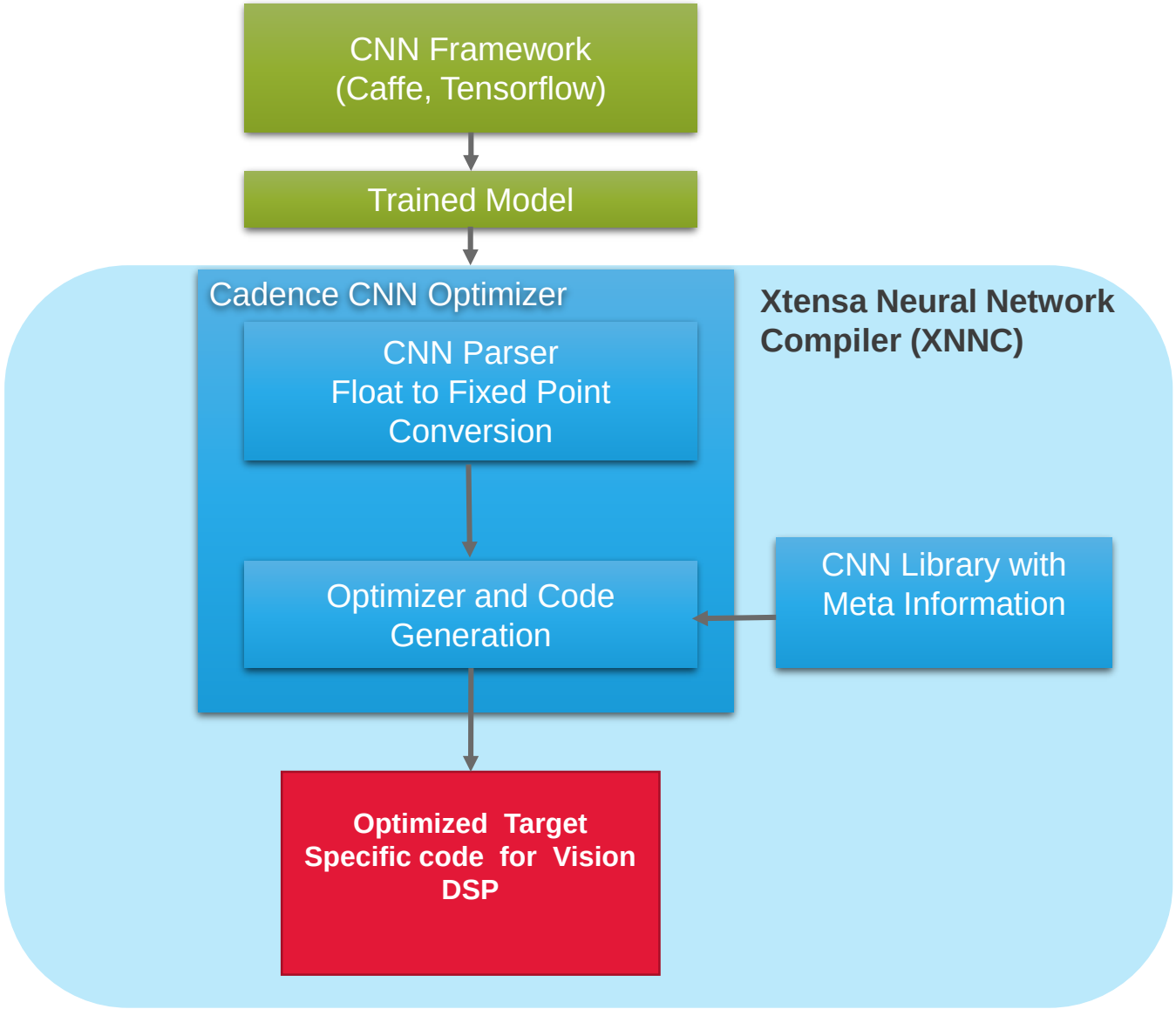
Vision C5: DSP Architecture

- 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

	Vision P6	Vision C5
Focus	Imaging and NN	NN
MAC (8x8)	256	1024
MAC (16x16)	64	512
Single and Half Precision VFPU (optional)	32 way FP16 16 way FP32	Not Required for Inference
Accumulators (to support higher MAC capability)	4 x 1536b	8 x 3072b
MAX SIMD Width	64 way SIMD	128 way SIMD
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


cadence®