Tensilica DNA 100 Processor:

A High-Performance, Power-Efficient DNN Processor for On-Device Inference

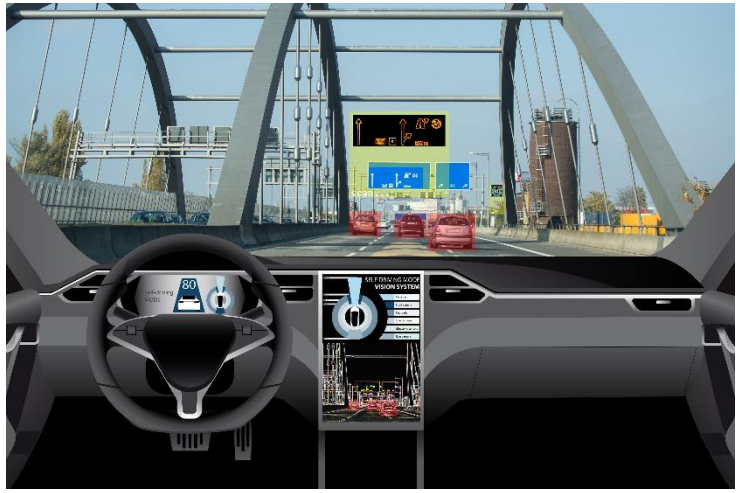
Megha Daga

Sr Manager, Product Marketing & Management, AI, Cadence IP Group

Feb 17, 2019

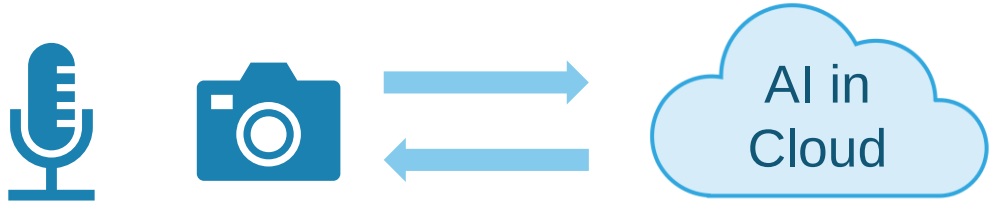
cadence®

AI Experiences are Everywhere



"*Alexa, Add a 2pm meeting to my calendar*"

Majority of AI Inferences Are in the Cloud



"*Alexa, when is my new camera arriving?*"

Smart Assistant
Voice search

Travel Assistant
Translation

Navigation Assistant
Store finder

On-Device AI – Why?

Low Latency requirements

- Natural dialogue in speech assistants - requires less than 200 ms latency
- Real time decision making in Automotive, Robots, etc. need low latency

Lack of Good Connectivity

- Smart City cameras are hard to connect to existing network
- Inspection drones for wind turbines and power lines operate in rural areas

Privacy

- Smart home video cameras and smart assistants – consumers desire privacy
  - Baby monitors, home security cameras – can devices generate alerts?
  - Voice assistants – keyword detection today → small vocabulary recognition?

On-Device AI Processing Needs Are Increasing

Mobile

- On-device AI experiences like face detection and people recognition at video capture rates

AR/VR headsets

- On-device AI for object detection, people recognition, gesture recognition, and eye tracking

Surveillance cameras

- On-device AI for family or stranger recognition and anomaly detection

Drones and robots

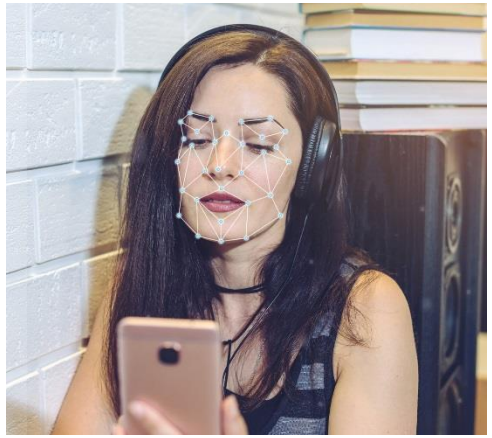
- On-device AI to recognize subjects, objects, obstacles, emotions, etc.

Automotive

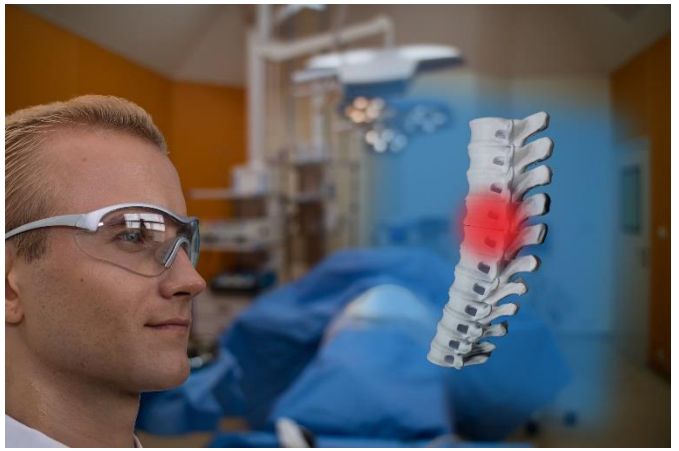
- On-device AI to recognize pedestrians, cars, signs, lanes, driver alertness, etc. for ADAS and AV

Target Markets for On-Device AI Inferencing

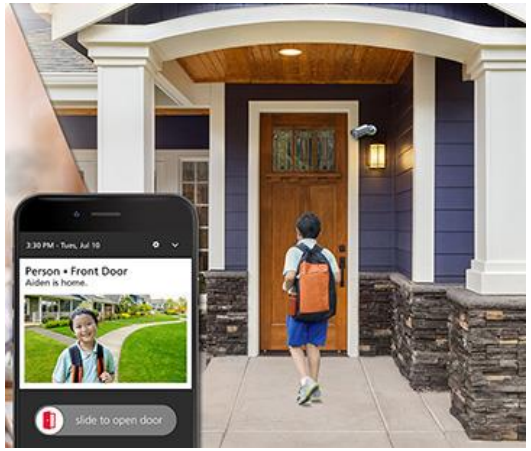
IoT
< 0.5TMAC



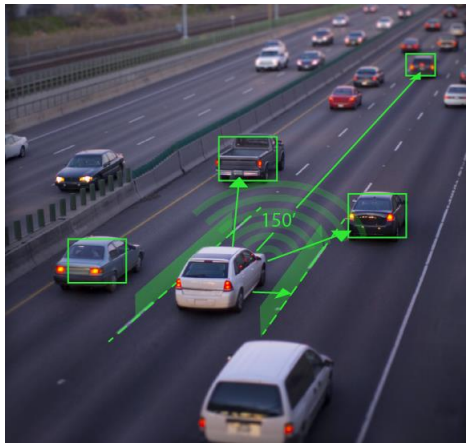
Mobile
0.5 - 2TMACs



AR/VR
1 - 4TMACs



Smart Surveillance
2 - 10TMACs

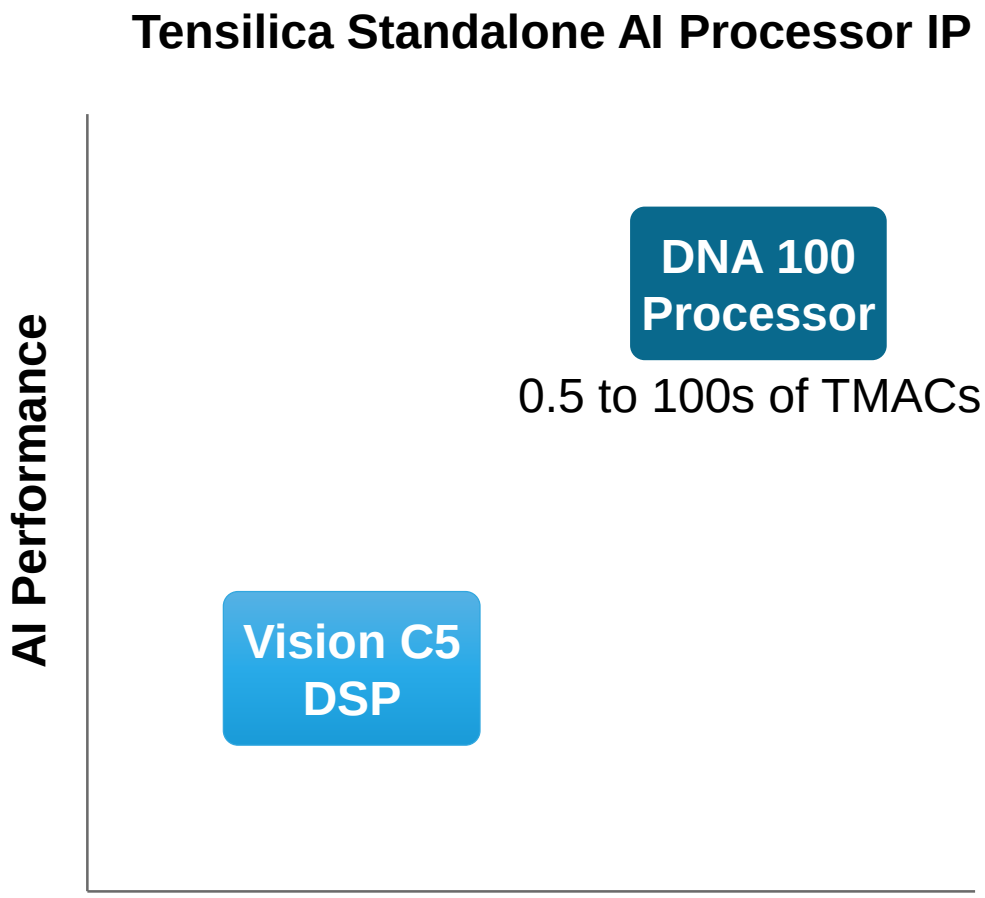


Autonomous Vehicles
10s - 100s TMACs

cadence

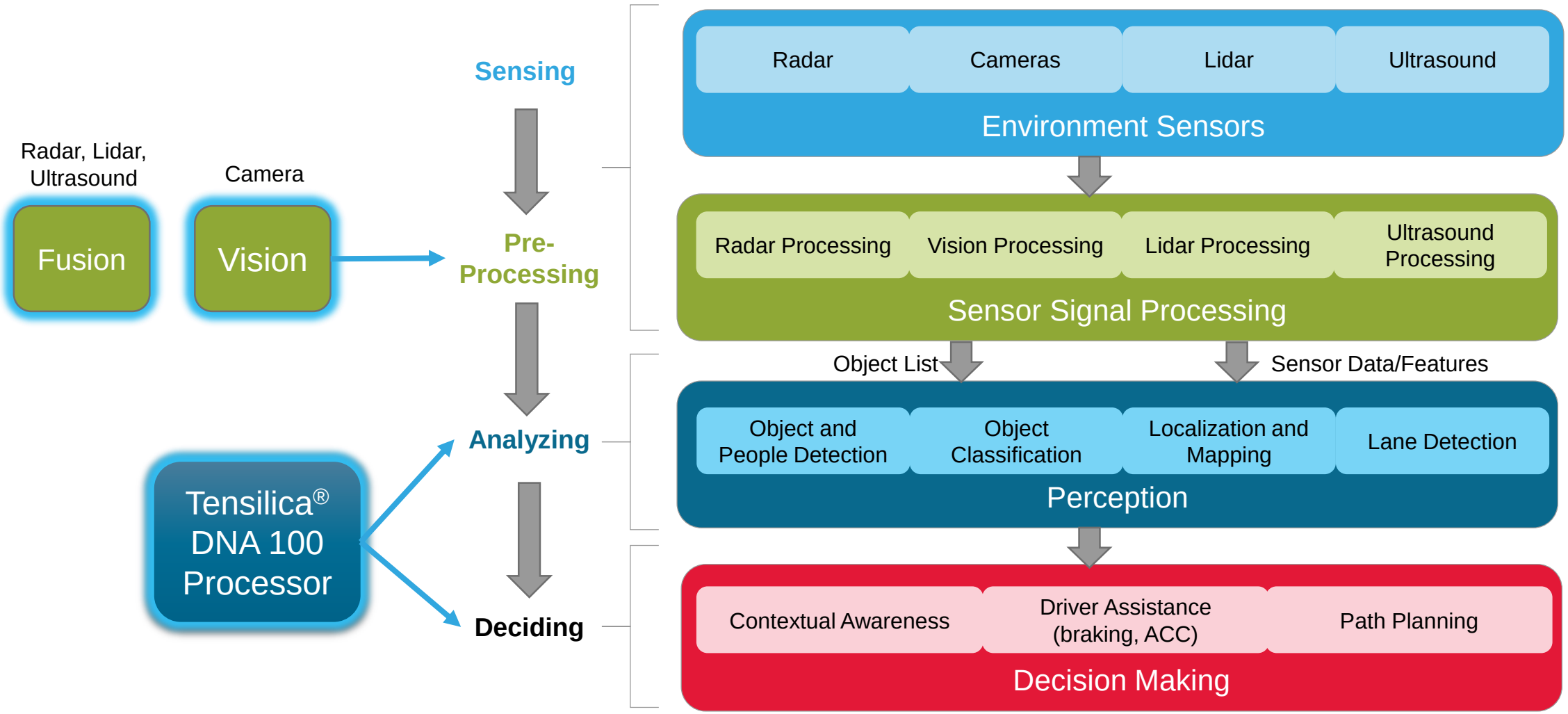
Tensilica DNA 100 Processor IP for AI

Tensilica® Application-Specific DSPs + **Tensilica AI DSPs**



Use Case: AI in Automotive

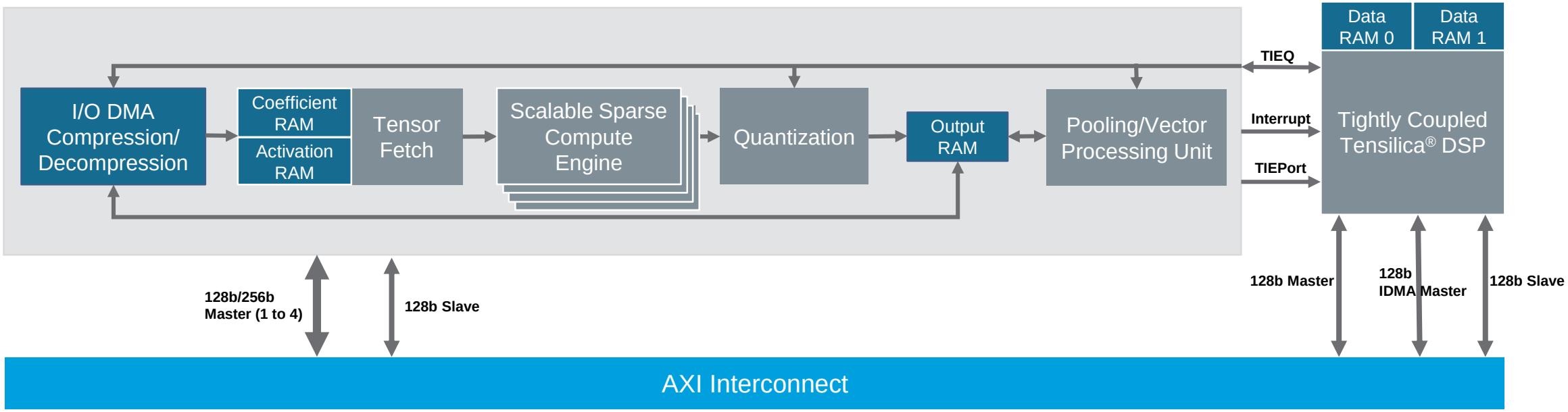
Perception and decision-making with cameras, radar, lidar, and ultrasound



cadence®

Tensilica DNA 100 Processor Block Diagram

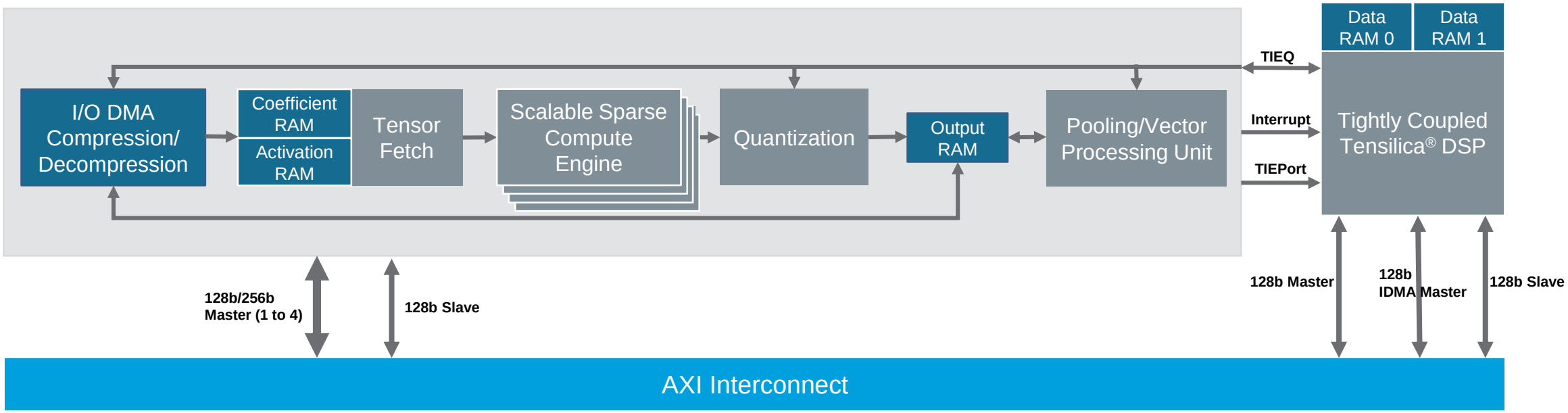
Sparse compute engine with high MAC utilization



cadence®

Tensilica DNA 100 Processor Block Diagram

Sparse compute engine with high MAC utilization

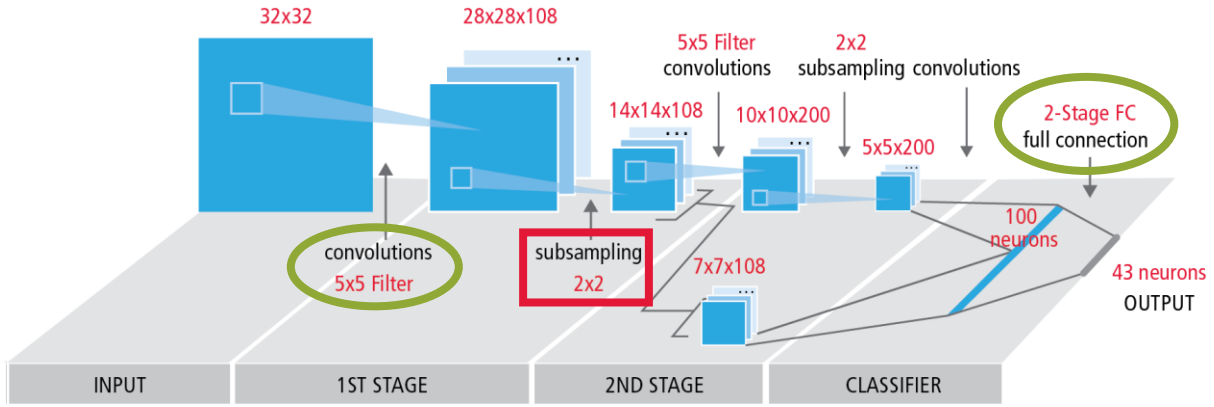


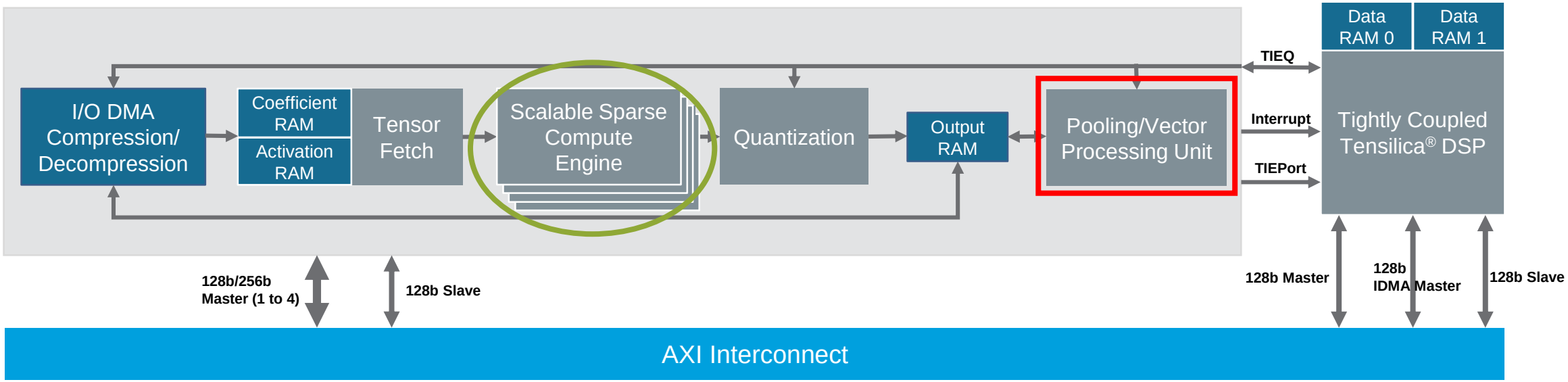
- ✓ Bandwidth reduction
  - Weight compression
  - Activation compression
- ✓ Compute reduction
  - Non-zero MAC operations only
- ✓ Efficient convolution and fully connected
  - High MAC occupancy rate
- ✓ Non-convolution layers
  - Pooling
  - Sigmoid, tanh…
  - Elementwise add, sub…
- ✓ Programmable
  - Future proof
- ✓ Extensible
  - Custom layers

Tensilica DNA 100 supports key requirements for on-device AI processing

Neural Network Mapping onto Tensilica DNA 100 Processor

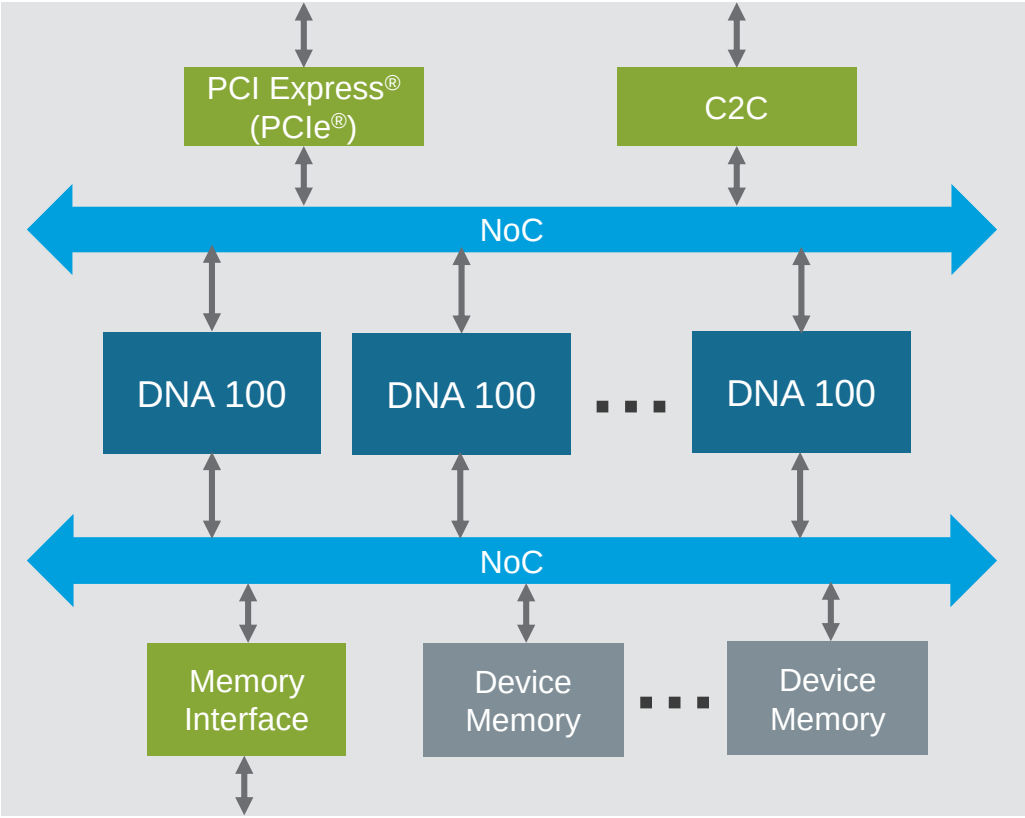
An example



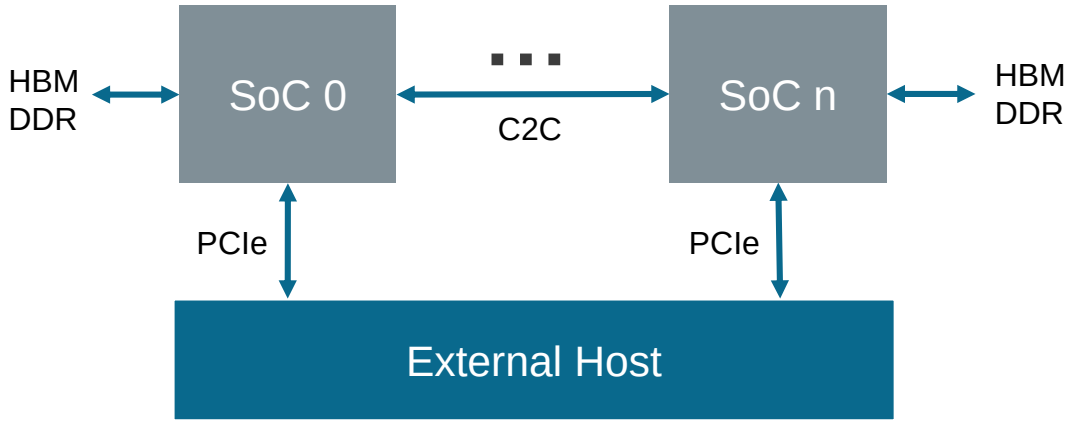


cadence®

SoC Concept: Scaling to 100s of Effective 8b TMAC



- Array of Tensilica® DNA 100 cores are subject to area and power targets



- Chip to Chip (C2C) link used to scale beyond a single chip

Effective MAC on Tensilica® DNA 100 Processor

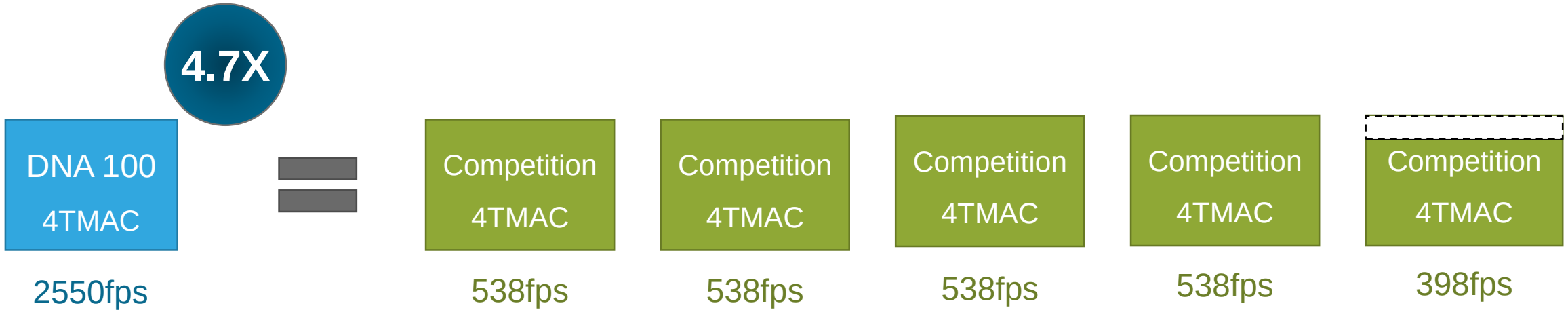
Enabled by sparse compute engine

Physical Array Size	Effective TMAC or Performance @ 1GHz			
	No Network Pruning[1]		With Network Pruning[2]	
256MAC	2X	0.5TMAC	3X	0.75TMAC
1K MAC		2TMAC		3TMAC
4K MAC		8TMAC		12TMAC

1 For 15% sparse weights and 50% sparse activation
2 For 35% sparse weights and 60% sparse activation
Network pruning and re-training provides higher sparsity

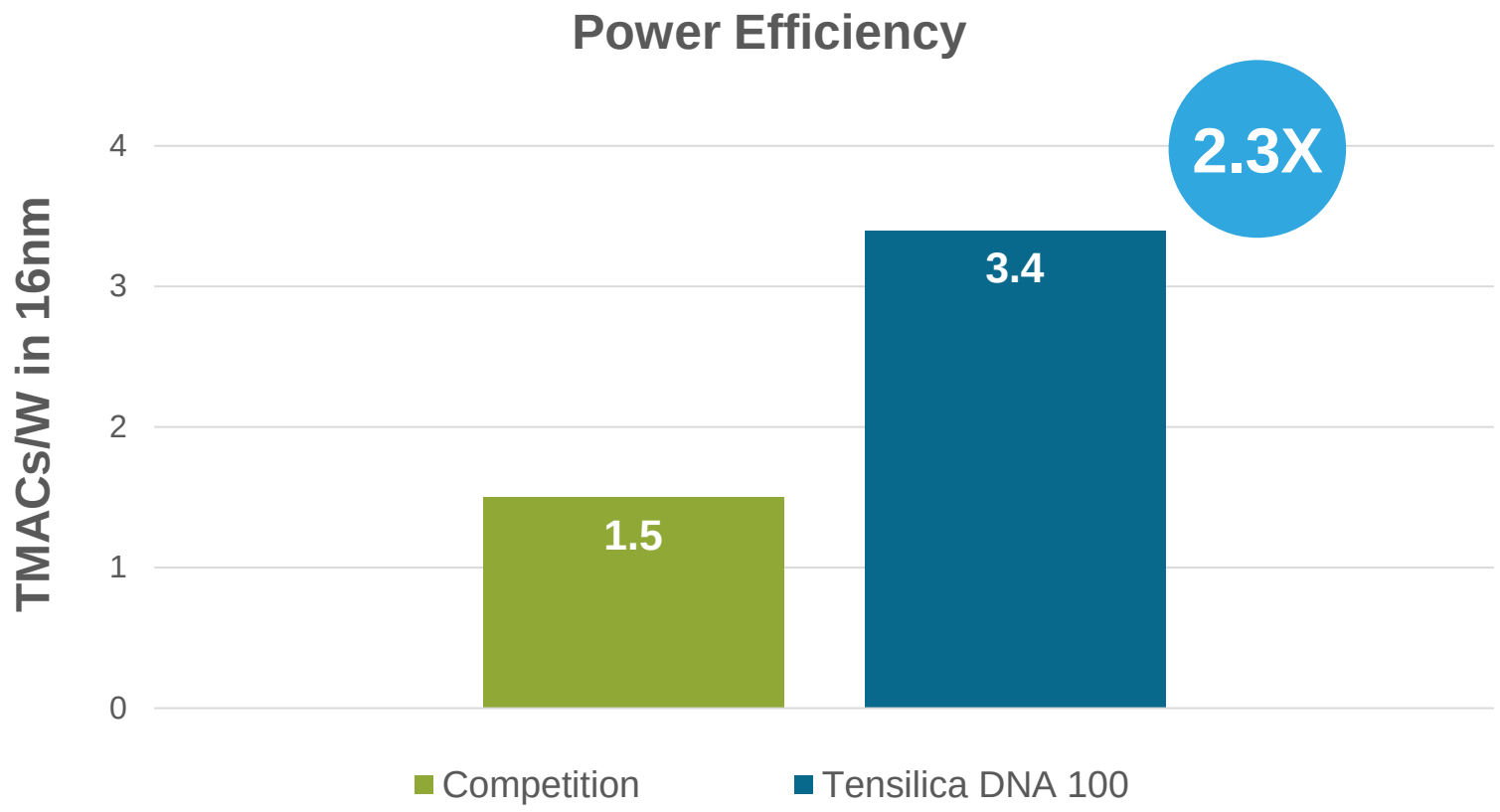
Tensilica DNA 100 Processor Performance – Up to 4.7X Competition

Enabled by sparse compute and high MAC utilization



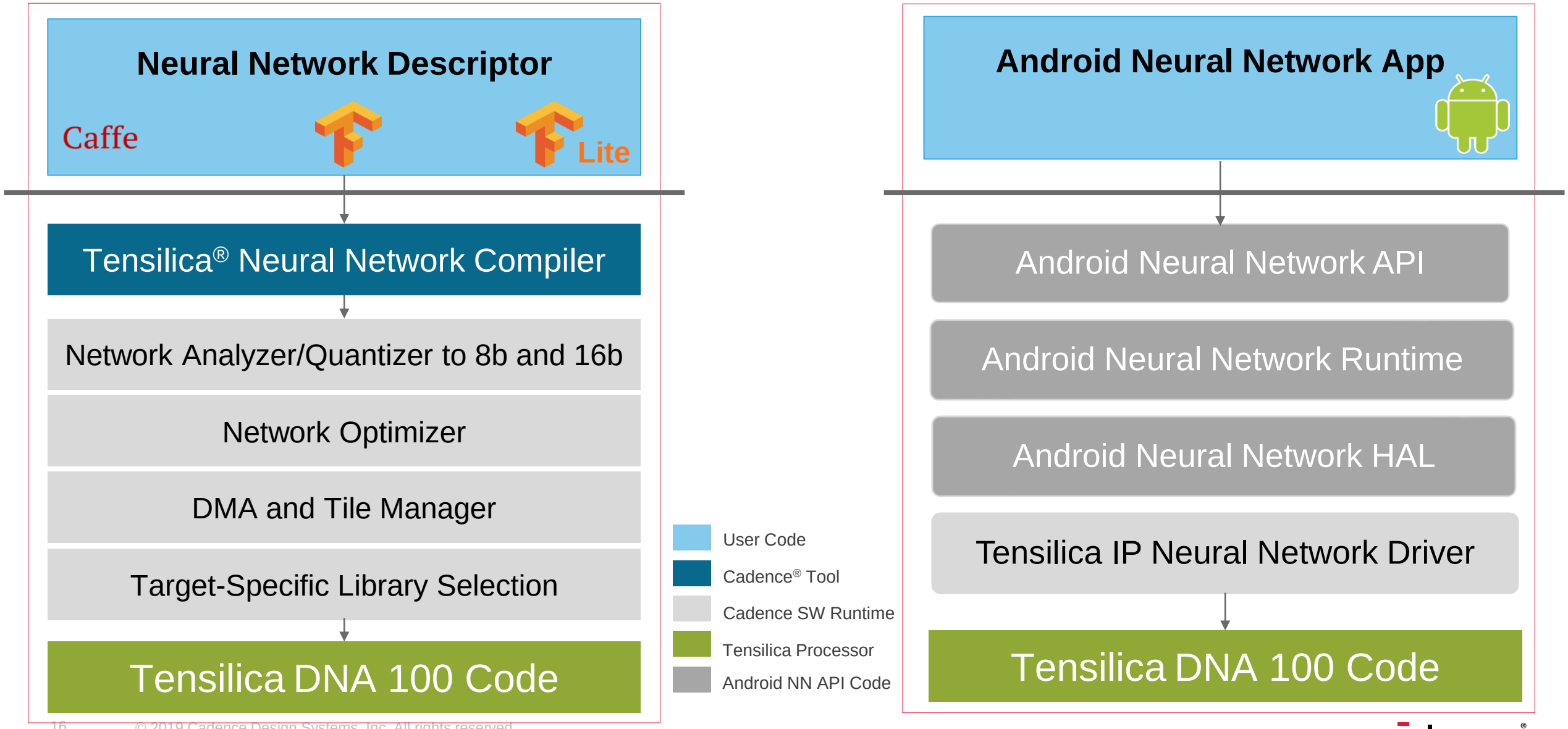
- Tensilica® DNA 100 processor performance on ResNet50 @ 1GHz is 2550 frames per second (fps)
- Tensilica DNA 100 processor and competition are both 4KMAC physical array configuration
- Tensilica DNA 100 processor numbers are with network pruning
  - Assuming 35% sparse weights and 60% sparse activation

Tensilica DNA 100 Power Efficiency – Up to 2.3X Competition



*Tensilica® DNA 100 processor is with network pruning for 4TMAC configuration

AI Software Platform



Cadence Tensilica® Supports Facebook's Glow

POSTED ON SEP 13, 2018 TO AI RESEARCH, ML APPLICATIONS

Glow: A community-driven approach to AI infrastructure


VIJAY RAO

NADAV ROTEM

Today, we are announcing the next steps in Facebook's efforts to build a hardware ecosystem for machine learning (ML) through partner support of the Glow compiler. We're pleased to announce that Cadence, Esperanto, Intel, Marvell, and Qualcomm Technologies Inc, a subsidiary of Qualcomm Incorporated, have committed to supporting Glow in future silicon products.

Integrating Facebook's Glow, an open-source machine learning compiler based on LLVM, to enable *a modular, robust, and easily extensible approach*

https://code.fb.com/ml-applications/glow-a-community-driven-approach-to-ai-infrastructure/

Tensilica® DNA 100 Processor

Industry-leading performance and power efficiency

Up to 4.7X performance for similar array sizes

Enabled by sparse compute and high MAC utilization

Up to 2.3X power efficiency for similar array sizes

Up to 3.4 TMACs/W*

Complete AI software platform and strong partner ecosystem

*In 16nm

cadence®