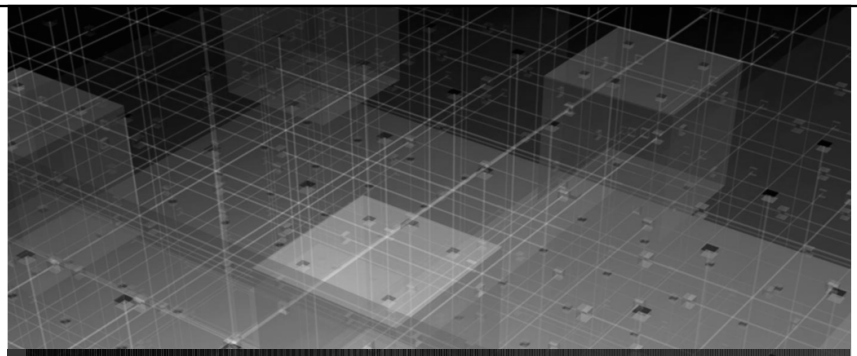


Pushing the Energy-Efficiency Envelope for "Always Alert" Sensor Nodes

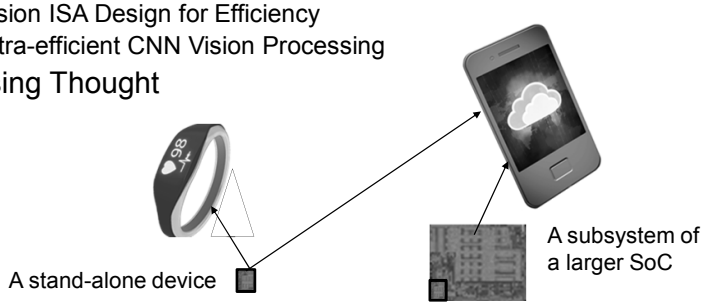
Chris Rowen
CTO – IP Group
Cadence Design Systems, Inc.

MPSOC July 2015

cadence

Agenda

- Always Alert: sensing the real world
- Energy and layered cognition
- Feeling and Hearing
  - Tensilica Fusion DSP
  - Ultra Low Energy
- Always Alert Vision
  - Vision ISA Design for Efficiency
  - Ultra-efficient CNN Vision Processing
- Closing Thought



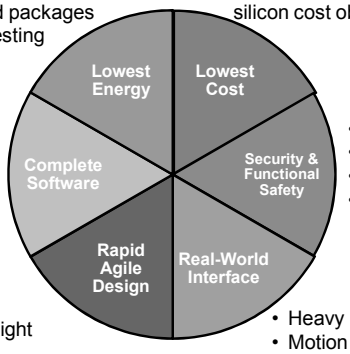
cadence

cadence

Key design challenges for **Always Alert** nodes

- Long life
- Limited battery capacity
- Silent, sealed packages
- Energy harvesting

- Large volumes + consumer targets = silicon cost obsession



- Mandatory interoperability with wireless, cloud and apps interfaces
- Rich experiences

- Sensitive data
- Wide deployment
- Wireless access
- Control of physical systems

- Rapid innovation and tight market windows
- Continuous software update

- Heavy [analog] data gathering
- Motion + sound + video
- Lots of MEMS
- Small form-factor packaging

cadence

Energy and layered cognition

Active Power

>> 1Watt
"What is Madonna's biggest hit?" — Cloud application processing

50mW-1000mW
"Play 'Like a Virgin'" — System application processing

1mW - 50 mW
"Turn up volume" — Subsystem event processing

1-1000 uW
"Hello audio" — Always Alert -detect activity of interest

nW
Quiescent system

Cloud server farm

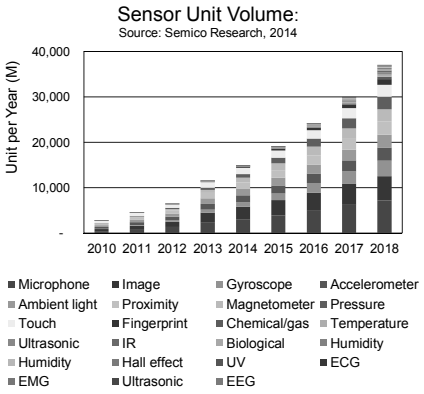
Application Filter

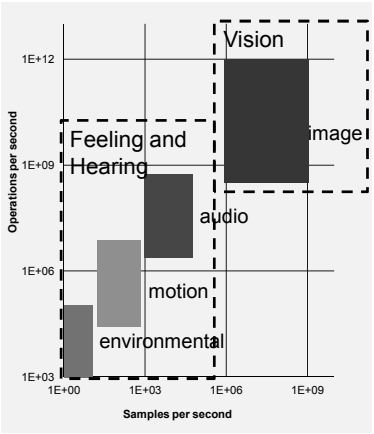
Always Alert

cadence

cadence

Sensing the real world
The evolution of a rich electronic ecosystem





cadence

The Always-On Design Problem

Sensing | **Computation** | **Communications**

- Three distinct functions lumped together. Processors cores not typically suited for all 3 functions
- Microcontrollers have limited DSP for wireless comms
- Signal processing is significant and growing in complexity over time – especially from sensor fusion
- Energy efficiency is extremely important
- For ease of use – one core efficiently spans all functions

cadence

cadence

Introducing Tensilica Fusion DSP

Ultra-low power processing for mobile and wearable applications

Leading low power DSP performance	• Lower power for always on • Efficient floating point support for sensor fusion • Quad MAC performance for narrowband
Configurability – Get the right core immediately	• Multiple configuration options allow core to be configured exactly for the targeted applications • No waste, maximum efficiency
Comprehensive SW and ecosystem	• SW compatible with HiFi. 70+ partners. 140+ SW packages • Comms ecosystem support • Optimized DSP library with fixed and float kernels • Ease of use for SW developers

7

cadence

Typical Tensilica Fusion Processing Applications

Mobile, Wearables and IoT

Sensing	• Sensor Fusion • Biometric monitoring • Pedestrian Dead Reckoning
Wake-Up Processing	• Voice Trigger • Face Trigger • Gesture Trigger
Audio Voice Speech	• Speech Recognition • Speech Pre-Processing • Audio Playback
Comms (Narrowband Wireless)	• BLE • LTE CAT 0 • WIFI 802.11n, 802.11ah • GNSS

Common Sensors

- Accelerometer Magnetometer
- Gyroscope Barometer
- Ambient Light
- Optical Proximity
- Microphone

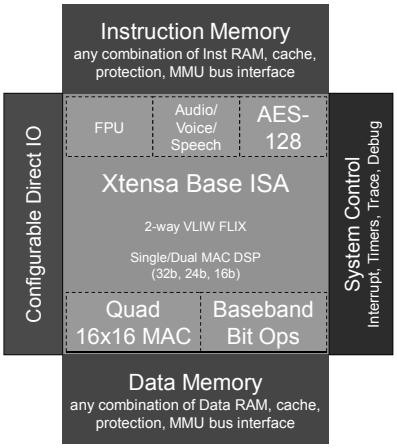
8

cadence

cadence

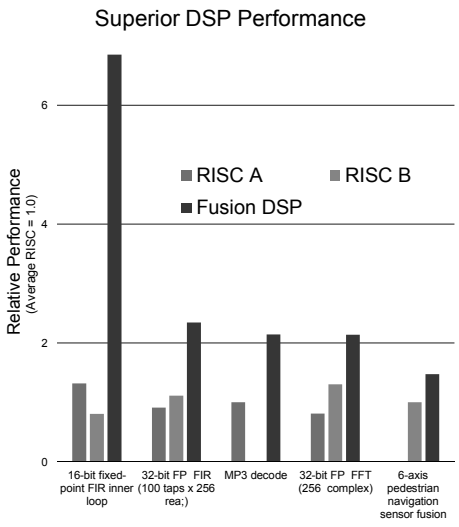
Tensilica Fusion Configuration Options
Click to configure

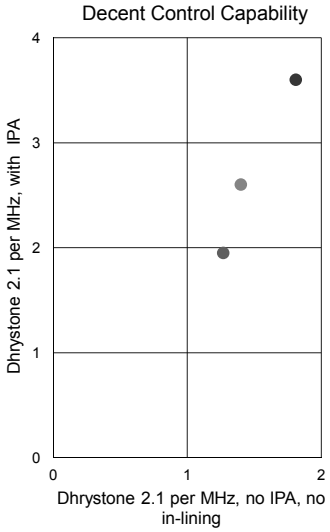
- Single-precision FPU
  - Floating Point instructions issued concurrently with 64-bit load/store
  - Speeds S/W porting
- AVS (Audio/Voice/Speech)
  - SW compatibility with HiFi 3 Audio DSP
  - Access to 125+ HiFi Audio/Voice software packages
- Quad 16x16 MAC
  - Accelerates communications standards like BLE/Wi-Fi
  - Accelerates voice algorithm performance
- BLE/Wi-Fi AES-128
  - Encryption acceleration for wireless
- Baseband bit operations
  - Accelerates performance of bit operations (CRC, convolutional coding, interleave) for implementation of Baseband MAC/PHY



cadence

Tensilica Fusion benchmarks



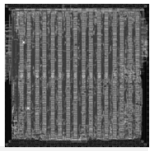


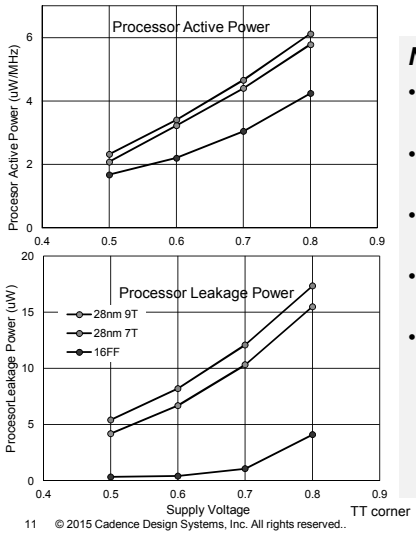
cadence

cadence

IOT processing and ultra-low-energy

Xtensa core scales from ~1uW/MHz to 1.5GHz





Near-threshold core development

- *SPICE (SPECTRE APS) simulation of full Tensilica base 32b processor core*
- *Verifies correctness, speed and power independent of library characterization.*
- *Full parasitics extracted circuit gives accurate power and functionality*
- *Running basic power and functionality diagnostics*
- *Optimally-pipelined core runs >1.5GHz in16FF, but same RTL also achieves micro-power level. Sustains high MHz at lower voltage*

cadence

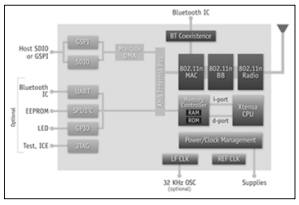
Hundreds of millions of Xtensa cores already shipped in Always Alert devices

EPSON
EXCEED YOUR VISION

GPS watches extend battery life from 12 to 30 hours

Qualcomm
ATHEROS

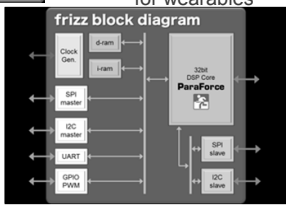
Qualcomm IoE devices (QCA4002, 4004) powered by Xtensa® DPUs running AllJoyn announced September 2013.



MegaChips

frizz

Xtensa-based *ParaForce* DSP for wearables

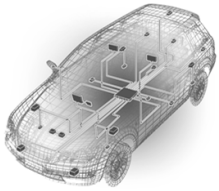


Audience/Knowles is the leader in advanced voice, audio and motion sensor processing for mobile devices.

Audience

cadence

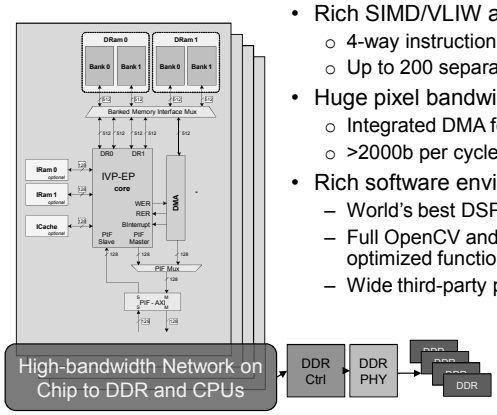
Always Alert Vision



- Imaging and video represents virtually all IoT content by data volume
- Unique computing, power and bandwidth challenges
- Rapid evolution of new applications driving new algorithms and architectures

cadence

Vision Processors

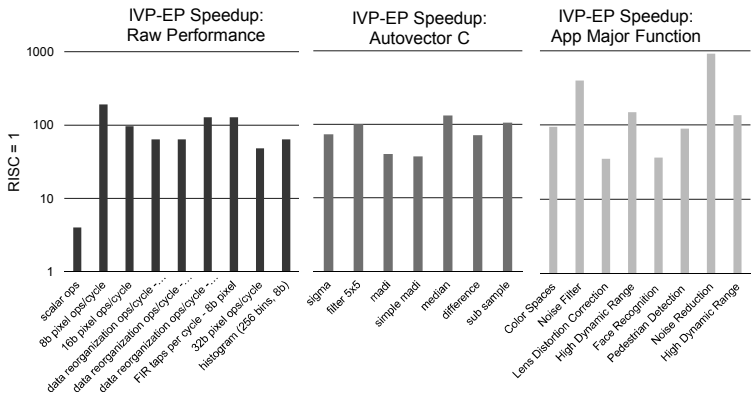


- A family of high-performance DSPs for imaging, video and vision
- Rich SIMD/VLIW architecture
  - 4-way instruction issue
  - Up to 200 separate ALU operations per cycle
- Huge pixel bandwidth
  - Integrated DMA for data streaming
  - >2000b per cycle data memory bandwidth
- Rich software environment
  - World's best DSP C compilers: 0 assembly code
  - Full OpenCV and OpenVX support with 800 optimized functions
  - Wide third-party program and support

cadence

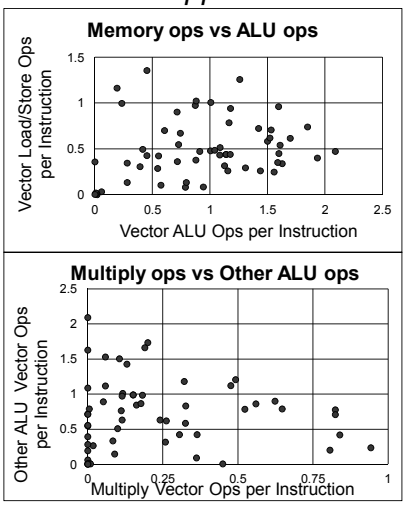
cadence

Vision Processor Family
A new category of application performance



cadence

Application Diversity drives Vision ISA flexibility
based on 45 application kernels

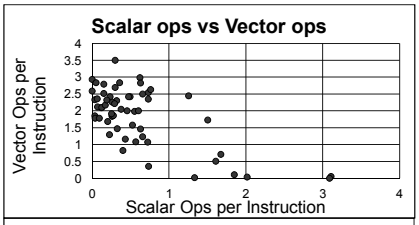


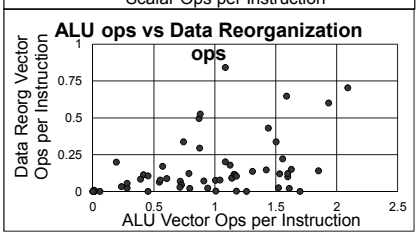
- Typically several ALU ops per load operation
- Wide range of ALU : Load/store ratio (1:2 to 5:1)
- Many important functions don't do multiplies
- A fraction have very heavy multiply usage – e.g. convolutions
- ISA should handle wide range of ratios efficiently

cadence

cadence

Application Diversity drives ISA flexibility

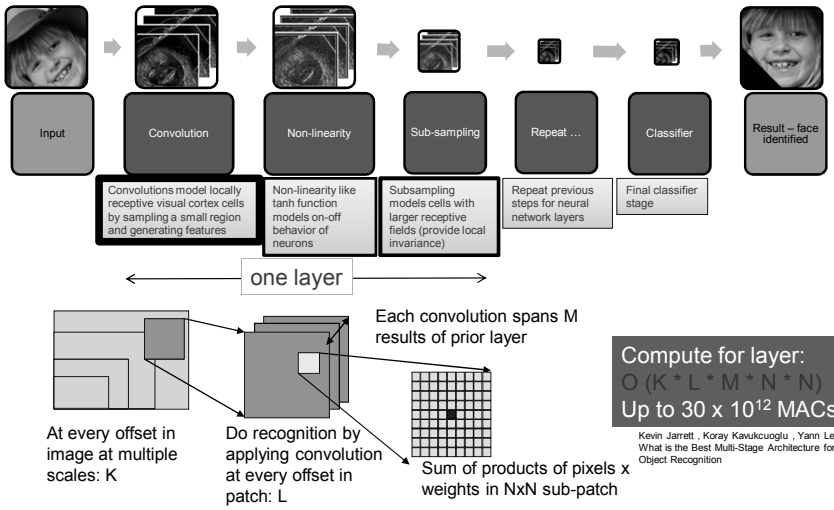




- A successful architecture maximizes the fraction of kernels that can be vectorized
- A small number of functions may still use scalar ops heavily
- On-the-fly data reorganization may be important in a few kernels
- ALU : Reorg ratio varies from 10:1 to 1:1
- Efficient data reorganization boosts benefit of vectorization

cadence

Example: Convolutional Neural Network (CNN)
Up to 30T Multiply-adds per VGA frame



Compute for layer:
O (K * L * M * N * N)
Up to 30 x 10^{12} MACs

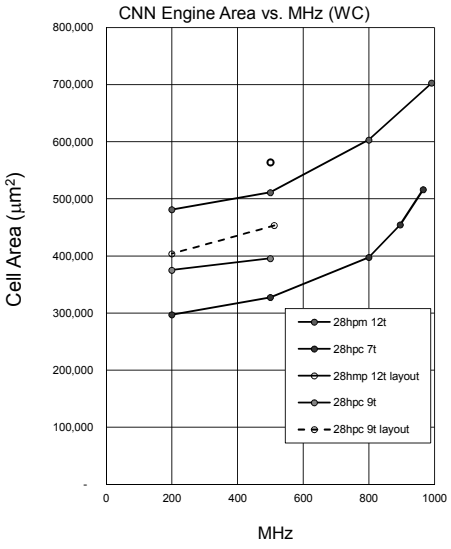
Kevin Jarrett , Koray Kavukcuoglu , Yann Lecun
What is the Best Multi-Stage Architecture for Object Recognition

cadence

cadence

Specialized Vision Processing:
CNN Engine Example

- Compute-intensive specialized DSP architecture using TIE on Xtensa
- Very high throughput (256 MAC/cycle) on filter and convolution-intensive algorithms for vision and object recognition applications
- High memory bandwidth to handle huge stream of 16b weights
- Architecture: SIMD/VLIW architecture with 47 general-purpose DSP vector ops on top of 80 base ops
- Free TIE source distribution: Built to be freely adapted and evolved by customers
- Total floorplan area with memory: 0.8 – 1.0 mm^2
- Scalable multi-core

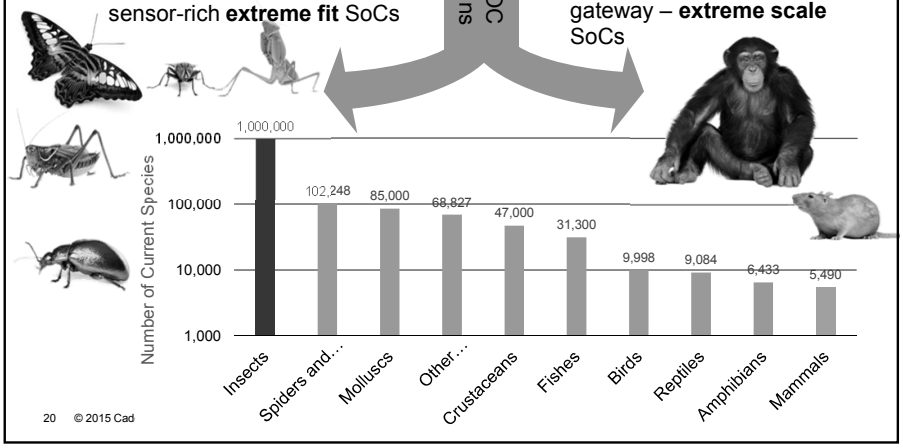


A closing thought
Specialization ➔ proliferation of designs

Specialized low-cost, low-energy platforms: wearables, smart-home, medical, industrial, sensor-rich **extreme fit** SoCs

All SOC Designs

General-purpose all-in-one system platforms: cellphone, PC, server, gateway – **extreme scale** SoCs



cadence®

cadence®

