Standalone, Scalable, Low Power, and Highly Flexible Neural Network DSP – Vision C5 DSP

IP Group

cadence®

Cadence Tensilica Processor and DSP IP Business

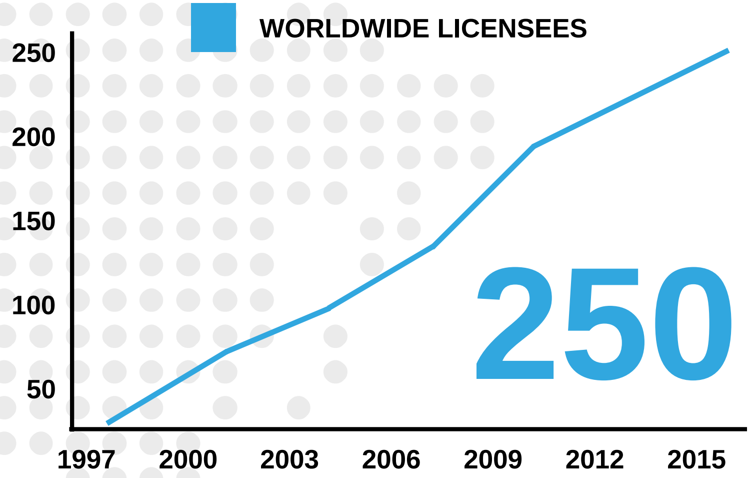
TENSILICA® CUSTOMERS

4B+ Processors SHIPPING Annually

DSP LICENSING REVENUE

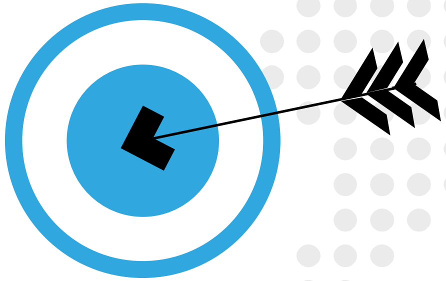
#1 DSP IP LICENSING REVENUE

TENSILICA LICENSEES



LEADING AUDIO DSP IP

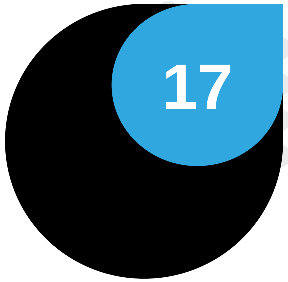
TOP AUDIO DSP CHOICE



GLOBAL ECOSYSTEM

200+ ECOSYSTEM PARTNERS

SEMICONDUCTORS



17 of the Top 20 SEMICONDUCTOR VENDORS USE TENSILICA

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

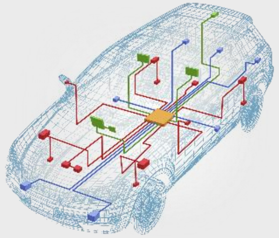
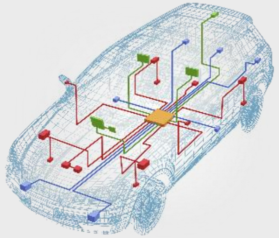
Low Power

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

Neural Network Workloads Vary by End Market

Pick the right inference platform for the market—One size does NOT fit all!

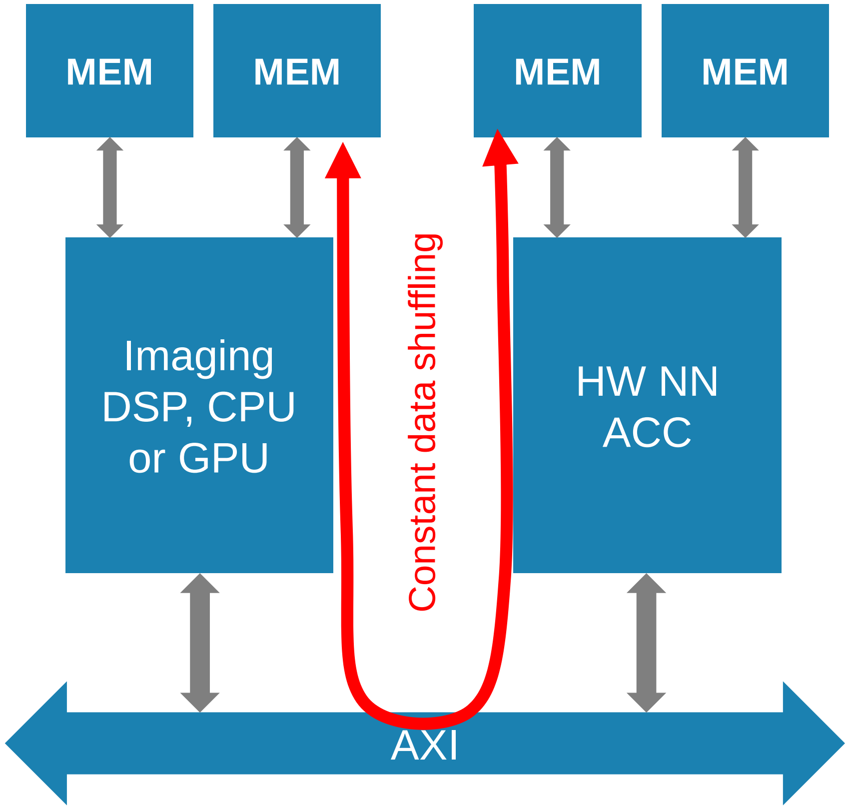
Processing Power			
Up to 10TMAC/sec	Automotive (towards autonomous)		Runs multiple NNs all the time
1TMAC/sec	Surveillance / Automotive (semi-autonomous)	 	Runs few NNs all the time
<200 GMAC/sec	Mobile, AR/VR		Runs a NN once in a while

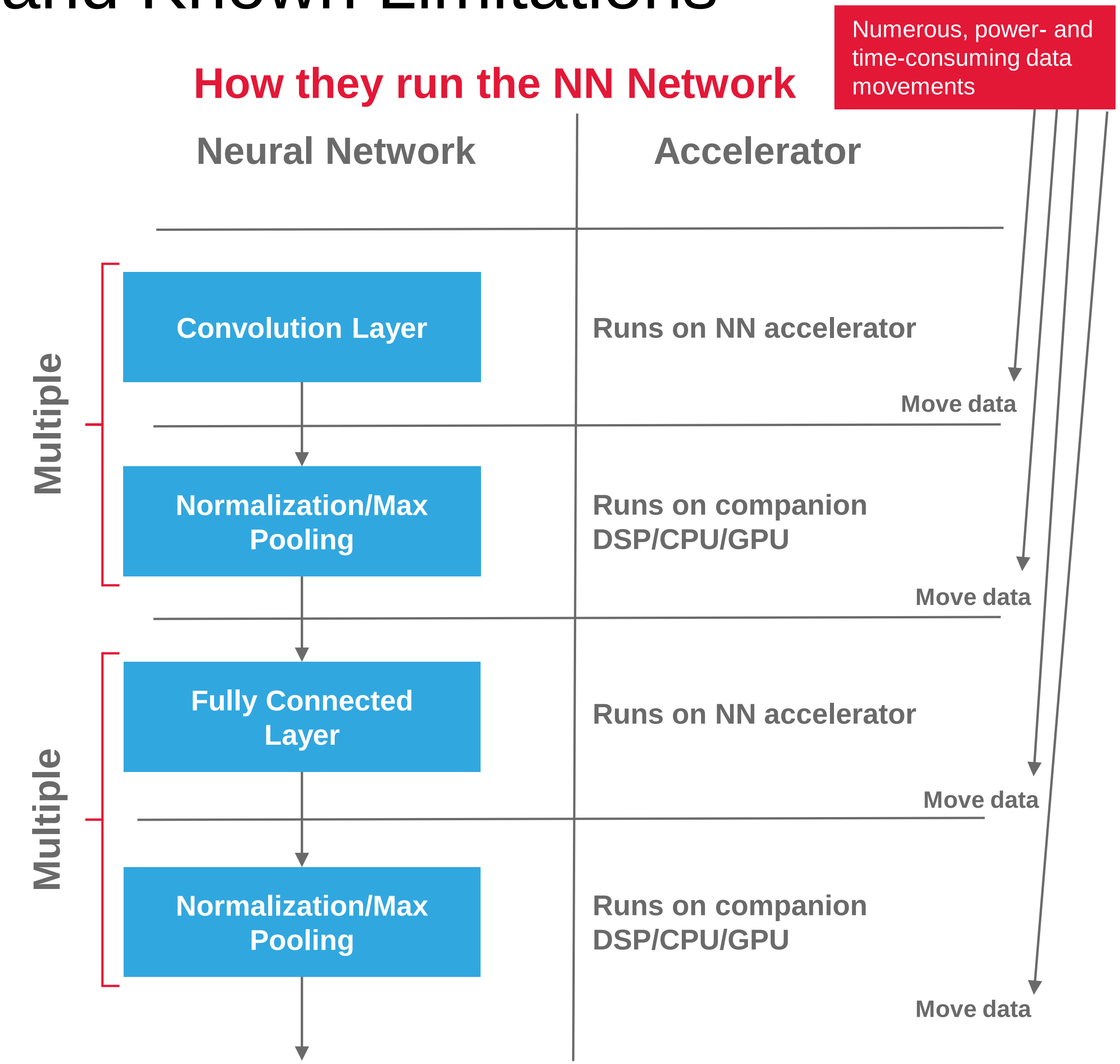
cadence

Current Alternatives for Implementing NNs in Embedded Systems

	CPUs	GPUs	Neural Network Hardware Accelerators	Imaging / Vision DSP (such as: Tensilica® Vision P6 DSP)
Ease of Development	Easy, pure SW, good tools, off-the-shelf IP	Easy, pure SW, good tools, off-the-shelf IP	Difficult, HW fixed at tapeout, SW must be partitioned between programmable core (CPU, GPU or DSP) and accelerator	Easy, pure SW, good tools, off-the-shelf IP
Power Efficiency	Poor	Better than CPU, but still poor	Great for the offloaded layers, not all layers offloaded, adds significant data movement overhead	Up to 10X better than GPU
Future Proofing	Yes, always reprogrammable	Yes, always reprogrammable	No, high risk since as NNs evolve, current accelerator choices will become a poor fit for future NN styles	Yes, always reprogrammable
Max NN Performance per Core (TMAC/s)	<200GFLOP	~200GFLOP	Up to 1TMAC	200-250GMAC

NN Accelerators: How They Work and Known Limitations





- Designed to offload/accelerate only convolution layers
- All other NN layers are run on an imaging DSP, control CPU or GPU
- Both DSP/CPU/GPU and NN accelerators are busy while running NN
- Excessive data movement between two processing elements
- Scales badly – to get 2X NN performance requires 2x (DSP + ACC)

Introducing Vision C5 DSP for Neural Networks

cādence®

Tensilica® Vision C5 DSP for Neural Networks

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

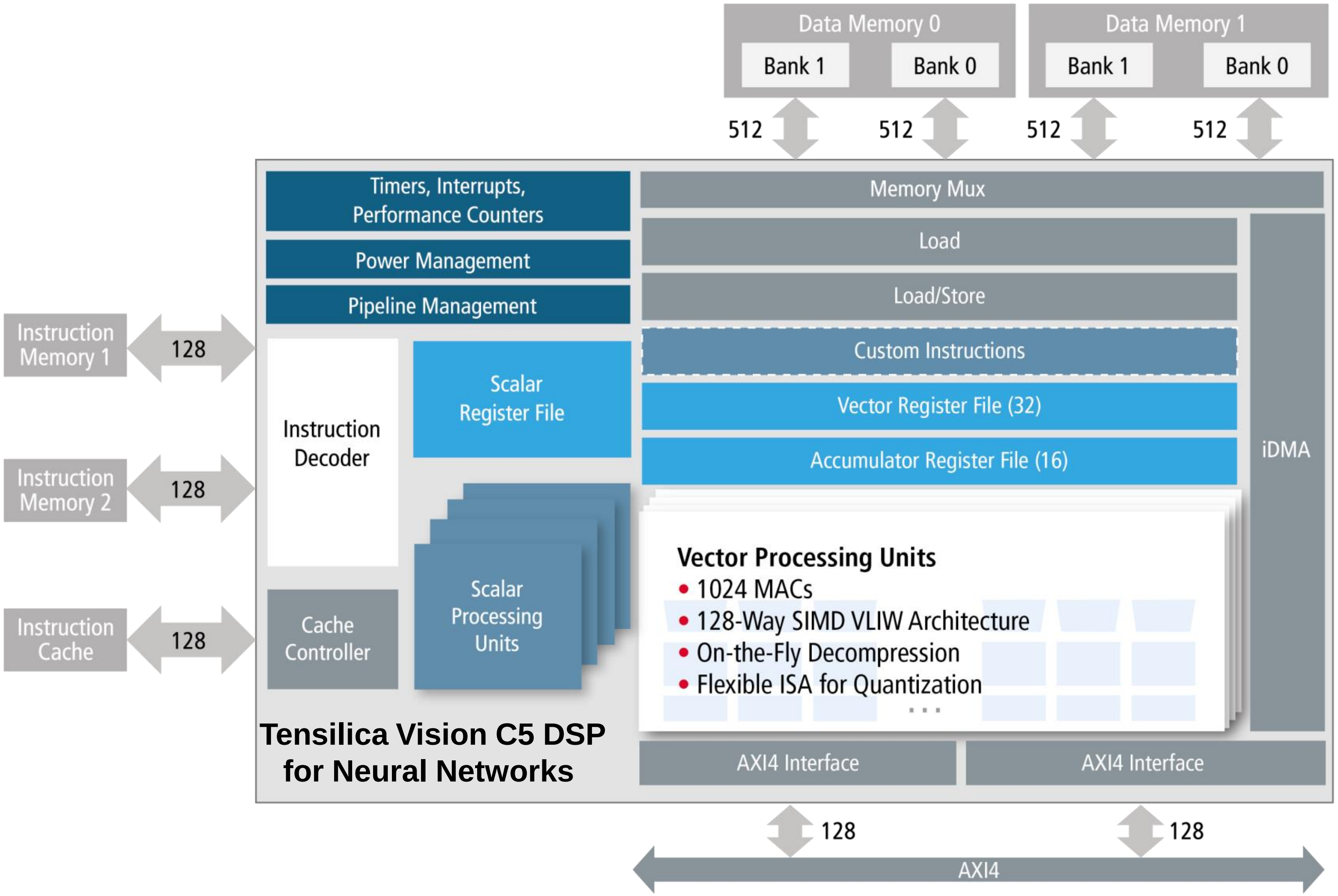
Building a DSP for changing NN field – general purpose and programmable

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

Tensilica® Vision C5 DSP for Neural Networks

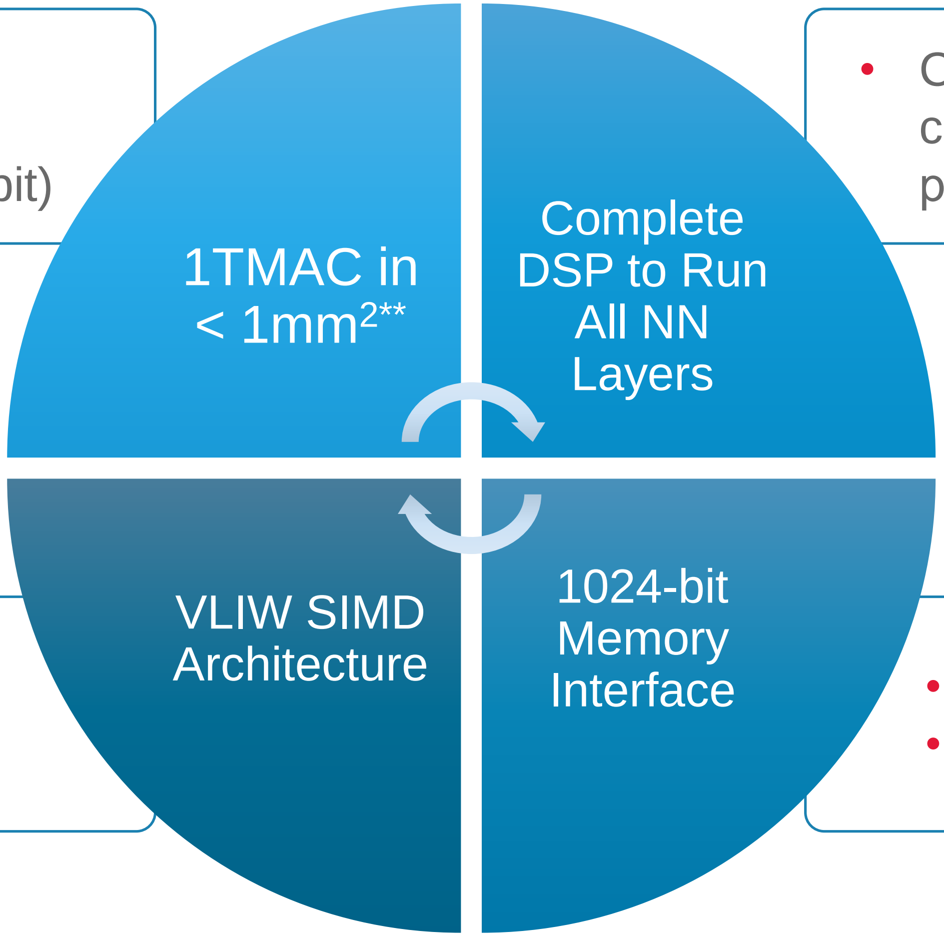


cadence®

Tensilica® Vision C5 DSP for Neural Networks

- 1024 MAC (8-bit)
  512 MAC (16-bit)

- Convolution, fully connected, pooling,…



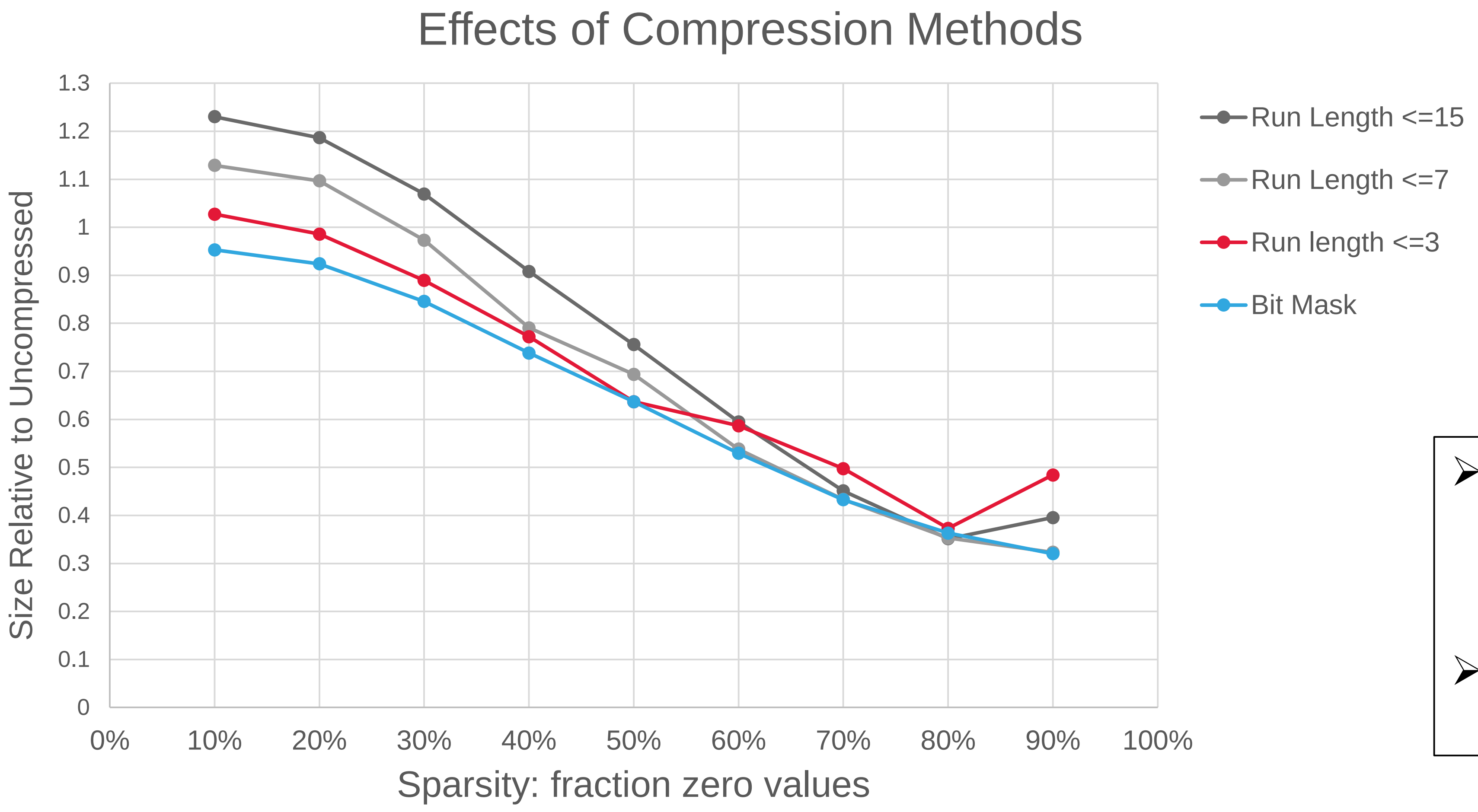
- 4-way VLIW
- 128-way 8-bit SIMD

- Integrated DMA
- Dual load/store

**16nm

Optimization for Sparsity – Coefficient Compression & Support for on the fly Decompression

Achieve 60% memory storage reduction @75% sparsity



- Compress Coefficient offline to save bandwidth
- Vision C5 support on the fly decompression

cādence®

Tensilica® Vision C5 DSP vs NN Accelerator

Vision C5 DSP	NN Accelerator
A complete processor that stands on its own: **Accelerates all NN layers**	Built to accelerate **only NN convolution functions**
Flexible and future-proof solution: • Supports variable kernel sizes, depths, input dimensions • Supports different compression/ decompression techniques • Support for new layers can be added as they evolve	HW accelerators are mostly designed based on current needs and hence provide a rigid and not future-proof solution
Main vision/imaging DSP free to run other applications while NN DSP runs NN	While running NN, **main vision/imaging DSP cannot run other applications**
Simple **(single-processing)** programming model for NN	Complicated **multi-processor** programming model
No need to move data between NN DSP and main vision/imaging DSP	**Need to move** data between NN DSP and main vision/imaging DSP (wastes power)

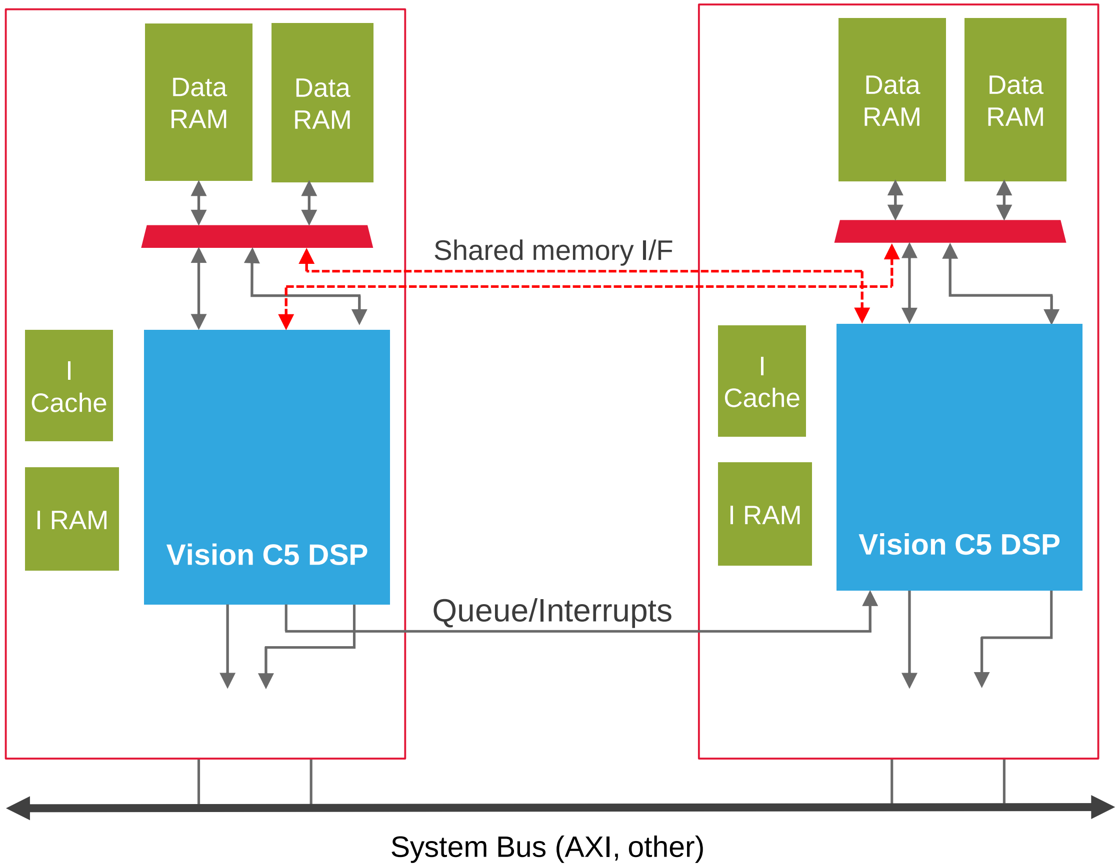
cādence®

Tensilica® Vision C5 DSP vs NN Accelerator

	Neural Network	Accelerator	Neural Network DSP
MULTIPLE	Convolution Layer	Runs on NN accelerator **Move Data**	Entire NN runs on Vision C5 DSP No extraneous data movement Does not occupy the imaging/vision DSP
	Normalization/ Max Pooling	Runs on companion imaging/vision DSP **Move Data**	
MULTIPLE	Fully Connected Layer	Runs on NN accelerator **Move Data**	
	Normalization/ Max Pooling	Runs on companion imaging/vision DSP **Move Data**	

cadence®

Tensilica® Vision C5 DSP Architected for Multi-Processor

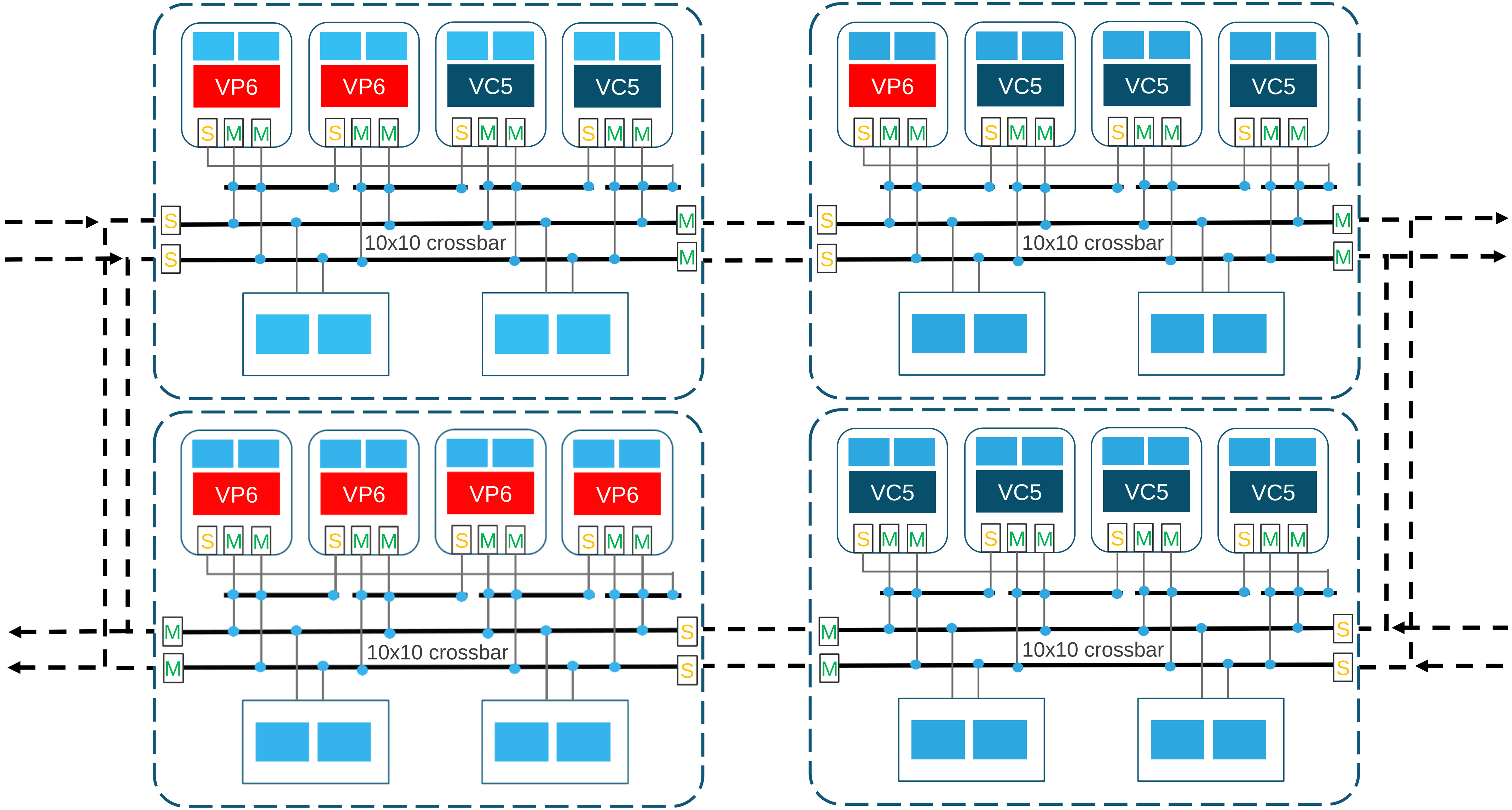


- Builds upon >17yrs of Xtensa® multi-processor experience
- Allows multi-TMAC/s solution
- Shared memory architecture
- Interrupts and queues for synchronization
- Automated creation of multi-processor SystemC® model
- Synchronous multi-processor debugging

Multi-core with shared memory I/F and queue/interrupts to synchronize

Scale Vision Sub-system Heterogeous Multi-core

Flexibility to customize MP cluster under the same programming model



cadence

Multi-core NN Load Partition Example

Split load across layer/batch/kernel

Operation:

$$F_O(x,y,n) = \sum_{z=0}^{Z-1}\sum_{r=-R}^{R}\sum_{r=-R}^{R} W(r,r,z) * FI(x+r,y+r,z)\ \forall x \in X, y \in Y, n \in N$$

Implementation as nested loops:

layers → ~~for (l=0; l<L; l++)~~

if(layers[l] == CONV) {

Ifmap batch → ~~for (b=0; b<B; b++)~~

kernels → ~~for (nu=0; nu<NU; nu++)~~

→ for (nv=0; nv<NV; nv++)

Ifmap height → for (y=0; y<Y; y+=S)

Ifmap width → for (x=0; x<X; x+=S)

SIMD vectorization in NV or X

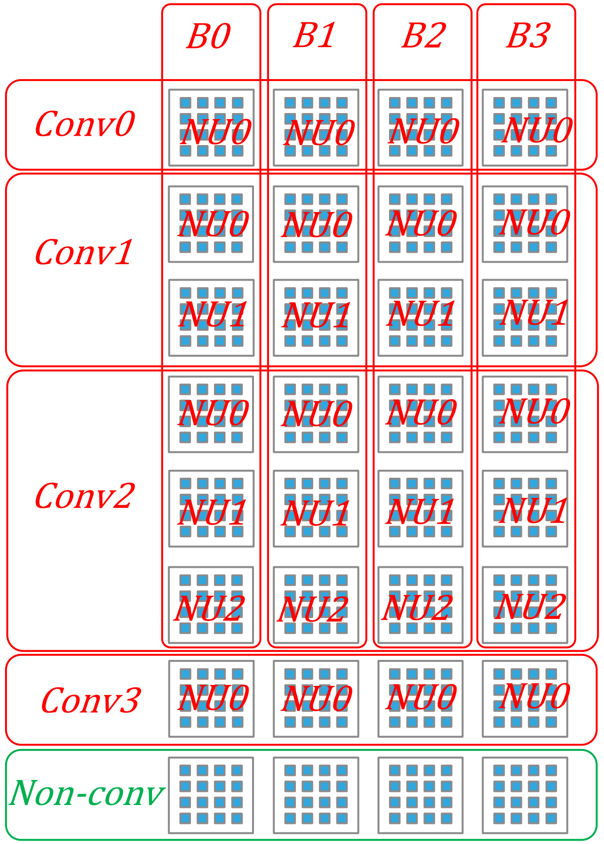
Ifmap channels → for (z=0; z<Z; z++)

Kernel width → for (ky=0; ky<KY; ky++)

Kernel height → for (kx=0; kx<KX; kx++)

$$\boldsymbol{F_O(x,y,n) \mathrel{+}= Wl_{b,n}(kx,ky,z) * FI(x+kx,y+ky,z)}$$

} #end if



- L and B loops are distributed across MP cores
- N is split into two loops, NU, NV and N = NU*NV
  - NU is distributed across cores
  - NV is handled by vectored SIMD

Cores are designated to certain layers and batches. Most efficient if layers can be load-balanced, no overlap in weight coefficients across cores.

cadence

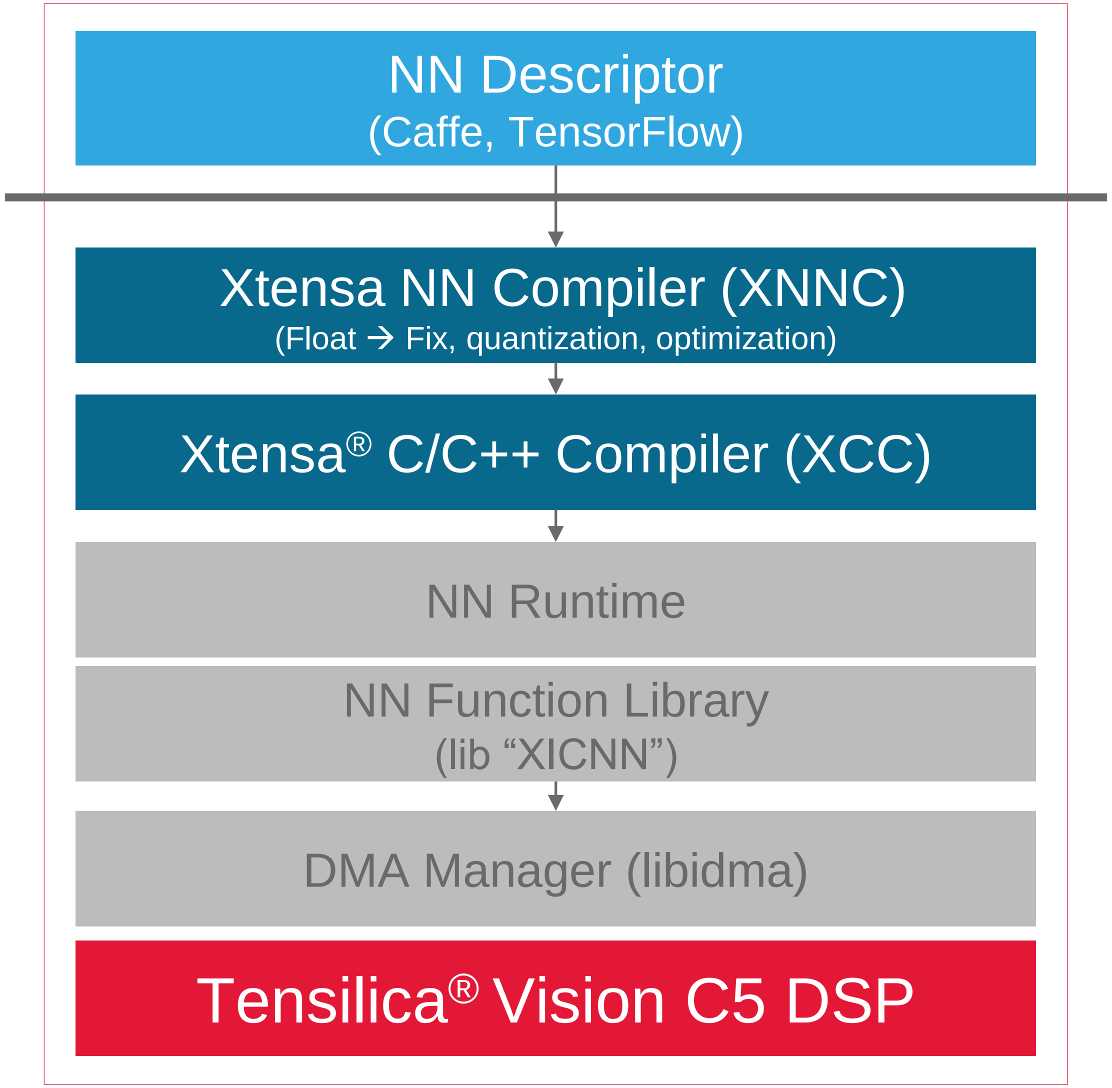
Vision C5 DSP vs Commercially Available GPUs

AlexNet
Performance up to
6X* faster

Inception V3
Performance up to
9X** faster

Note:
Faster is measure of cycle count requirements
* AlexNet data with 8 batch
** Inception V3 data with single batch

Automated Software Flow for Various NN Frameworks



- User Code
- Cadence® Compiler / Tool
- Cadence SW Library / Runtime
- Tensilica Vision C5 DSP

Xtensa Neural Network Compiler (XNNC)

- Push button solution to generate code for NN from Caffe or TensorFlow
- Hand optimized library to get maximum performance for each CNN functions

cadence®

Vision C5 DSP – Preferred Solution for NNs in Embedded Systems

	CPUs	GPUs	Neural Network Hardware Accelerators	Imaging/Vision DSP (such as: Tensilica® Vision P6 DSP)	Vision C5 DSP
Ease of Development	Easy, pure SW, good tools, off-the-shelf IP	Easy, pure SW, good tools, off-the-shelf IP	Difficult, HW fixed at tapeout, SW must be partitioned between programmable core (CPU, GPU or DSP) and accelerator	Easy, pure SW, good tools, off-the-shelf IP	Easy. Pure SW. Good tools
Power Efficiency	Poor	Better than CPU, but still poor	Great for the offloaded layers, not all layers offloaded, adds significant data movement overhead	Up to 10X better than GPU	Optimized for NN. No wasted HW. No wasted data movement
Future Proofing	Yes, always reprogrammable	Yes, always reprogrammable	No, high risk since as NNs evolve, current accelerator choices will become a poor fit for future NN styles	Yes, always reprogrammable	Yes. Always reprogrammable
Max NN Performance per Core (TMAC/s)	<200GFLOP	~200GFLOP	Up to 1TMAC	200-250GMAC	1TMAC

cadence®

Vision DSP Partner Ecosystem (Public)

Morpho
- WDR (wide dynamic range)
- Super video image stabilization

ArcSoft
- Live Beautify
- HDR / Low-light Enhance
- Facial Recognition
- Dual-camera Solutions

uurmi SYSTEMS creating waves for tomorrow
- ADAS suite
- Fog removal, object detection
- System integrator

sensory
- Face and voice authentication
- Face detection

KHRONOS GROUP CONNECTING SOFTWARE TO SILICON

OpenVX

Cadence Chair of OpenVX WG at Khronos Group

MULTICORE WARE
- CNN neural networking
- Imaging algorithm expertise

Almalence
- Super-resolution zoom, HDR
- Camera processing

IRIDA LABS
- Imaging and vision experts
- Low light enhancement
- Advanced noise reduction
- Face detection

cādence®

Tensilica® Vision C5 and Vision P6 DSPs:
Cadence Addressing All Market Segments

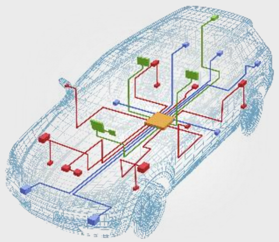
Processing Power

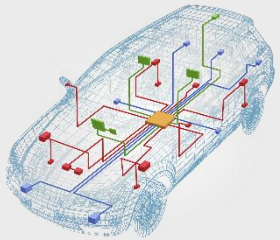
Up to 10TMAC/sec

1TMAC/sec

<200 GMAC/sec

Segment	DSP	
Automotive (towards autonomous)	Multiple Vision C5 DSPs	**Runs multiple NNs all the time**
Surveillance / Automotive (semi-autonomous)	Vision C5 DSP	**Runs a couple of NNs all the time**
Mobile	Vision P6 DSP	**Runs a NN once in a while**





cadence®

Summary

Cadence® Tensilica® Vision C5 DSP for neural networks

- Not an “accelerator”—industry’s complete DSP designed for CNN to run all neural network layers
- 1 TeraMAC/sec computational capacity in less than 1mm^2
- General purpose and programmable to meet evolving requirements
- Optimized for vision, radar/lidar and fused-sensor applications with high-availability (always-on) neural network (NN) computational needs
- Architected for multi-processor design—scales to multi-TMAC/sec solution
- Targeted at surveillance, automotive, drone and mobile/wearable markets

cādence®