

Heterogenous Multicore Design – an Always-On Use Case

Pat Harmon, Director of Engineering
Synopsys ARC[®] Processor Summit 2022



Agenda

- Always-on AIoT Applications Introduction
- Typical Implementation Requirements
- SoC Architecture & Use Case
- Programming Model, Software and Tools
- Synopsys - One-stop Shop
- Key Take-Aways

Always-on AIoT

Typical application domains

Always Monitoring



- Heart rate monitor
- Human activity recognition
- ...

Always Listening



- Smart speakers
- Smart phones
- ...

Always Watching



- Face trigger
- Gesture recognition
- ...

- Processing requirements vary with
 - Application domain
 - Input data rate
 - Model complexity

User Expectations

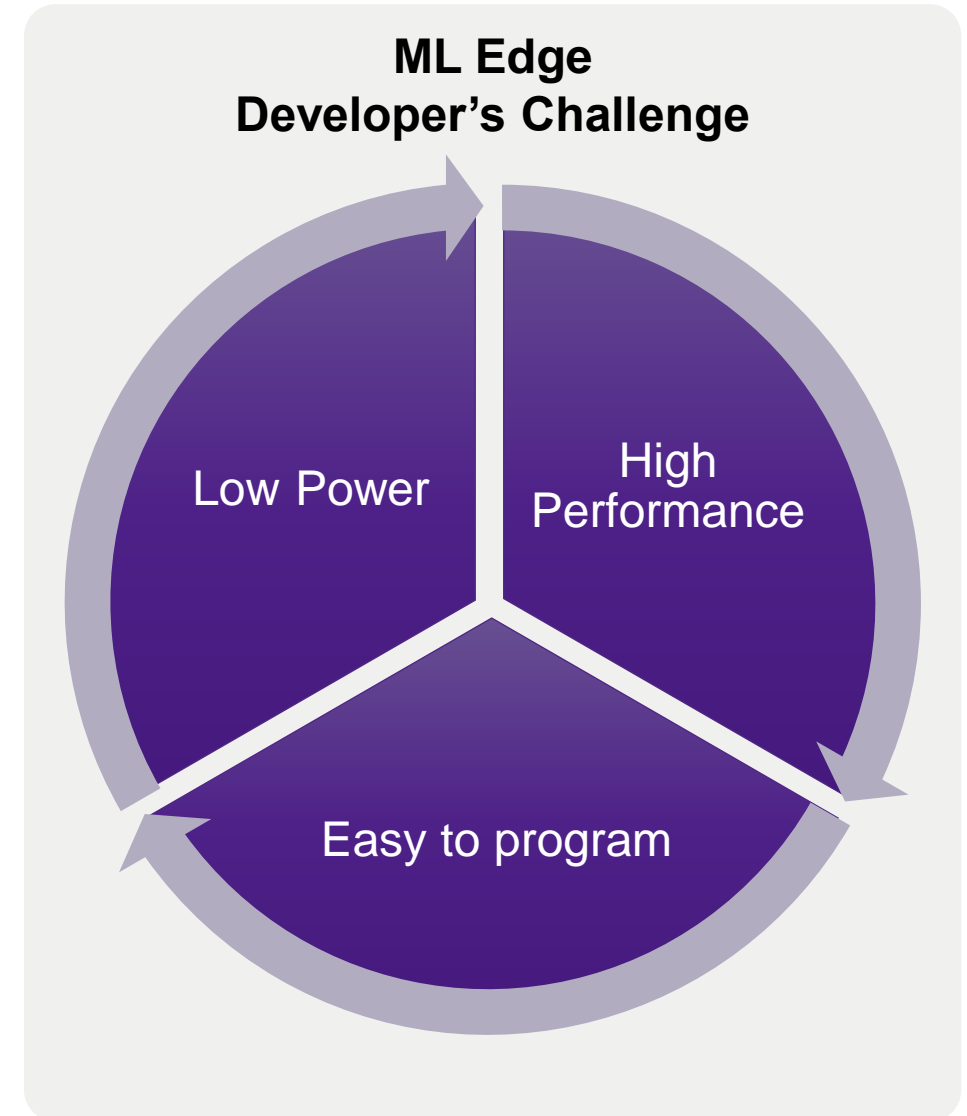
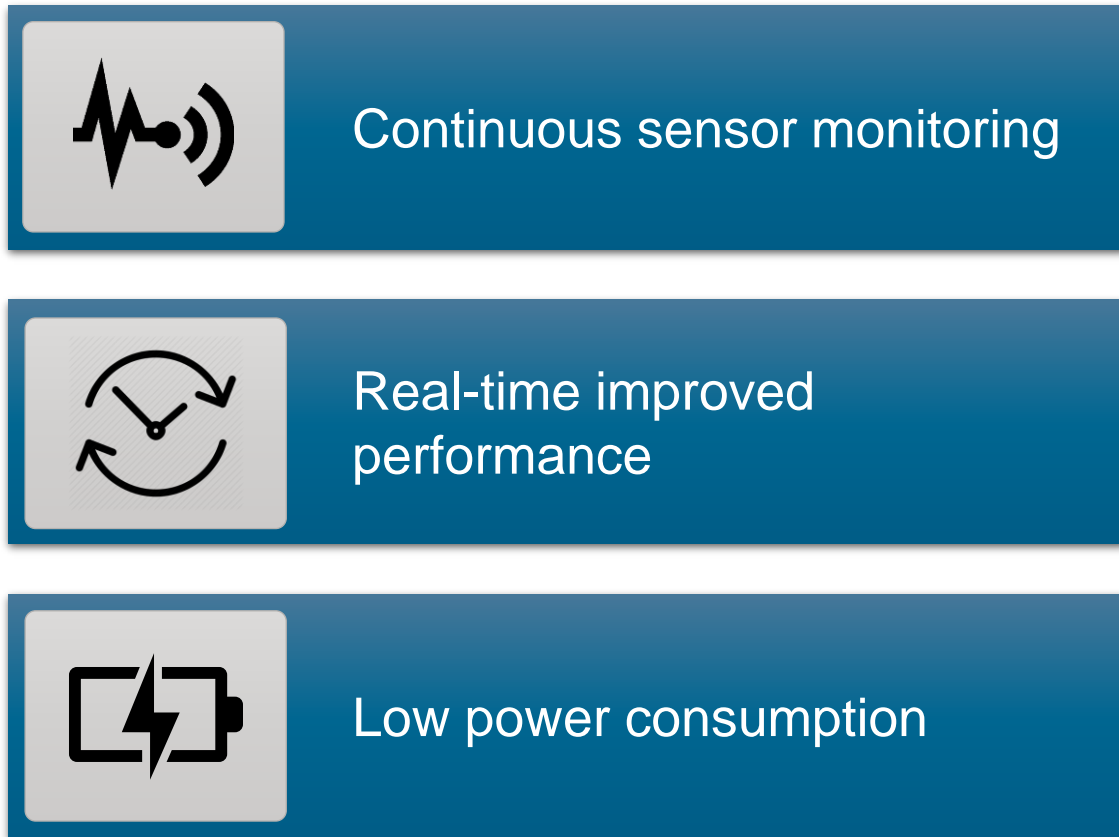
- More interest in vision-based systems in deeply-embedded environments
- Example Applications
 - Gaze detection to turn on a TV
 - Smart appliances
 - Laptop wakeup
 - Smartphone wakeup
 - Remote Surveillance
- Processing needs are increasing... but users still expect
 - Responsive systems
 - Long battery life
 - Improved Performance



Agenda

- Always-on AIoT Applications Introduction
- **Typical Implementation Requirements**
- SoC Architecture & Use Case
- Programming Model, Software and Tools
- Synopsys - One-stop Shop
- Key Take-Aways

Deeply-Embedded AIoT Systems Must Achieve...



AON Application looking at Machine Learning Inference

Wide range of processing requirements

- Processing requirements of machine learning inference can differ by orders of magnitude
- Key factors impacting processing requirements are the input data rate and the complexity of the trained model

Machine Learning Application	Input data rate	Complexity of trained model
Human activity recognition	10s Hz (few sensors)	Low to medium
Voice control	10s kHz (e.g. 16 kHz)	Low to medium
Face detection	100s kHz (low resolution and frame rate)	Low to medium
Natural speech recognition	10s kHz	High
Mid-range image processing (RADAR, LiDAR, thermal camera, ..)	Up to 100s MHz	Low to medium
Point cloud processing (RADAR, LiDAR)	Up to several MHz	Medium to high
Advanced computer vision	100s MHz (high resolution and frame rate)	High

DesignWare ARC Processor IP

Unrivalled Efficiency for Embedded Applications



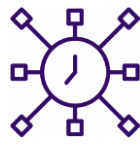
EM Family

- Optimized for ultra low power IoT
- 3-stage pipeline w/ high efficiency DSP
- Power as low as 3uW/ MHz
- Area as small as 0.01mm² in 28HPM



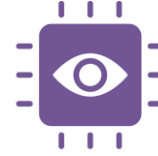
SEM Family

- Security processors for IoT and mobile, including DSP
- Protection against HW, SW, and side channel attacks
- SecureShield for Trusted Execution Environments



HS Family

- Highest performing CPUs, CPU + DSP
- 32- & 64-bit ISAs
- High-speed 10-stage pipeline
- SMP Linux support
- Single- and multi-core configurations



EV Family

- Heterogeneous multicore for vision and AI processing
- DNN (Deep Neural Network) Engine
- High productivity, standards-based tool suite



VPX Family

- High performance vector DSP
- SIMD/VLIW design for massive parallel processing
- Multiple vector FP engines for high precision results



NPX Family

- Scalable neural processor units
- Up to 250 TOPS (440 TOPS with sparsity)
- Supports latest AI applications
- High productivity, standards-based tool suite

Functional Safety (FS) Processors



- Integrated hardware safety features for ARC EM, SEM, HS, VPX, EV and NPX processor families
- Accelerates ISO 26262 certification for safety-critical automotive SoCs

AON Application looking at Machine Learning Inference

Wide range of processing requirements

- Processing requirements of machine learning inference can differ by orders of magnitude
- Key factors impacting processing requirements are the input data rate and the complexity of the trained model

Machine Learning Application	Input data rate	Complexity of trained model	
Human activity recognition	10s Hz (few sensors)	Low to medium	} ARC EM / HS VPX
Voice control	10s kHz (e.g. 16 kHz)	Low to medium	
Face detection	100s kHz (low resolution and frame rate)	Low to medium	
Natural speech recognition	10s kHz	High	} EV7x } NPUx
Mid-range image processing (RADAR, LiDAR, thermal camera, ..)	Up to 100s MHz	Low to medium	
Point cloud processing (RADAR, LiDAR)	Up to several MHz	Medium to high	
Advanced computer vision	100s MHz (high resolution and frame rate)	High	

Typical Implementation Requirements

AON System Blocks

- Low-power IoT applications, battery-powered, always-on
 - Vision (face detection, face recognition, ..)
 - Voice (voice trigger, ..)
 - Radar (in-room, detection / tracking / gesture recognition)
 - NN inference (INT8)

Typical Requirements	Synopsys Solution
<p>Two Power Modes domains to maximize efficiency and reduced power consumption:</p> <ul style="list-style-type: none"> - Ultra-low-power always-on / detection mode (Always-On domain) <ul style="list-style-type: none"> • Face detection (5-10 fps, 100x100 resolution) / voice trigger • Power budget 1 mW - Performance mode (High throughput domain) <ul style="list-style-type: none"> • Vision (20-40 fps, ~500x500 resolution), Radar, .. • 500-800 MHz 	<p>ARC IP Portfolio Unrivaled Efficiency for Embedded Applications</p> <p>ARC EMxD - Ultra-low-power processor with advanced DSP capabilities for AON monitoring, wake detect and/or person detection</p> <p>ARC VPX High-Performance DSP for processing-intensive applications ARC NPX Neural Processing Unit with industry's highest performance</p>
<p>Tight Integration of System Level Components</p>	<ul style="list-style-type: none"> • MIPI CSI-2 Host Controller for camera interconnection (Vision AON) • DMA, UART, GPIO, I3C, WDT, Timers • AMBA components
<p>Fast Time-to-Market</p>	<ul style="list-style-type: none"> • Synopsys One-Stop Solution

Agenda

- Always-on AIoT Applications Introduction
- Typical Requirements
- **SoC Architecture & Use Case**
- Programming Model, Software and Tools
- Synopsys - One-stop Shop
- Key Take-Aways

Requirements driving Architectural modeling and analysis

Requirements



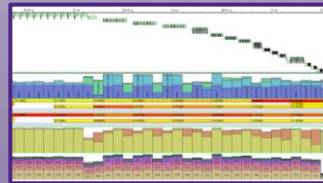
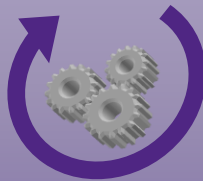
- Hardware and Software
- **Performance, Low Power**
- Security, Safety, Test, Memory, ...

Analytic Modeling



- Analysis of the kernels in application
- Conceptual mapping on multicore architecture
- Conceptual memory allocation and dataflow

Exploration

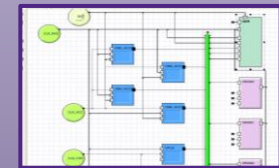


- Simulation and Analysis
- Iterative refinement of design parameters in hardware and software

PA Ultra Modeling



software



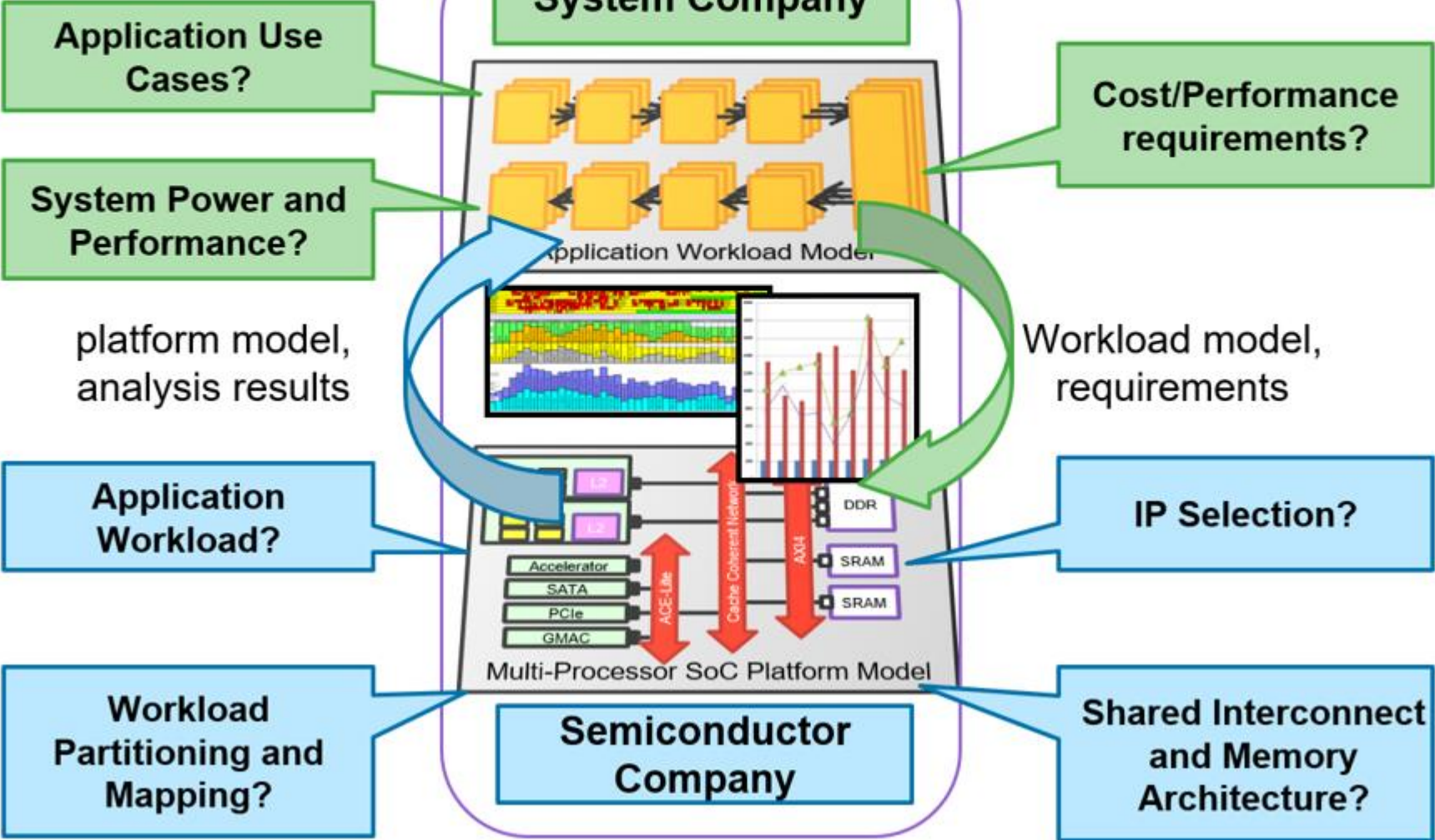
hardware

- Parameterized model of the software
- Parameterized model of the multicore architecture
- Mapping of software- on the hardware model

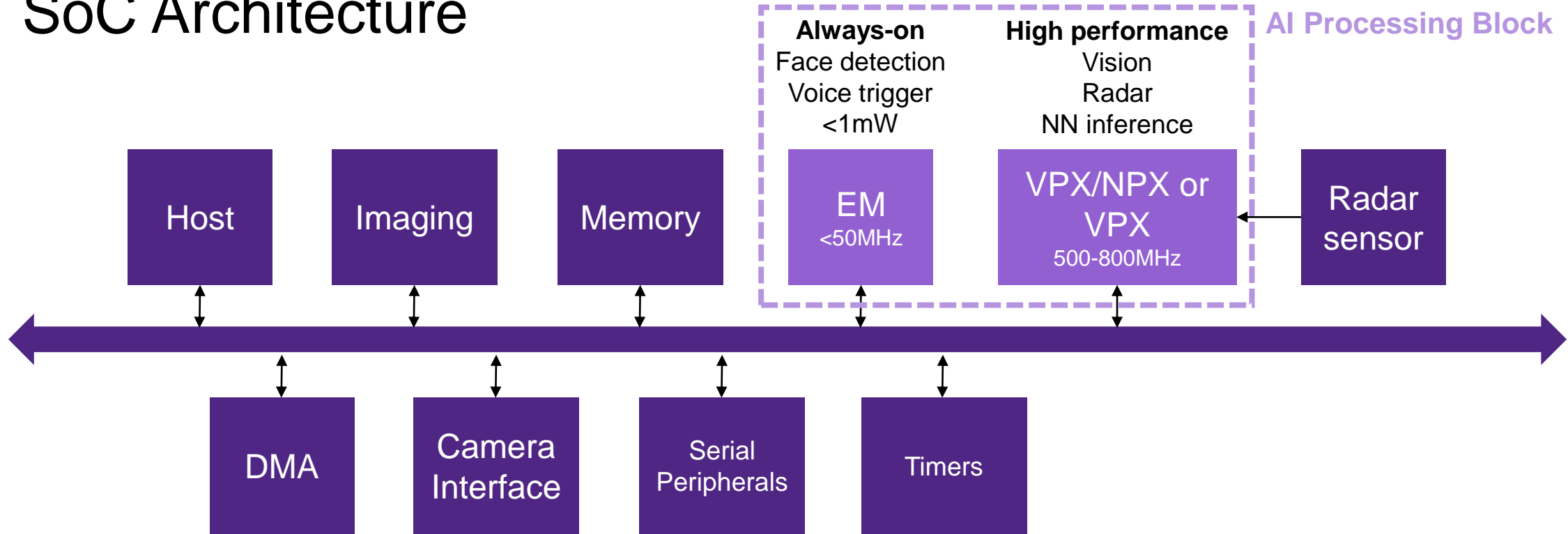
Early Collaboration and Architecture Analysis

Use-cases & workload

SoC requirements & realization



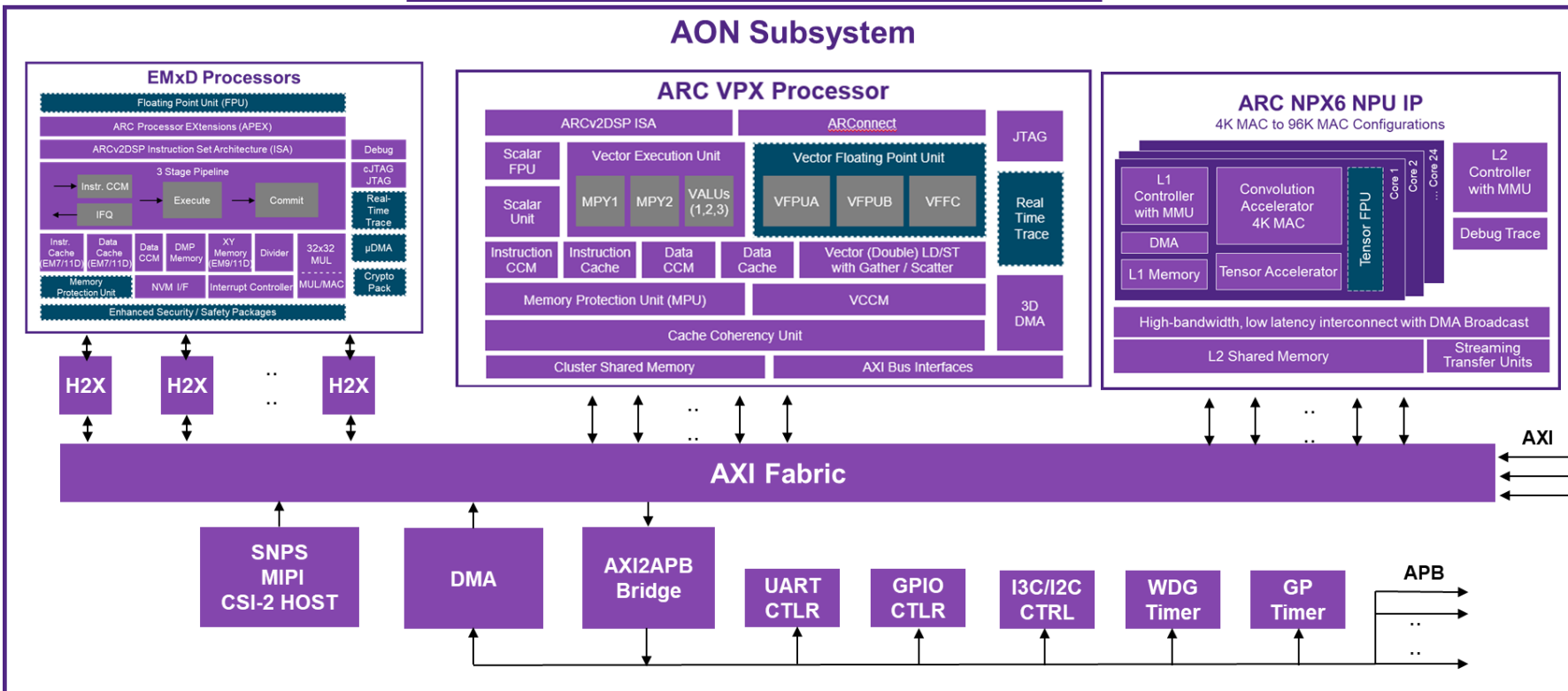
SoC Architecture



- Meeting 1mW for Always-On demands very low-power processor running at low frequency – **EM Processor**
 - For example: <20 uW/MHz x <50MHz
- High performance demands few orders of magnitude more performance – **VPX or VPX + NPX**
 - Vector DSP that meets diverse requirements of Vision, Radar and NN inference
 - Powered down when system is in Always-On mode

AON Vision Architecture

Proposed Solution with NPU + DSPs



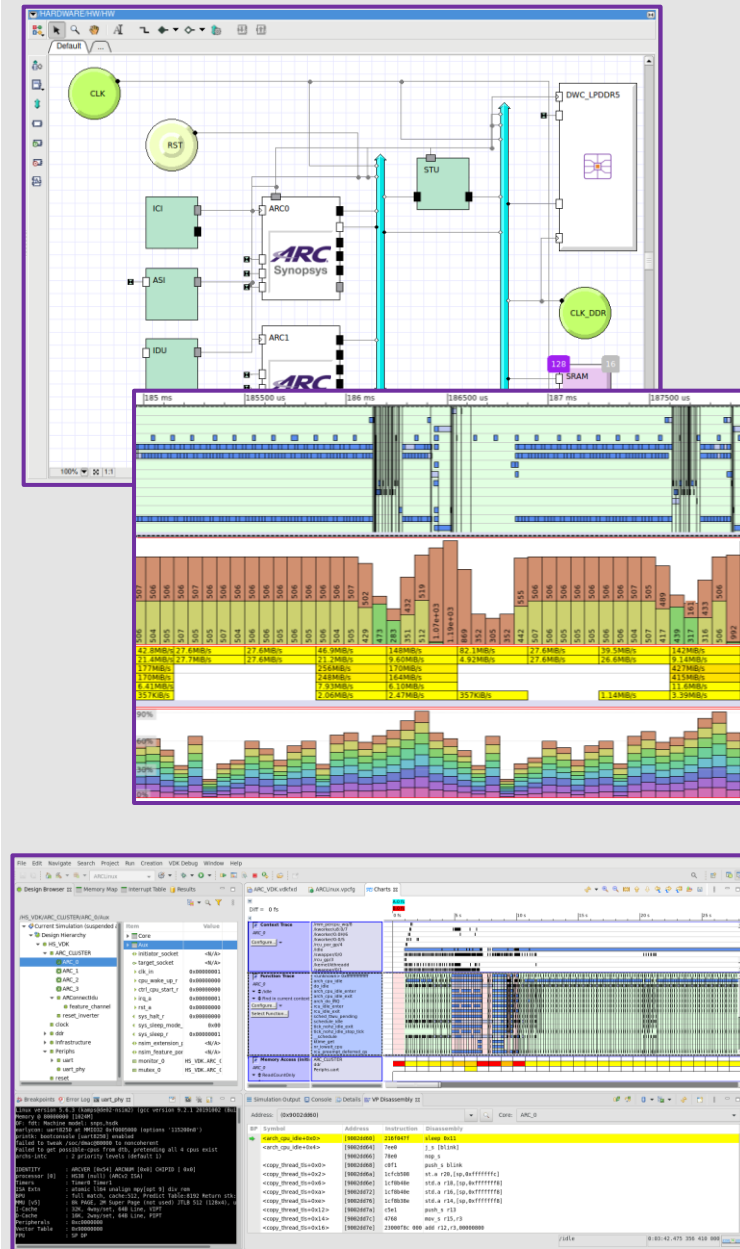
AON Subsystem Highlights

- Integrated pre-verified IoT subsystem ideal for always-on trigger functionality
- Ultra-low power EMxD DSP processor with high-efficiency control and signal processing operations
- VPX DSP w/ Vector processing units (VPUs) for high DSP-performance functionality
- NPX NPU with High performance and power efficiency of neural networks (NN)

VPX5 Simulation Models

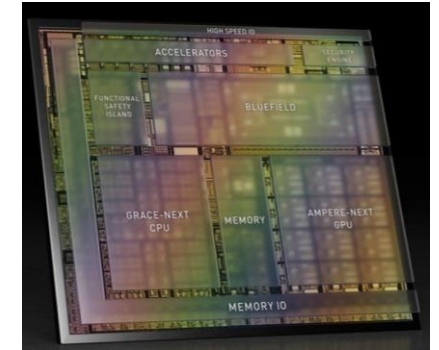
Support for building virtual prototypes

- nSIM/nCAM has
 - SystemC wrapper
 - Supports Accellera SystemC, Platform Architect, and Virtualizer
 - Model Libraries for Platform Architect and Virtualizer
 - For easy deployment in Synopsys Virtual Prototyping tools
 - Instrumented for debug and analysis
- Allows for easy creation of your own Virtual Platform
- Integration of MetaWare Debugger (mdb) into Platform Architect and Virtualizer
 - For debugging complete systems containing VPX5 and other processors

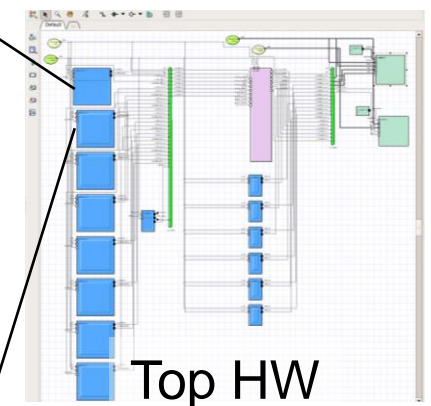
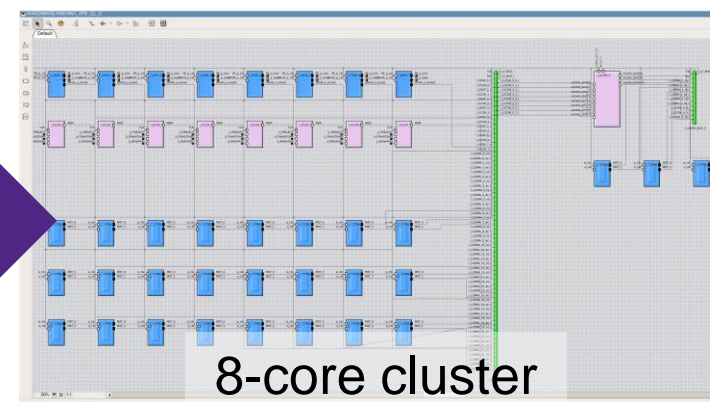
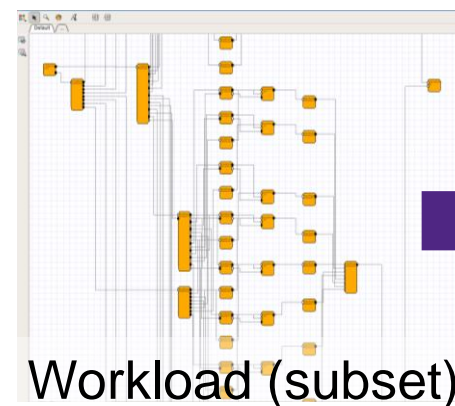
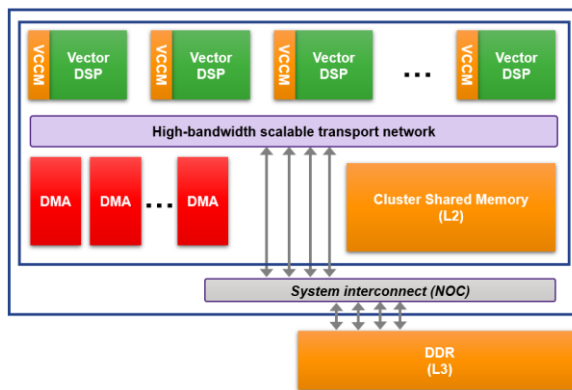


Autonomous Driving Compute Challenge

- Autonomous Driving today consumes 1.5 to 3 kW*
- 2022 Nvidia 4-Orin system consumes 250-300 W*



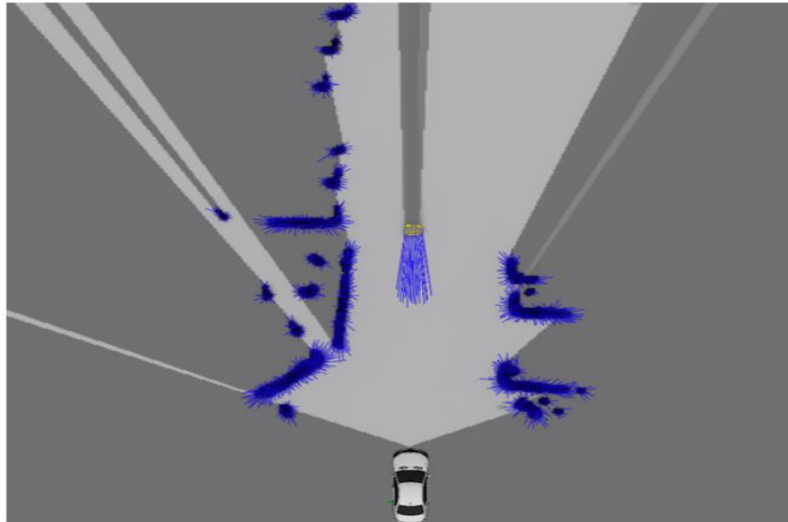
- Customer project: GPU prototype of new camera fusion algorithm by itself consumes 300W
- Synopsys proposes 64-core ARC vector DSP with SRAM/DMA, estimated to consume ~5W
- Created Platform Architect model to prove 30fps target performance



* <https://www.forbes.com/sites/samabuelsamid/2021/04/12/nvidia-launches-1000-tops-automated-driving-chip-volvo-to-launch-orin-powered-system-in-2022>

Particle Filter: the workload and the proposed cluster architecture

Particle Filter Workload



- Operates on large data-structures stored off-chip
 - 4M cells and 8M particles
- ~2T DSP instructions/second
- <12ms latency requirement

Main challenge:
bandwidth requirements

GPU Architecture Nvidia GTX 980 Ti

Optimize algorithm to fit
GPU parallelization

Caching

250W (40fps)

~8 GB per frame

25 ms Latency

400 mm² in 28nm

Proposed Cluster Architecture

Optimize algorithm to limit
external bandwidth

Explicit memory management

~6.3W (at 40 fps)(*)

~2.5 GB per frame

10.8 ms Latency

~50 mm² in 7nm(*)

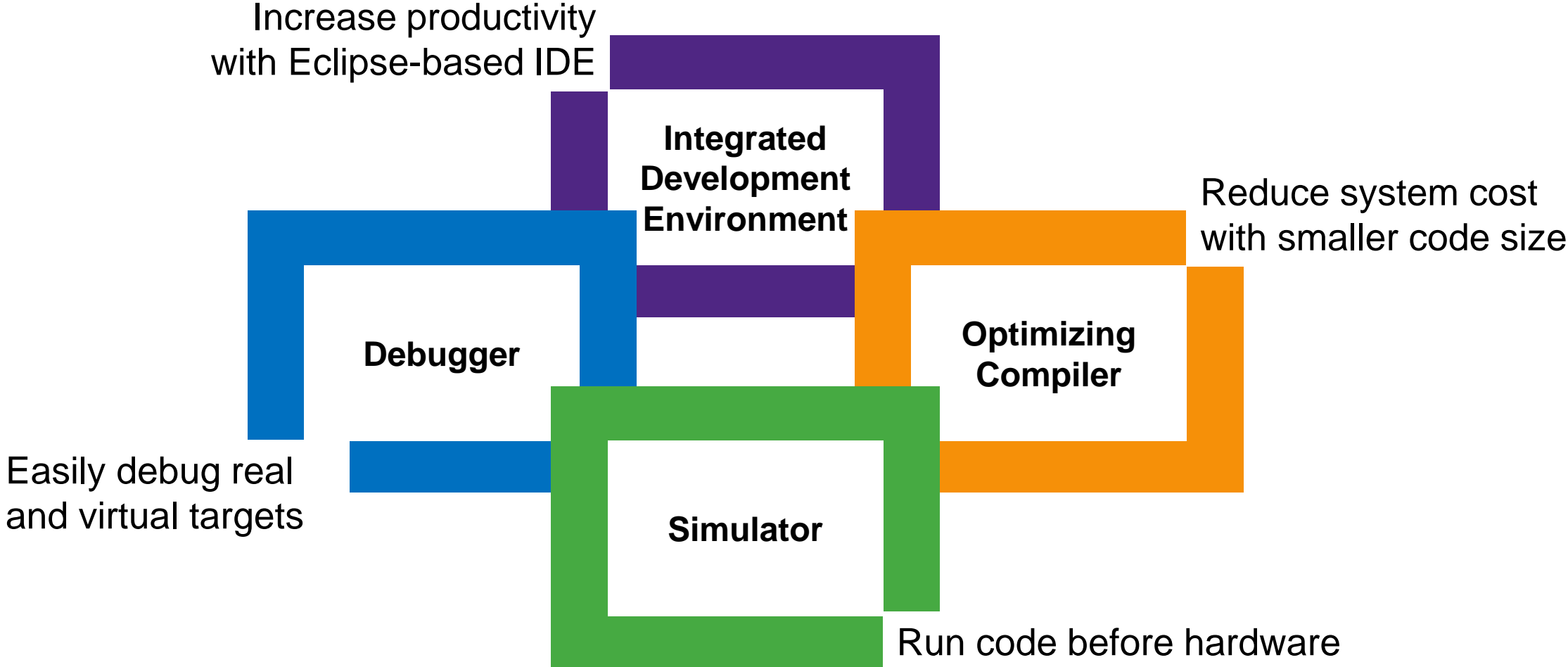
(*) Area and power incl. logic and on-chip SRAM,
excluding DDR PHY's

Agenda

- Always-on AIoT Applications Introduction
- Typical Requirements
- SoC Architecture & Use Case
- **Programming Model, Software and Tools**
- Synopsys - One-stop Shop
- Key Take-Aways

ARC MetaWare Development Toolkit – Unified Story

Integrated, Best-in-Class Tool Chain for Compilation, Debugging & Simulation



Machine Learning Inference S/W Library

Optimized for ARC EMxD and VPX

Group	Functions	Short Description
Main operations	<ul style="list-style-type: none"> • 2D convolution • Depth wise 2D convolution • LSTM • Simple RNN • Fully connected 	Convolve input features with a set of trained weights
Pooling	<ul style="list-style-type: none"> • Average pooling • Max pooling 	Pool input features with a function
Transform / activation functions	<ul style="list-style-type: none"> • ReLU • SoftMax • Leaky ReLU • Sigmoid • TanH 	Transform each element of input set according to a particular function
Data routing operations	<ul style="list-style-type: none"> • Padding • Transpose • Concatenation 	Move input data by a specified pattern
Elementwise operations	<ul style="list-style-type: none"> • Vector arithmetic operations 	Apply multi operand function elementwise to several inputs

Software library targeted at ML inference on ARC DSP cores

- Library of kernel functions for effective inference of machine learning models

Supports easy implementation of layered NN graph topologies

- Library of optimized kernels for implementing multiple NN layer types
- High efficiency & small footprint kernel implementations
- C style APIs

Use models

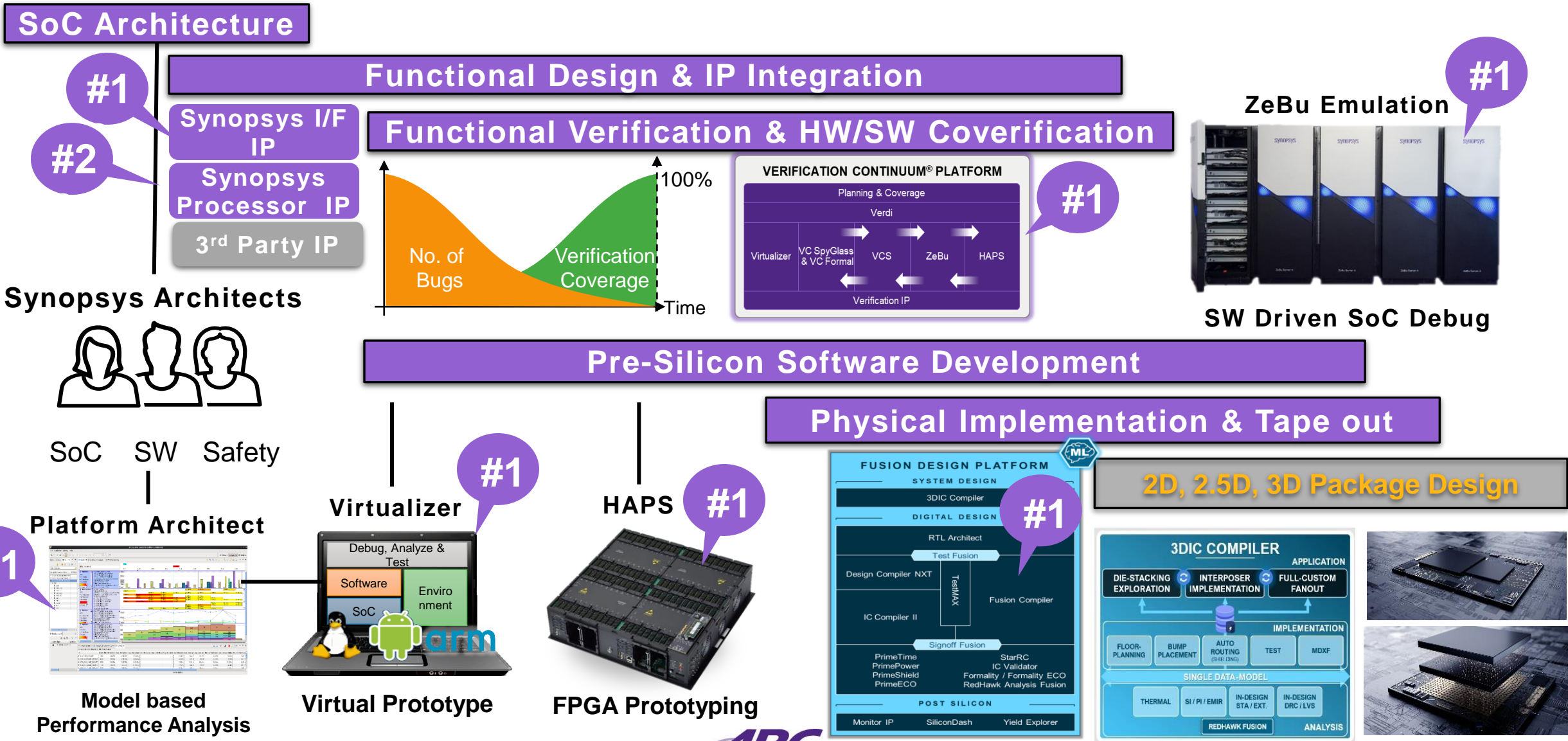
1. Using External ML framework (TensorFlow Lite Micro with MLI-accelerated backend)
2. User-callable APIs (manual graph mapping)
3. Automated graph mapping using MetaWare NN SDK (Available soon)

Available on embARC.org

Agenda

- Always-on AIoT Applications Introduction
- Typical Requirements
- SoC Architecture & Use Case
- Programming Model, Software and Tools
- **Synopsys - One-stop Shop**
- Key Take-Aways

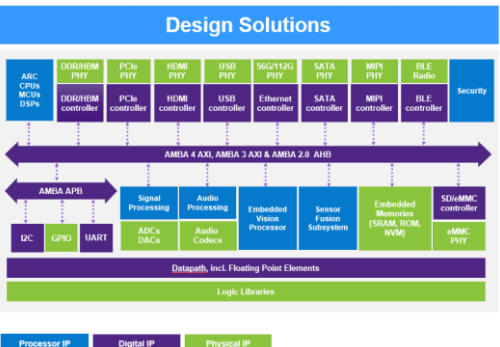
Synopsys Provides a Complete SoC Design Solution with Tools, IP and Services from Architecture to Silicon and 3D Packaging



Synopsys' Differentiation

Synopsys is the lowest risk solution with our IP, Tools, Services, Experience, and Partnership track record

IP



Design Solutions

ARC CPUs MCUs DSPs, DDR/DRAM PHY, PCIe PHY, HBM PHY, USB PHY, SDC/12G PHY, SATA PHY, MIPI PHY, BLE Radio, Security

DDR/DRAM controller, PCIe controller, HBM controller, USB controller, Ethernet controller, SATA controller, MIPI controller, BLE controller

AMBA 4 AXI, AMBA 3 AXI & AMBA 2.0 AHB

AMBA APB

DC, GPO, UART, ADCs DACs, Signal Processing, Audio Processing, Embedded Vision Processor, Sensor Fusion Subsystem, Embedded Memories (SRAM, DRAM, NVM), eMMC controller, SD/eMMC controller

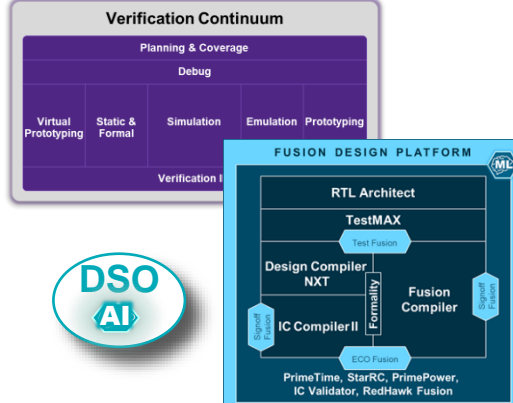
Datapath, incl. Floating Point Elements

Logic Libraries

Processor IP, Digital IP, Physical IP

- Broadest portfolio of silicon-proven IP
- Complete IP solution
 - Including **Processors, Security, Digital IPs, PHYs, Serdes, PPA Libraries, and Memories**
- #1 Provider of interface, embedded mem, & physical IP
 - Including **PCIe-related and DDR-related cores**

Technologies and Flows



Verification Continuum

Planning & Coverage, Debug, Virtual Prototyping, Static & Formal, Simulation, Emulation, Prototyping, Verification I


FUSION DESIGN PLATFORM

RTL Architect, TestMAX, Design Compiler NXT, Fusion Compiler, IC Compiler II, ECO Fusion, PrimeTime, StarRC, PrimePower, IC Validator, RedHawk Fusion

DSO AI

- #1 in EDA
- Highest performance verification and emulation solution in the industry
- Fusion Platform and Experienced Services deliver fastest convergence and best QOR
- #1 in Prototyping (ZeBu/HAPS)

Relationships and Partnerships



Outstanding Outsourcing Partner 2019

AWARD OF OUTSTANDING OUTSOURCING PARTNER

Presented to Synopsys in recognition of its strategic implementation services provided to TSMC


May, 2019

synopsys

- SoC and Processor subsystem repeat customers for over 20 years
- Deep foundry collaboration with TSMC, Samsung and Global Foundries
- DCA with TSMC
- Advanced Node Experience with 75+ at 7nm and below

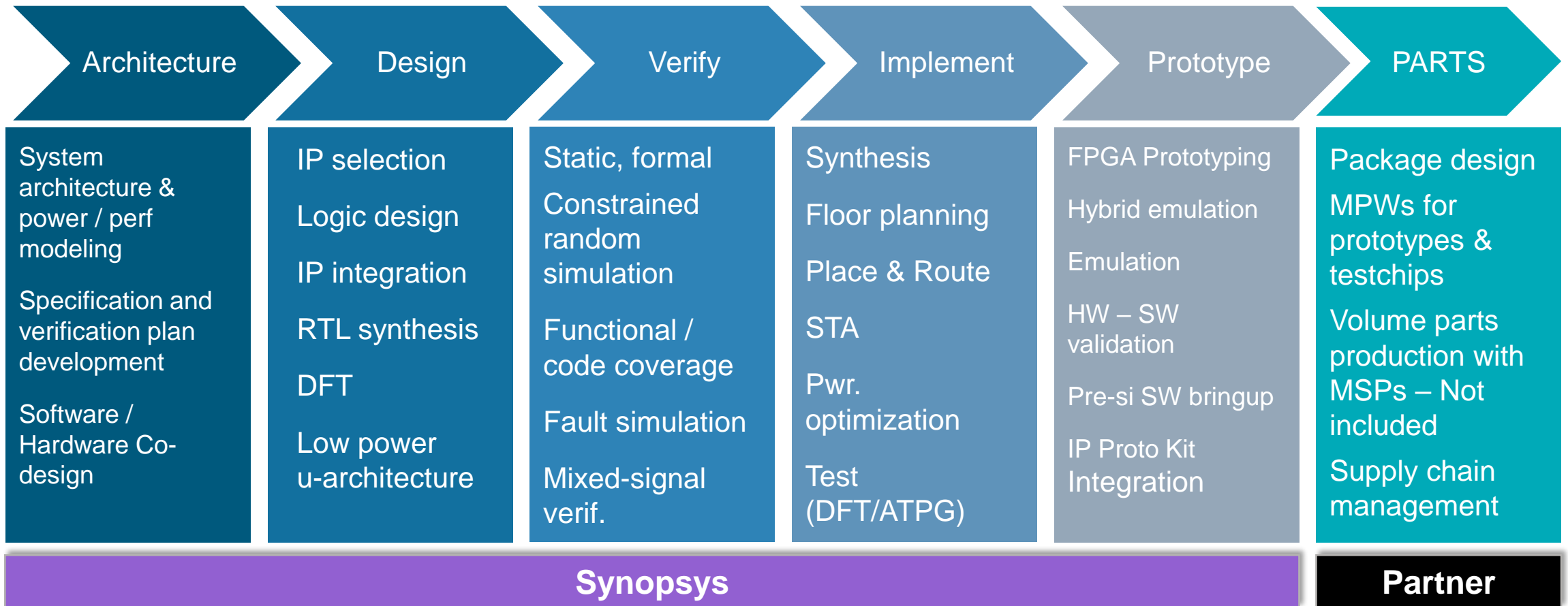
Service and Experience

- Over 30 years of high quality SoC design services
- SG IP teams are ISO9001 Certified
- Strong track record of delivering first pass silicon successes
- Successful IP, EDA and Services delivery to OEMs and Tier1's
- Successful SoC development for Automotive, AI, and HPC
- Ensuring SoC design knowledge transfer for COT needs
- Proven verification and physical design flows
- 99.2% customers "Satisfied" or "Very Sat." (1300+ surveys)



SoC Consulting Services

Expertise and experience in all aspects of SoC development



Reducing SoC Development Risk

Ultra Low Power Vision Block

Architecture to GDSII for image recognition subsystem

Design Challenge

- Tight schedule for Low Power Image Recognition Subsystem
- Need IP protocol, integration, verification expertise/experience
- Need an immediate and adaptable verification environment
- Need complete Spec2gdsii partner w/design methodology and flow

Services Delivered

- Early Architecture interaction on IP selection
- IP Configuration and Integration
- Power Aware Verification
- Design (Frame router and SRAM controller)
- Verification
- Physical Implementation

Project Results

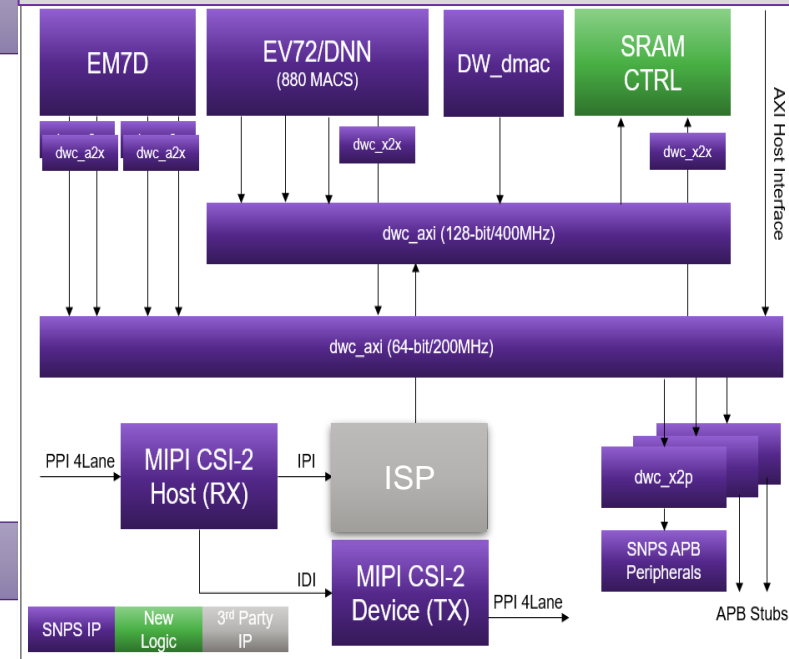
- Delivered on schedule

“I want to send the entire Synopsys team this personal note of my sincere gratitude and appreciation to you as you’ve done an impressive amount of work in a short time.

Your skillset and expertise are incredible.

Thank you for being a great partner”

- Technical/Project Lead



Agenda

- Always-on AIoT Applications Introduction
- Typical Requirements
- SoC Architecture & Use Case
- Programming Model, Software and Tools
- Synopsys - One-stop Shop
- **Key Take-Aways**

Key Take-Aways

- ARC Ease of Use & Broad Portfolio

- High Performance + Low Power ML is achievable!
- AON Architecture like ARC EM + VPX/NPX – Very efficient
- Efficiency via instruction-level parallelism operation
- ARC products provide Low-Active Power, Fast Response Times and High Performance

- Unified MetaWare Toolchain & Software

- IDE, Compiler, Debugger and Simulator
- MLI Library and NN SDK

- One-Stop Shop

- Complete SoC Design Solution
- Processors, Security & Interface Ips
- Industry leading EDA Tools and flows, IP, and Services from Architecture to Silicon and 3D Packaging
- SoC Consulting Experience

DesignWare ARC Processor IP

Unrivaled Efficiency for Embedded Applications

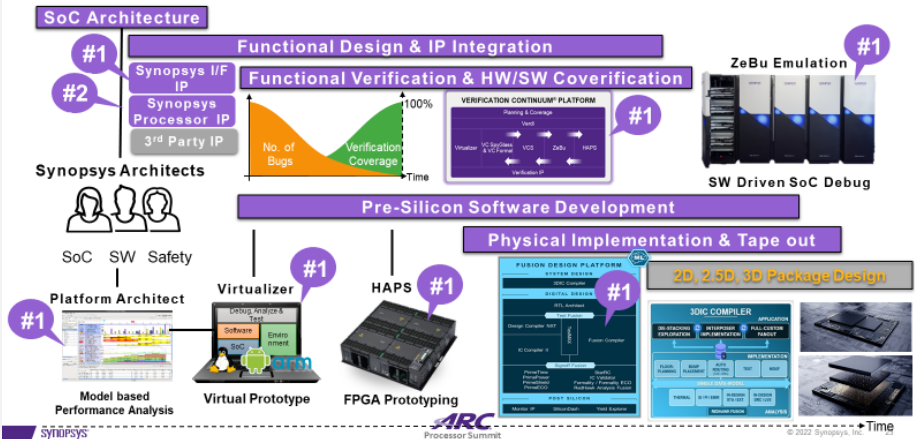
EM Family	SEM Family	HS Family	EV Family	VPX Family	NPX Family
<ul style="list-style-type: none"> Optimized for ultra low power IoT 3-stage pipeline w/ high efficiency DSP Power as low as 3uW/MHz Area as small as 0.01mm² in 28HPM 	<ul style="list-style-type: none"> Security processors for IoT and mobile, including DSP Protection against HW, SW, and side channel attacks SecureShield for Trusted Execution Environments 	<ul style="list-style-type: none"> Highest performing CPUs, CPU + DSP 32- & 64-bit ISAs High-speed 10-stage pipeline SMP Linux support Single- and multi-core configurations 	<ul style="list-style-type: none"> Heterogeneous multicore for vision and AI processing DNN (Deep Neural Network) Engine High productivity, standards-based tool suite 	<ul style="list-style-type: none"> High performance vector DSP SIMD/VLIW design for massive parallel processing Multiple vector FP engines for high precision results 	<ul style="list-style-type: none"> Scalable neural processor units Up to 250 TOPS (440 TOPS with sparsity) Supports latest AI applications High productivity, standards-based tool suite

Functional Safety (FS) Processors

- Integrated hardware safety features for ARC EM, SEM, HS, VPX, EV and NPX processor families
- Accelerates ISO 26262 certification for safety-critical automotive SoCs

© 2022 Synopsys, Inc.

Synopsys Provides a Complete SoC Design Solution with Tools, IP and Services from Architecture to Silicon and 3D Packaging



Thank You



SYNOPSYS[®]

Silicon to Software[™]