SYNOPSYS®
Silicon to Software™

# Virtualized AI Workloads – How and Why?

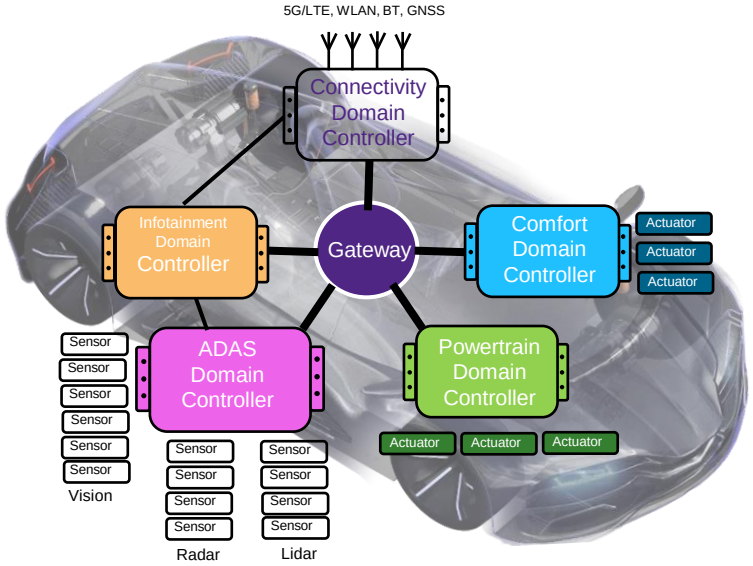
Fergus Casey
R&D Director, ARC Processors, Synopsys

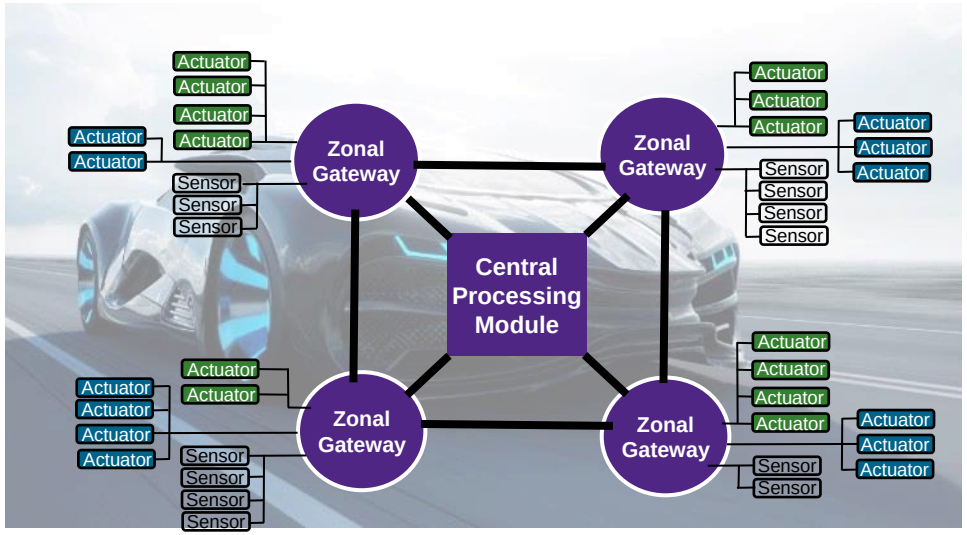
Synopsys ARC® Processor Summit 2022

# Agenda

- Evolution of Automotive Architectures
- Automotive SoC Consolidation
- Mixed Criticality -> Virtualization
- Automotive AI Use-cases
- Virtualization for AI Workloads
- Resulting Automotive SoC Architecture
- Summary

# Market Dynamics: Zonal Architecture Reshaping Automotive SoCs


Yesterday
30- 100+ ECUs in a car

Mainstream MCUs

Today
**Domain Logical Architecture**

Consolidating of ECUs
Integration of Functions, AI & ICs require 16/14nm & 7nm FinFET Class SoC

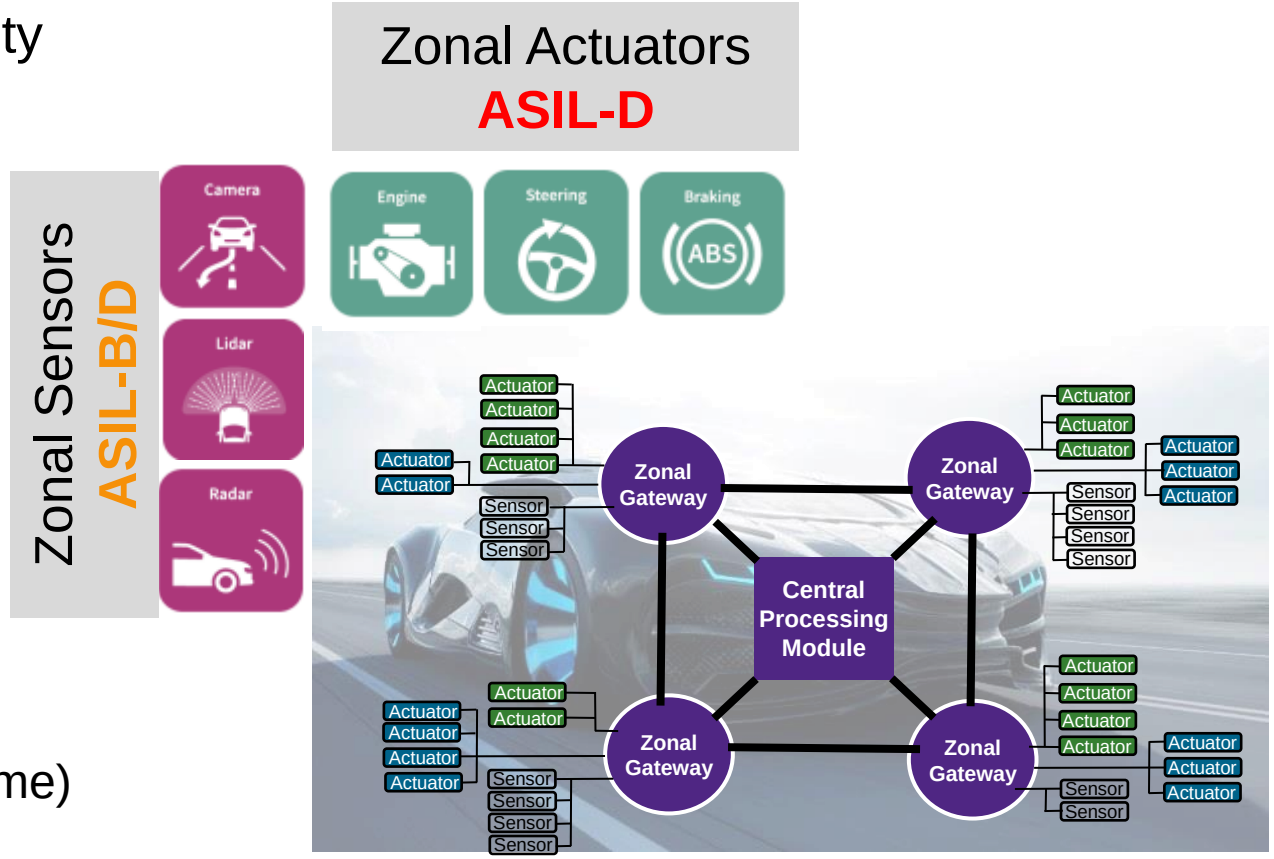
Tomorrow/Future
**Zonal Physical Architecture**

Multi-Applications Central Processing
**Multi-Chip** & Higher Complexity/Performance 5nm SoC

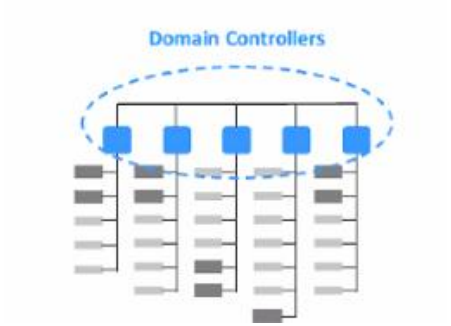
# What is Mixed Criticality

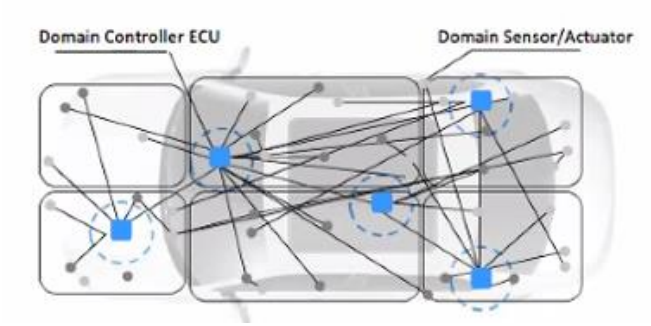

- Several **apps of different ASIL safety** levels running on the same SoC/Platform
- Each app requires **isolation** to maintain safety & security requirements

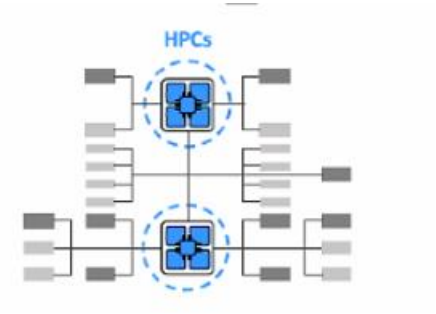
- **Spatial Isolation**
  - Resource partitioning
    - Achieved through HW Separation
      - Discrete ICs (Domain Controller architecture)
    - Achieved through SW Separation
      - Integrated multi-functional SoC
      - Integrated Multi-ASIL level support
- **Temporal Isolation**
  - Guaranteed WCET (Worse Case Execution Time)

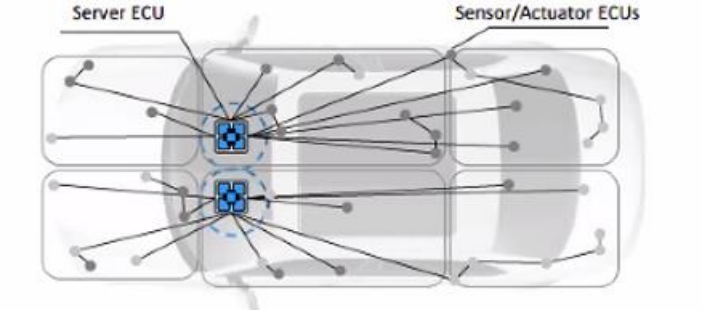


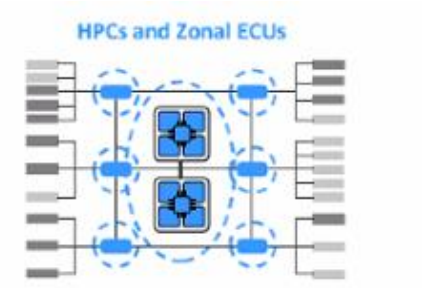
# Evolution of Automotive Architecture

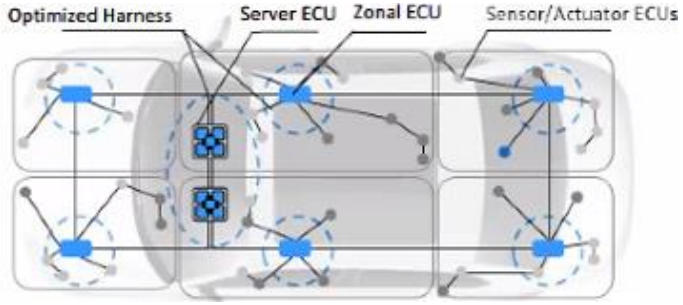



**Domain Based - Dedicated functional specific MCUs**

- **Scalable real time** cores (RTC)
- Scalable **DSP,** (GFLOPS)
- **Scalable NN accelerators** (TFLOPS)

– Fixed ASIL Level per SoC per Use-case

– Isolation achieved through physical HW separation

**Centralized - High Performance Multi-functional MPUs**

- High Performance Application Processors
- High NN enabled (TFLOPS)

– Supports multiple use-case processing

- **Mixed Criticality Architecture**

**Zonal Architecture - Zonal MCUs with Centralized backbone**

– Multi-functional SoC Integrated Processors

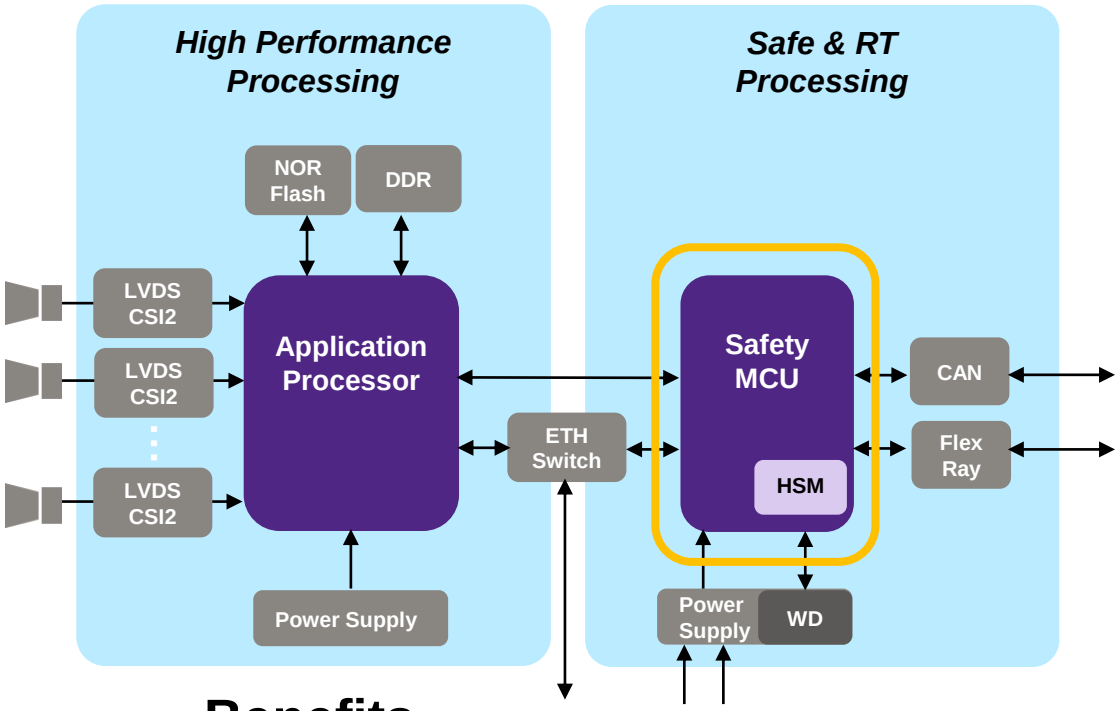
- Scalable real-time and application cores
- Scalable NN accelerators (TFLOPS)

– Supports multiple use-case processing

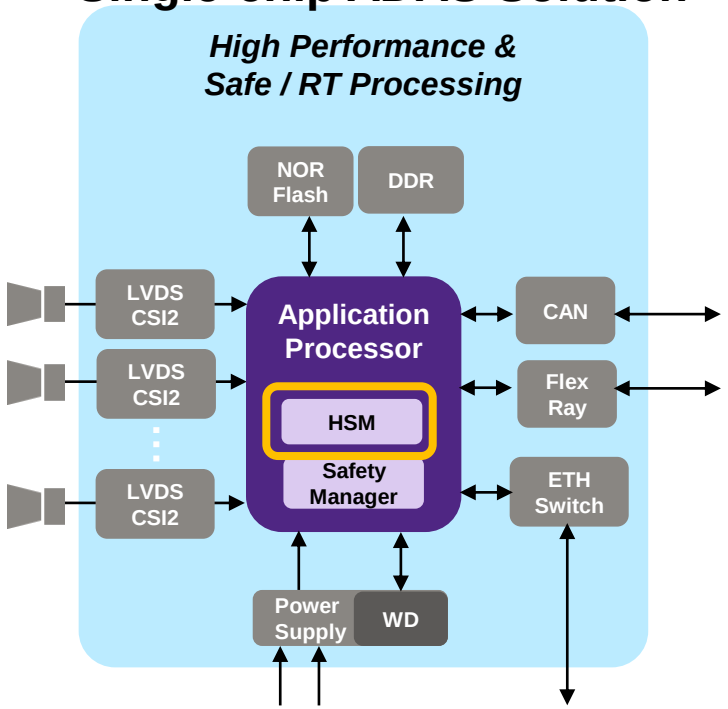
- **Mixed Criticality Architecture**

# Zonal Architecture => SoC Consolidation

## Multi-chip ADAS Solution




## Single-chip ADAS Solution




**Benefits**

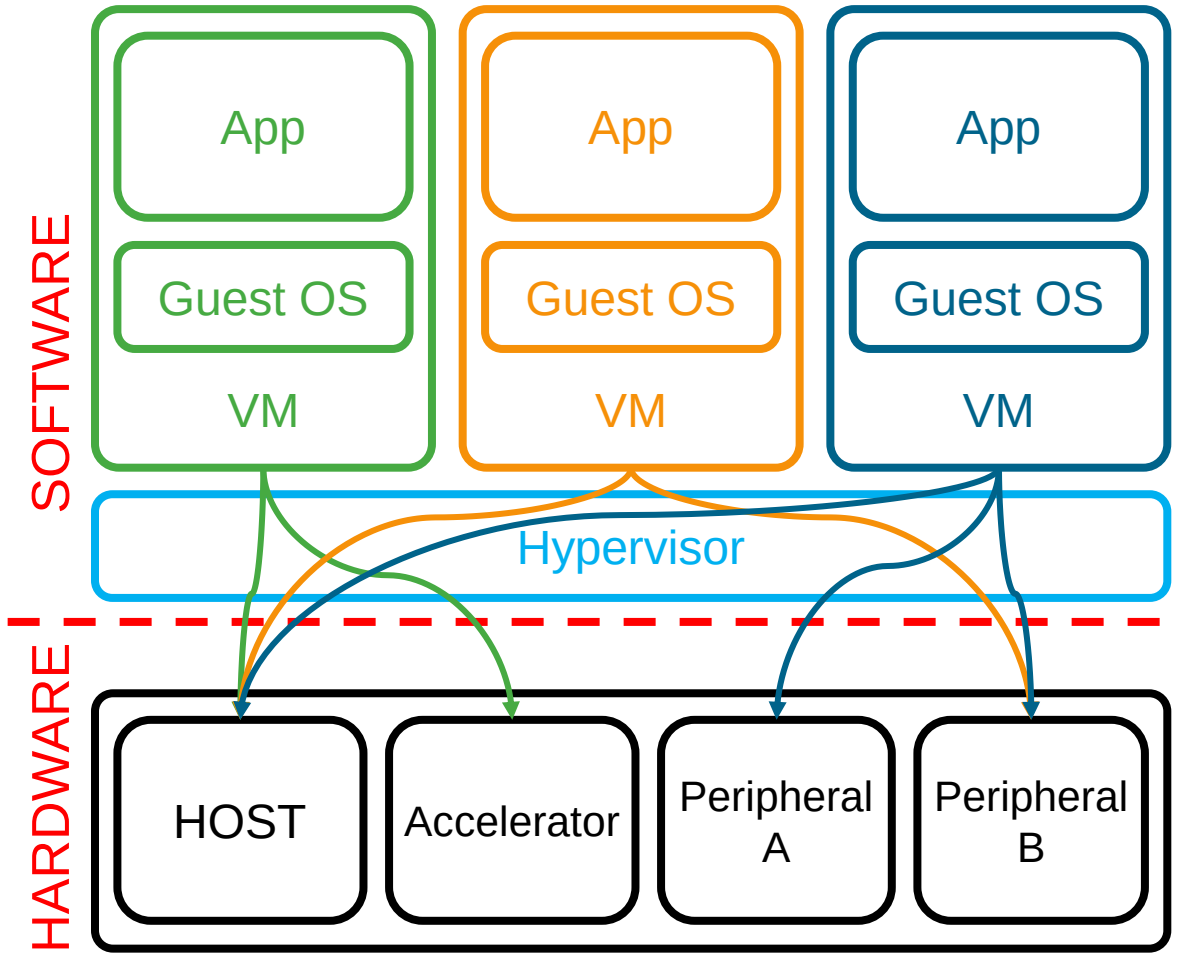
- Ease of Integration
- Reduced cost and weight

**Challenges**

- Added SoC Complexity – Freedom-from-Interference
- Multifunctional SoC demands Mixed Criticality
- Temporal and Spatial Isolation per use-case

**SoC Consolidation eases integration while increasing complexity to demonstrate Freedom-From-Interference**

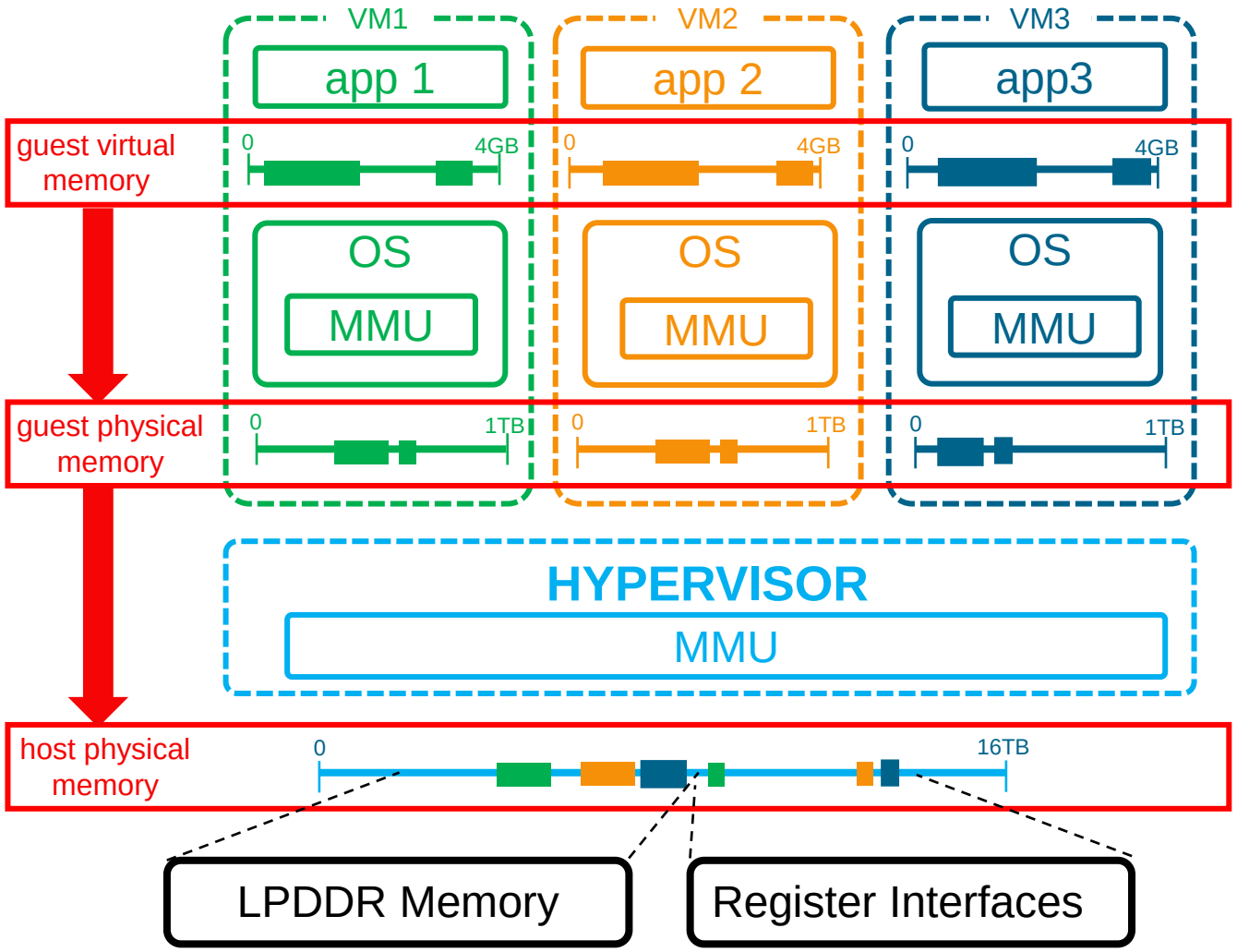
# Virtualization




- Multiple operating systems running in a virtual machine (VM) sharing the same hardware resources
- Hypervisor:
  - Manages hardware resources to virtual machines
  - Isolation between virtual machines
- Two layer MMU
  - OS: guest virtual → guest physical
  - Hypervisor: guest physical → host physical
- Accelerators and Peripherals
  - can be shared VM's
  - can be dedicated to a VM

**Virtualization as a safety mechanism for spatial and temporal isolation**

# Double Translation in the MMU




Two level translation by MMU

- Guest Virtual → Guest Physical owned by the OS in the VM
- Guest Physical → Host Physical owned by the Hypervisor

Spatial Isolation:

- Between Applications
- Between OS

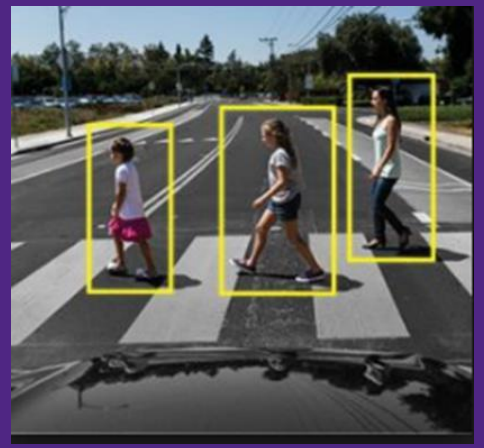
Physical Memory:

- External Memory
- Memory mapped IO of peripherals

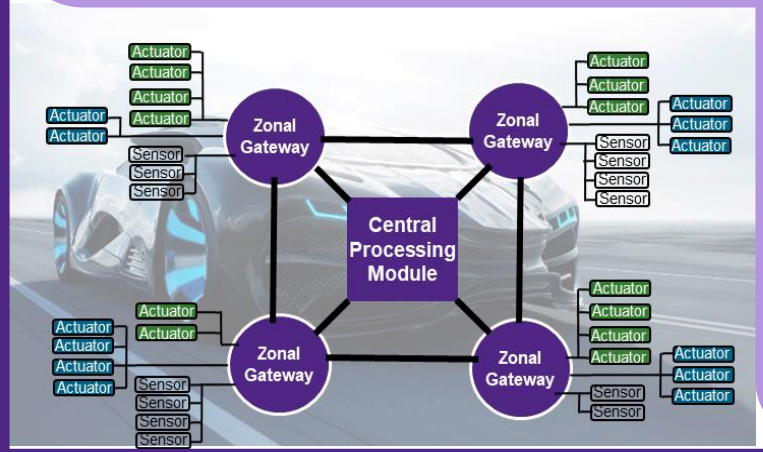
# Snapshot of Automotive AI Use-cases & Examples

## AI Enabled Sensors

- Multi-camera, Multi-Object Detection and Tracking
- Scene segmentation
- Lane Tracking
- Vision, Radar, LiDAR

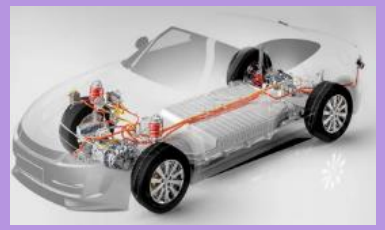


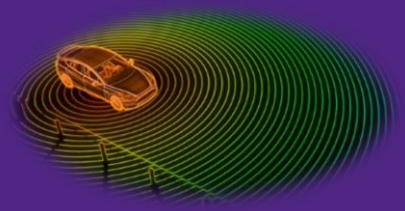

## Zonal Actuators


## AI Enabled Actuators

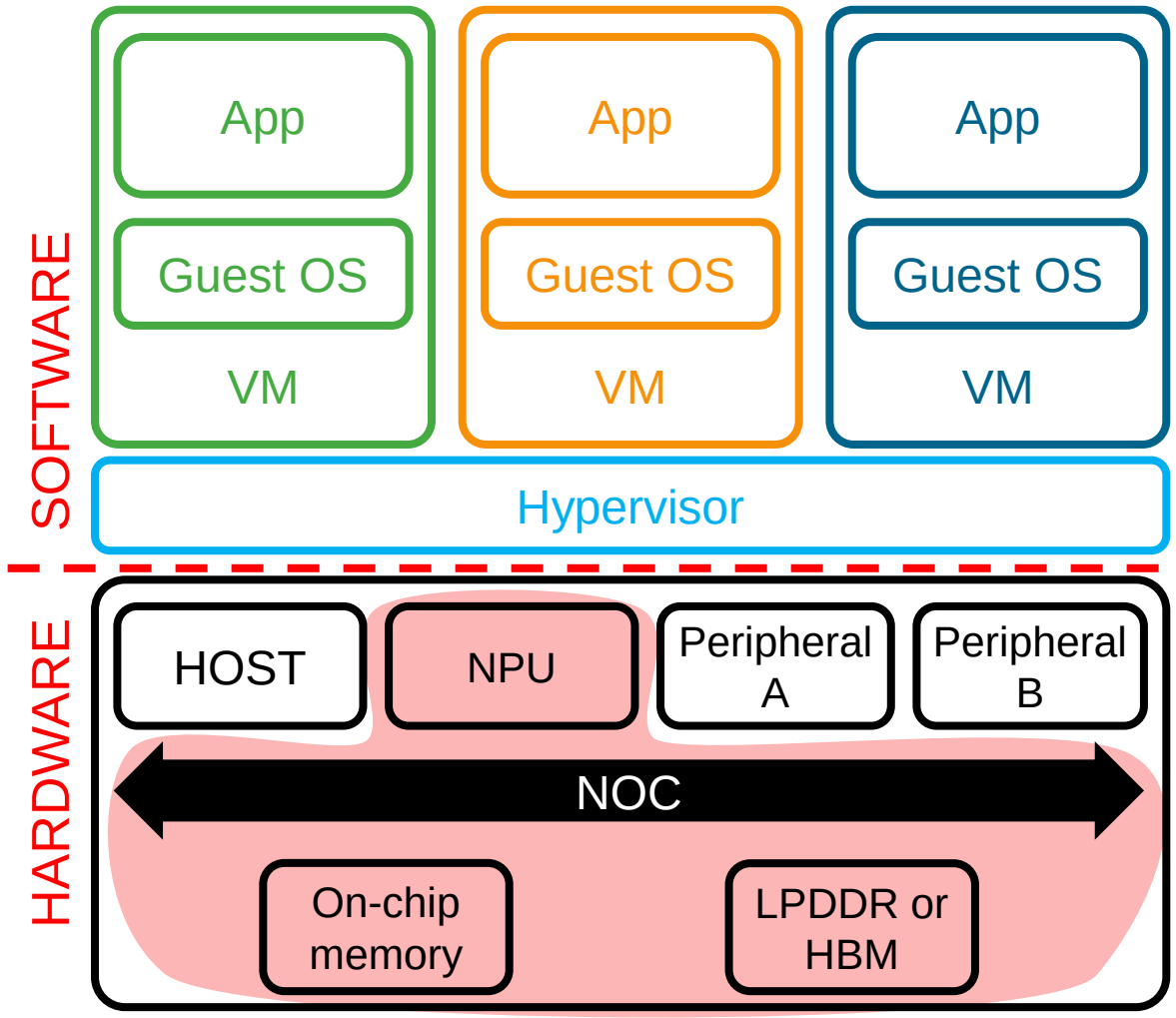
- Kalman Filter + NN replacing physical sensors with virtual sensors
- Battery Management – Predicting State of Charge and State of Health
- Predict Vehicle Motion Control



**>1 TOPs/S** **>10 TOPs/S** **100 TOPs/S**

# Virtualization for AI Workloads

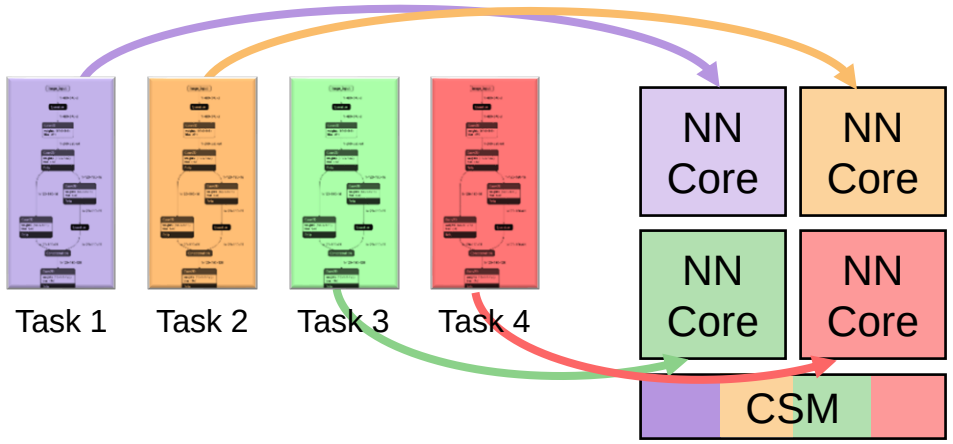


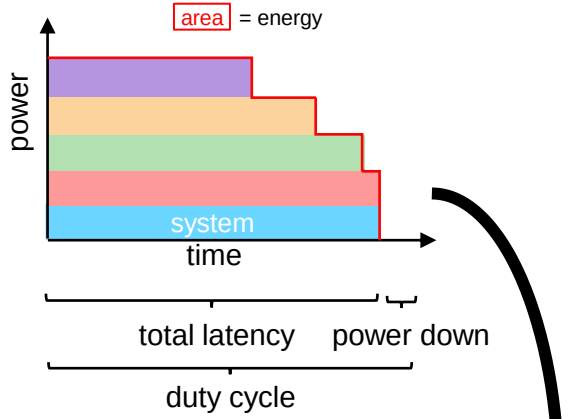

- Significant cost (area/power) saving by sharing:
  - AI Accelerator (NPX6)
  - NOC
  - On-chip memory (optional)
  - External memory interfaces (LPDDR/HBM)
- Isolation between applications and operating systems sharing the NPU
- Different ways to partition the resources
  - Time-slicing the compute and memory resources
  - Partitioning of the cores in the NPX and the memories
  - Mix of time-slicing and partitioning

**Virtualization provides significant cost reduction through shared AI (NPU) resources**

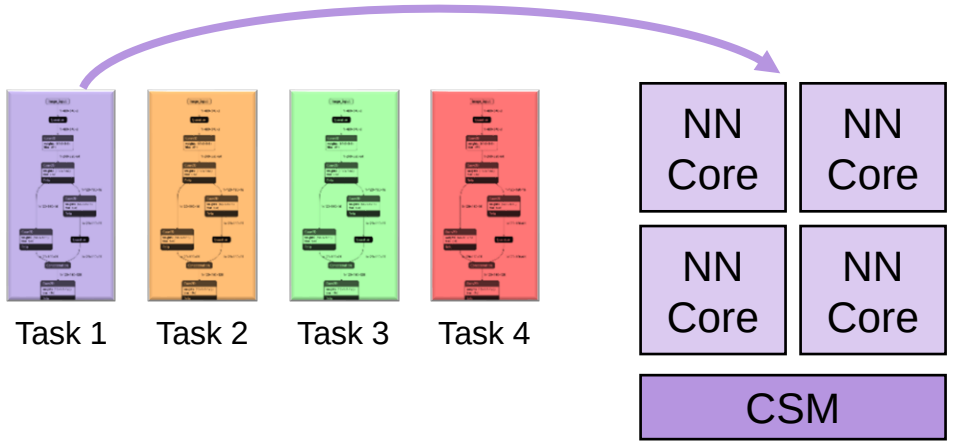
# Running Multiple Models on Multiple NN Cores

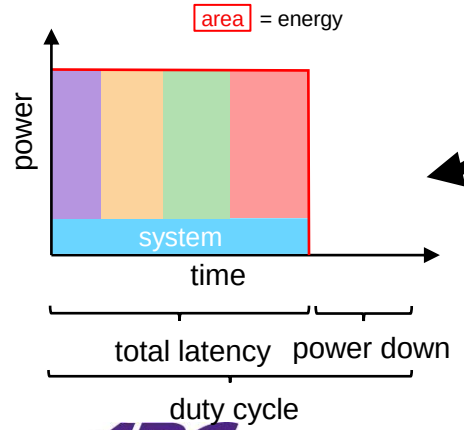
## Models mapped to separate cores, running in parallel



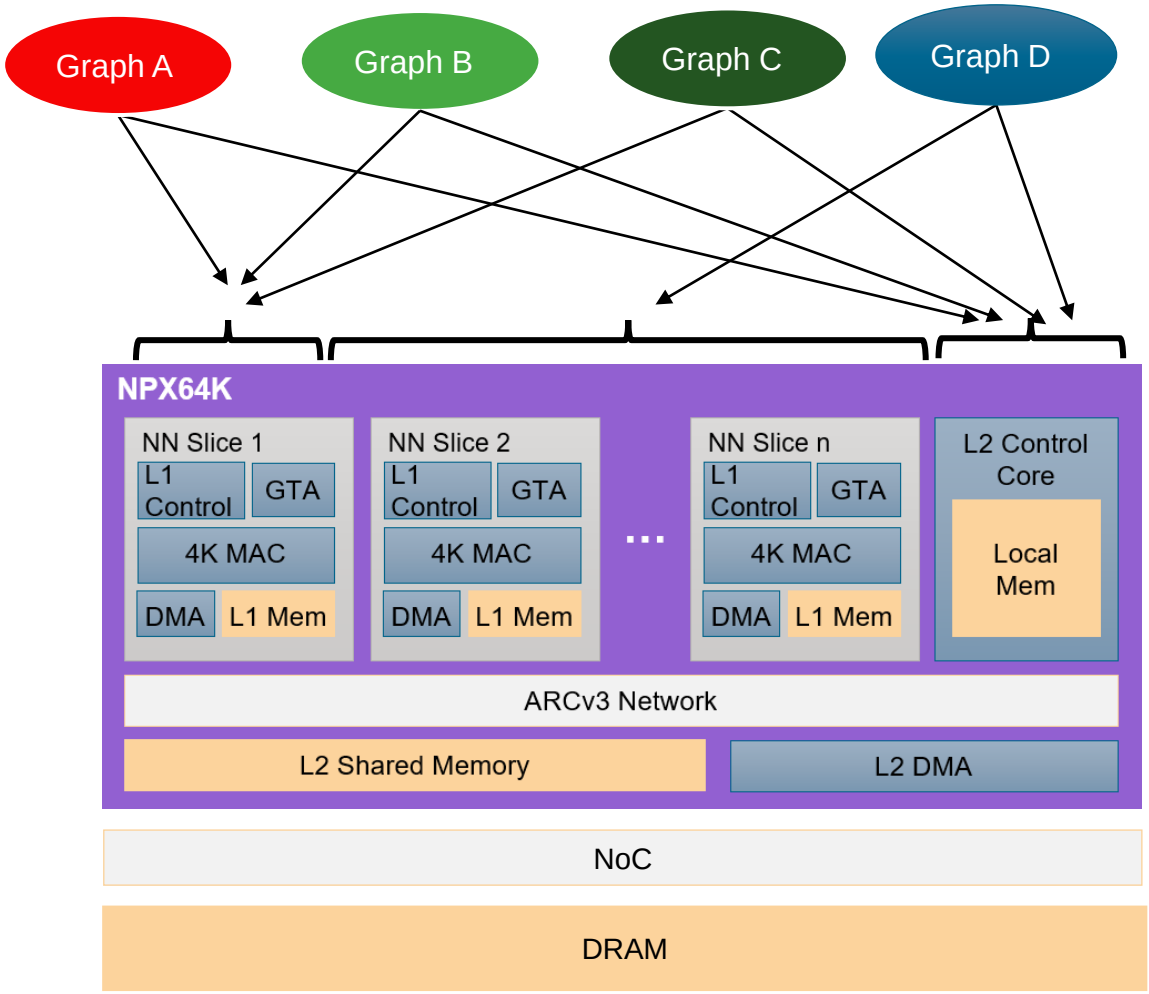

## Models mapped to all cores, running sequentially




- Pros of parallel model execution
  - Can handle model with unequal latency requirements
- Pros of serial model execution
  - Lower total latency
    e.g. for unequal models
  - Smaller L2 CSM needed:
    equal to the single worst-case requirement of all models
  - Lower external bandwidth:
    sequential execution allows to fully use the L2 CSM memory for each model. This allows larger tiles and segments, which imply lower DRAM bandwidth
  - Lower energy.
    Lower bandwidth decreases DRAM power

# Spatial and Temporal Isolation



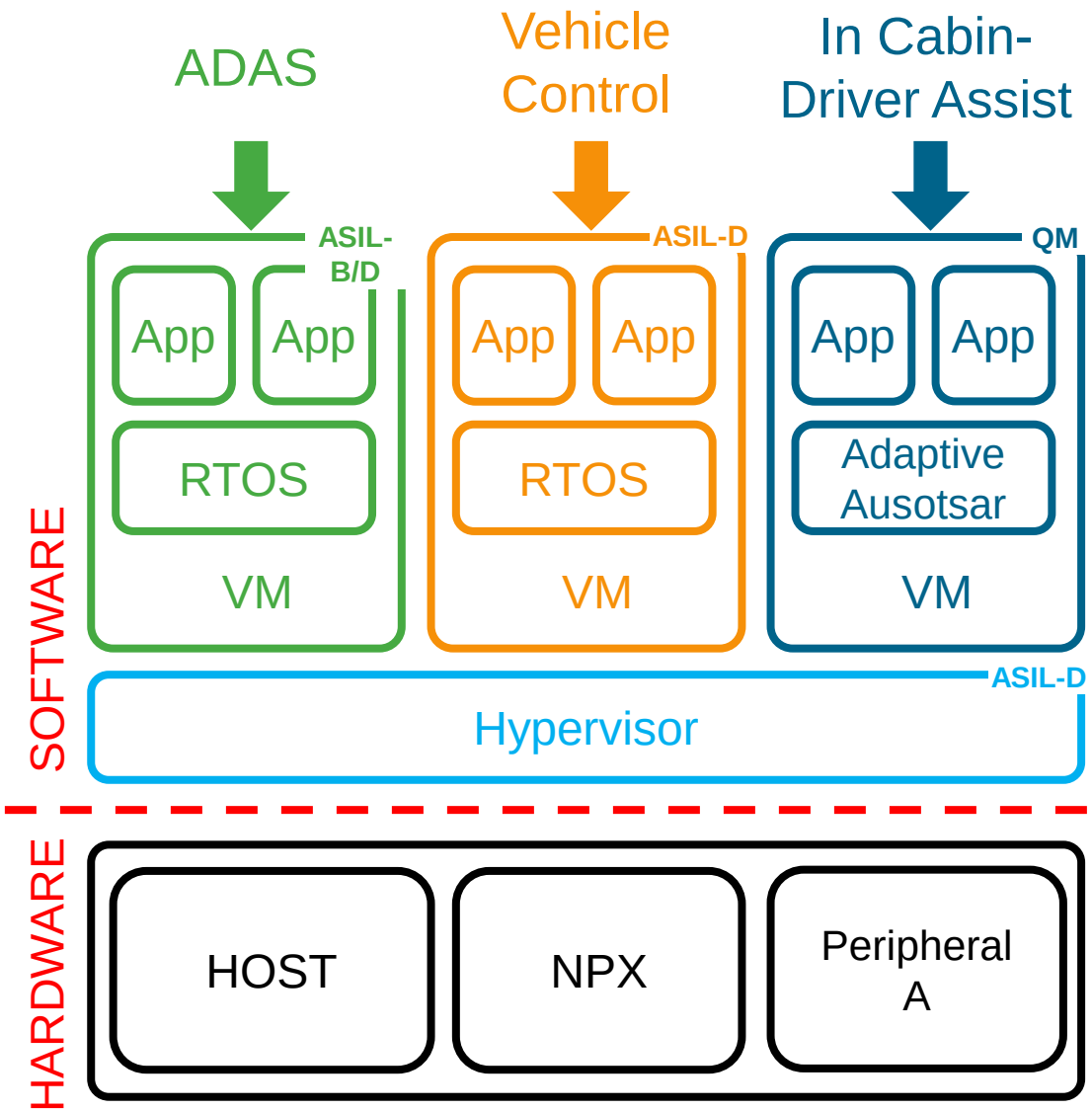

## Spatial Isolation

- Spatial isolation between graphs is required for functional safety, regardless of virtualization
- Shared resources:
  - NN Slices, L2 memory, L2 Control, DMA's, NoC and DRAM
- Spatial Isolation is implemented in ASIL-B firmware of the NPU

## Temporal Isolation

- Worst-case-execution-timing is monitored by NPU runtime.
- QM tasks may be interrupted to meet timing constraints of ASIL tasks

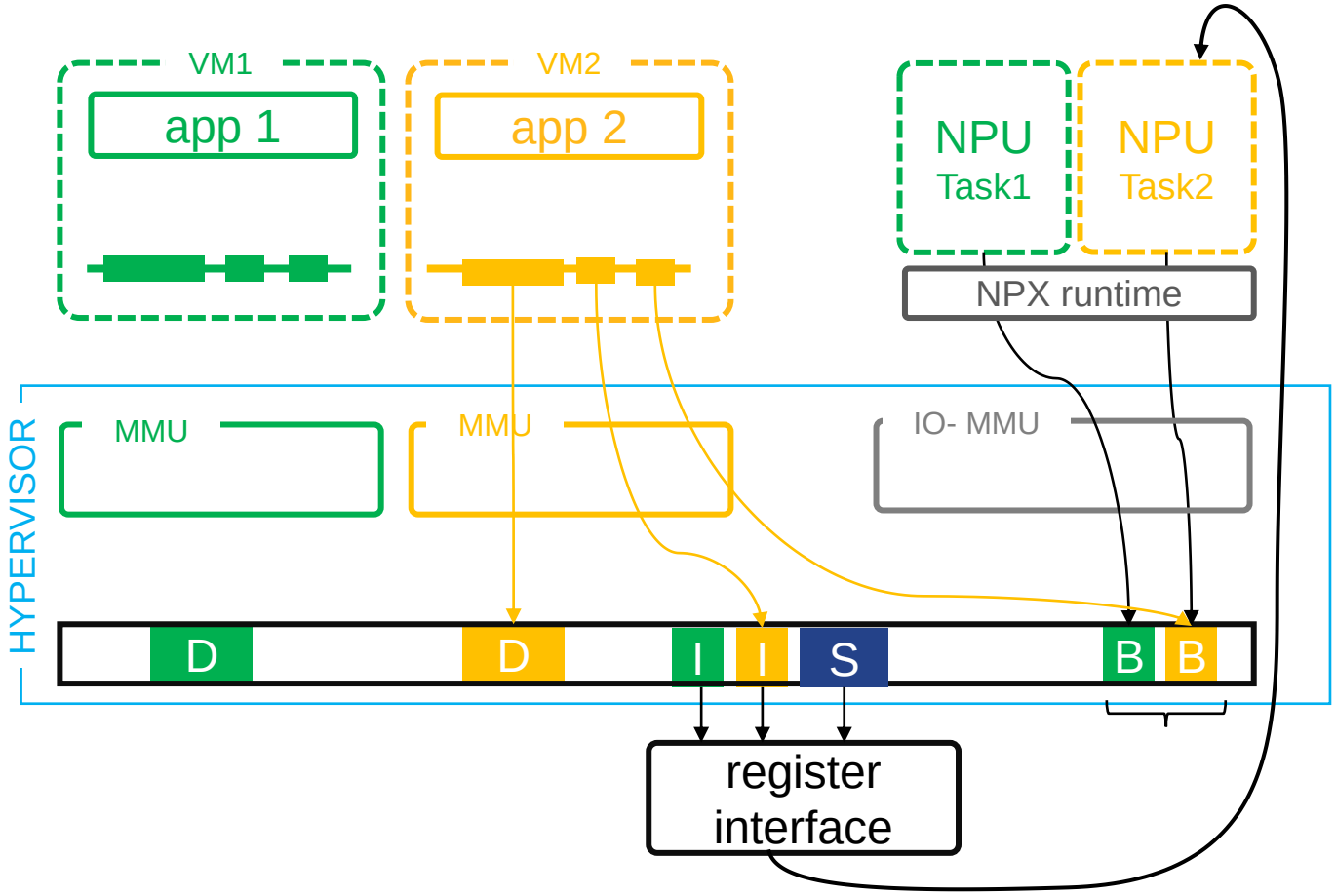
# Automotive Use Cases





- Software Partitioning in Virtual Machines
  - Isolated feature-rich virtual execution environment for each software partition
  - Suitable operating system for each partition
  - Independent development and certification
- Resource sharing with load balancing
- Software running in VM can be re-used on different hardware platforms
- Virtualization provides isolation for functional safety per ISO26262 standard:
  - Between ASIL-D software components
  - Between software components with mixed criticality levels

# Virtualization with Hardware Virtualization




Hardware virtualization requires:

- Independent hardware register interface for every VM
- IO-MMU integration on the initiator ports of the NPU

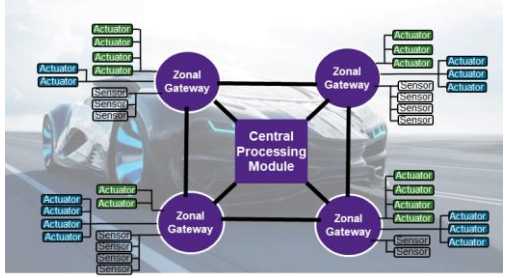
With these hardware features:

- Buffers used by NPU do not need to be contiguous in memory and does not need to be copied
- No Hypervisor intervention to trigger interrupt of the NPU
- To start a graph, the host will:
  - Put execution plan in shared buffer (B)
  - Put input data in shared buffer (B)
  - Trigger the interrupt (I)

# AI Driven Zonal Architecture Applied

## HW + SW Perspective

# SoC Zonal Architecture Overview




**Safe Compute Cluster**

Real-Time Cores

Shared L2 Memories

**Performance Compute Cluster**

App Cores

Shared L2 Memories

**Processing Accelerators**

DSP

NN

Interconnect Fabric

System Peripherals

Safety/Security Islands

Central SRAM

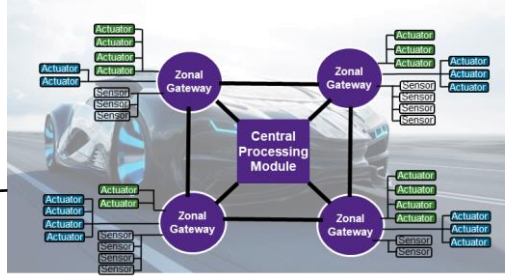
Communication

PCIe

Ethernet

USB

# Software Perspective- Virtual Machine Architecture




| **Virtual Machine** Secure Domain ASIL-B | **Virtual Machine** Open Domain QM | **Virtual Machine** Secure Domain ASIL-D | **Virtual Machine** Secure Domain ASIL-B..D | **Virtual Machine** Secure Domain ASIL-D |
|---|---|---|---|---|
| **Sensor Fusion** | **"User Apps"** | **Vehicle Operation** | **ADAS RADAR/Lidar** | **ADAS - AI** |
| **Trusted OS** | **Rich OS – Linux** | **Trusted OS** | **Trusted OS** | **Trusted OS** |

Hypervisor

Real Time CPU | Application CPU | NPU (NN) | VPU (Vector DSP) | Safety & Security Mgmt

Processor Interconnect with Protection, QoS, Availability

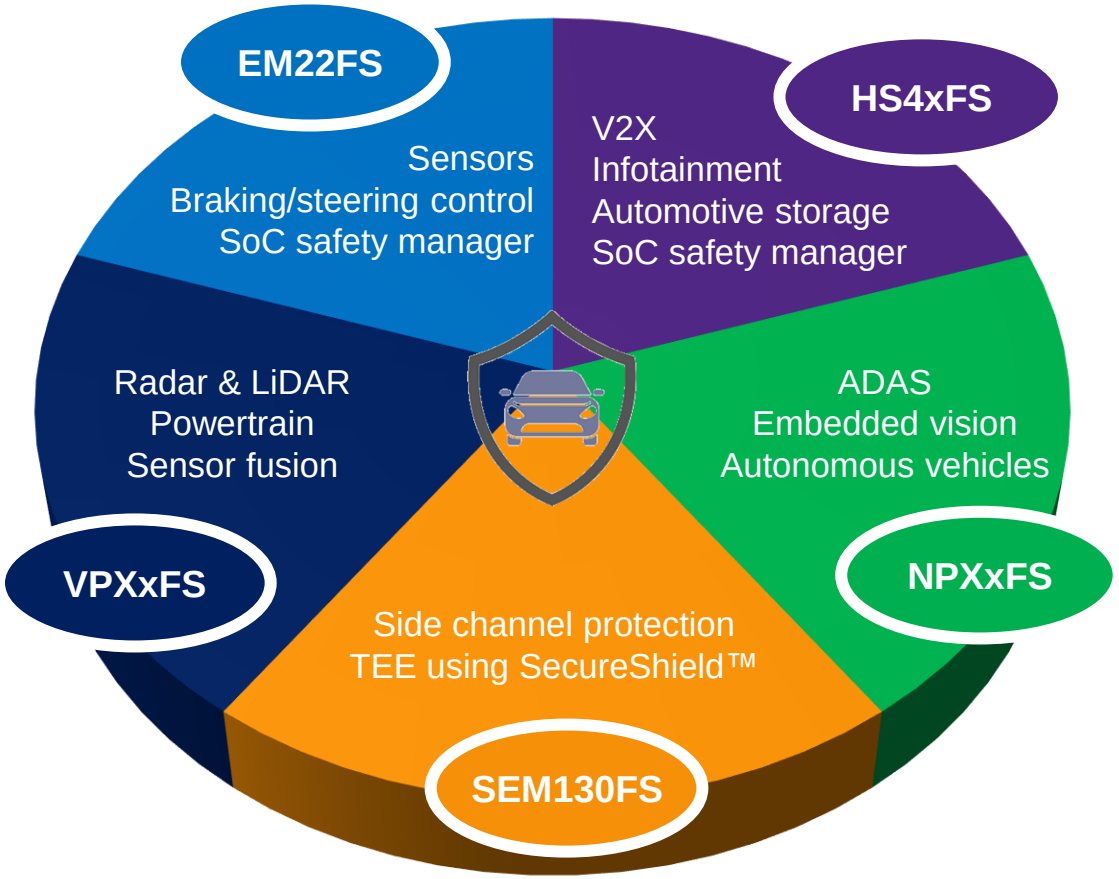
Memory

Ethernet Gateway

# ARC Functional Safety (FS) Processors

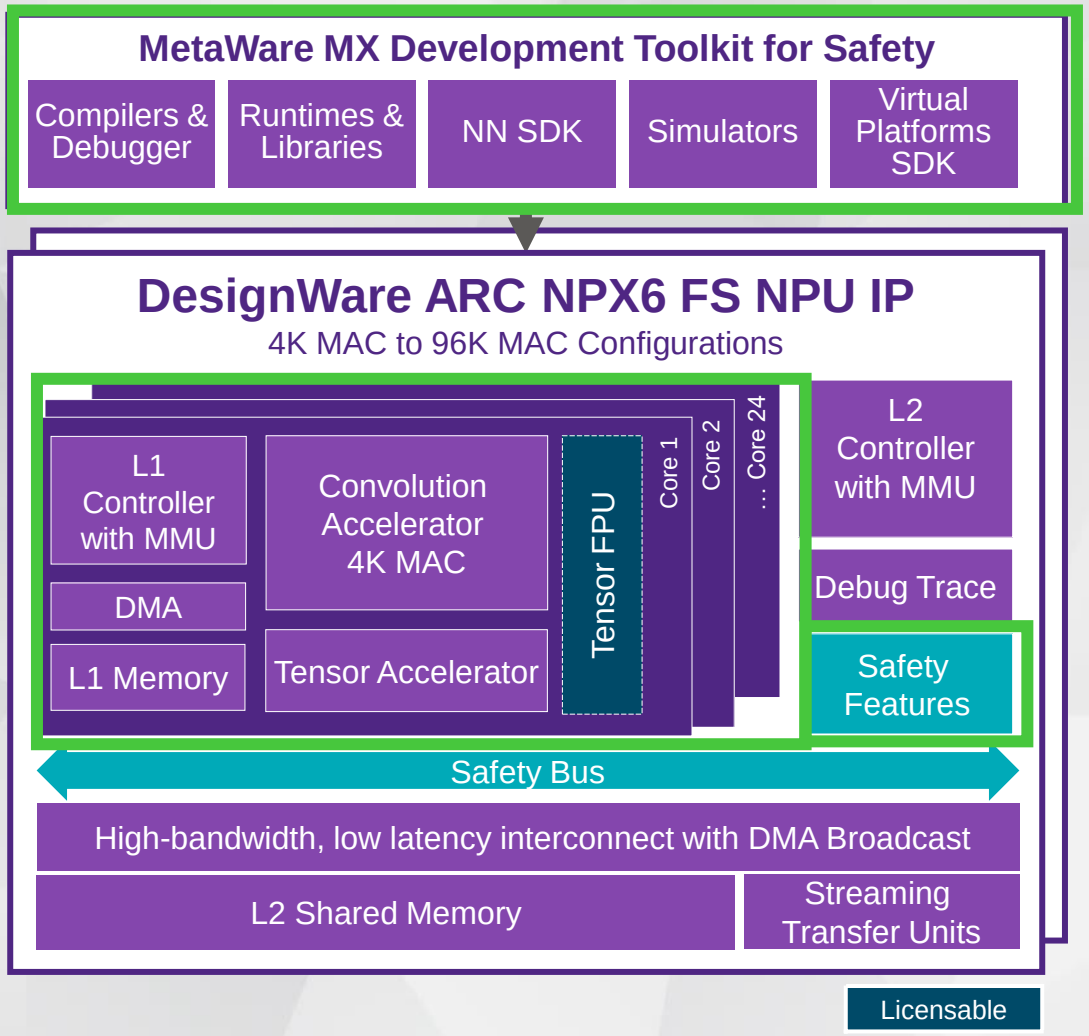
*ISO 26262 ASIL Compliant Cores for Automotive Applications*

SGS TÜV SAAR ASIL D COMPLIANT Functional Safety ISO 26262 www.sgs-tuev-saar.com




- Safety-enhanced cores span the ARC portfolio to address broad range of automotive applications
- Industry's First Processor IP Certified for Full ISO 26262 ASIL D Compliance
- ARC MetaWare Development Toolkit for Safety speeds ISO 26262-compliant software development
- FuSa Software Stack - ASIL certified embedded components for use in safety-critical applications
- Over 80 safety work products developed, accelerating customers' functional safety assessments

# DesignWare ARC NPX6FS w/ 340 TOPS* for Safety

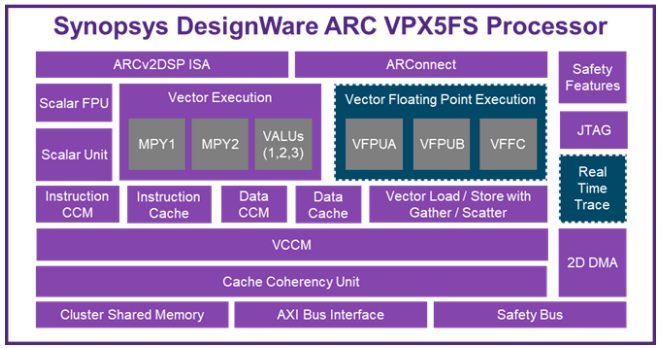



- **ISO 26262 compliant NPX6FS architecture**
  - 1 to 24 core NPU up to 96K MACS (340 TOPS*)
  - Multi-NPU support (up to eight for 2700 TOPS*)
  - **Functional Safety** support with ASIL B / ASIL D compliance
- **Integrated safety-critical hardware features** including Lockstep with time diversity, transient and permanent fault protection, memory ECC, Watchdog Timer, LBIST/diagnostic error injection, MMU, etc.
- **ARC MetaWare Development Toolkit for Safety** speeds ISO 26262-compliant software development
- **Safety documentation:** FMEDA reports & safety manuals speeds functional safety assessments
- **Virtualization** assistance – mixed criticality support for zonal architectures

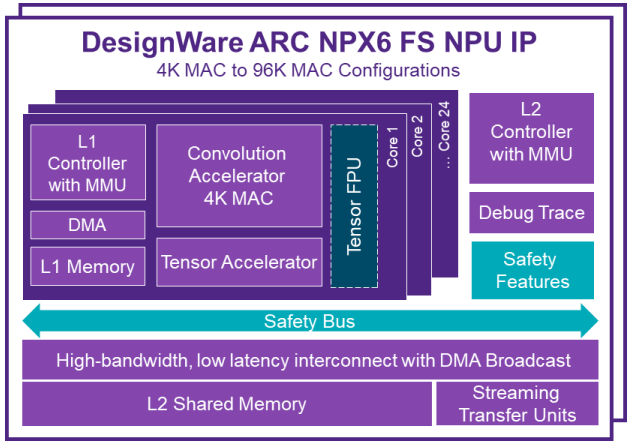
** 1 GHz,5nm FFC worst case with sparse EDSR model, SVT only*

# ARC FS Real-time, Application and Sensor Processing Cores

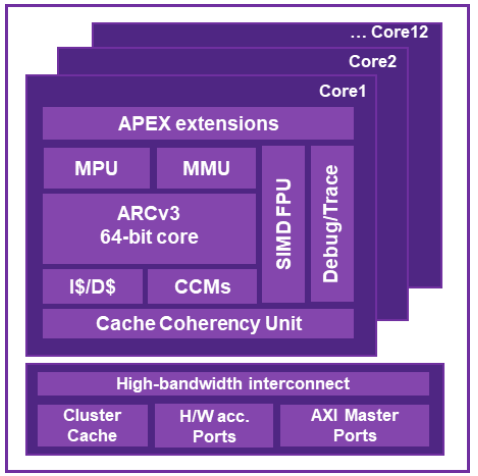
*ARC VPX5FS DSP, NPX6FS NPU and HS6x Processor*




**VPX5FS DSP**




**NPX6FS Neural Processing Unit**




**ARC HS6x Host Processor**

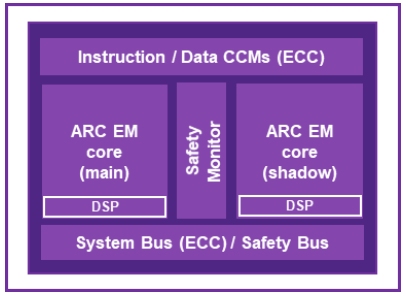
- Multicore vector DSP addresses ADAS sensors (LiDAR, RADAR), powertrain, sensor fusion, etc.
- SIMD/VLIW design for massive parallel processing
- Multiple vector FP engines for high precision results

- Addresses AI and vision applications: augmented reality, ADAS, surveillance, etc.
- 1 to 24 core scalable NPU up to 96K MACs executes graphs for object detection and scene segmentation up to 340 TOPS
- Automatic graph partitioning using MetaWare MX for improved performance, bandwidth, latency

- ARCv3 64-bit multi-core processor
- Configurable as real-time and/or application processor
- Support for up to 12 CPU cores and up to 16 user hardware accelerators
- 35% lower power (uW/MHz) than Cortex-A65AE

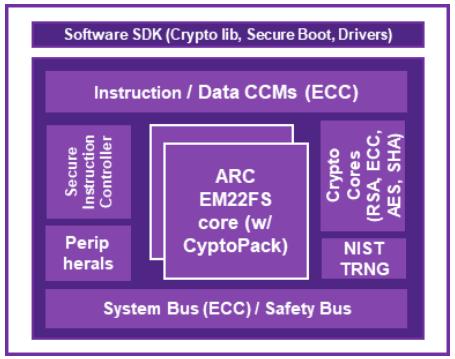
# Safety, Security Management & Real-Time Control

*ARC EM Safety Processor, tRoot HSM, ARC HS4xFS Real-time Processor*



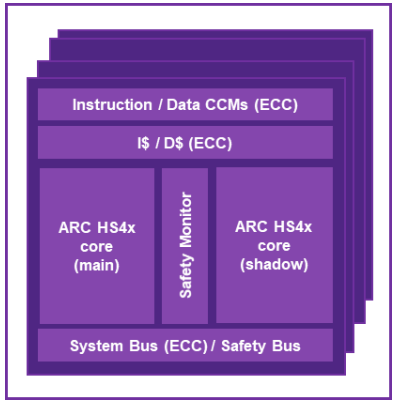

**EM22FS Safety Management Processor**

- Dual-core lockstep implementation with hybrid mode support
- Dedicated safety monitor validates DCLS operation and collects SoC level error info
- ECC for closely coupled memories, MPU, user Programmable Watchdog Timers
- FuSa safety management S/W stack available




**tRoot Hardware Security Module**

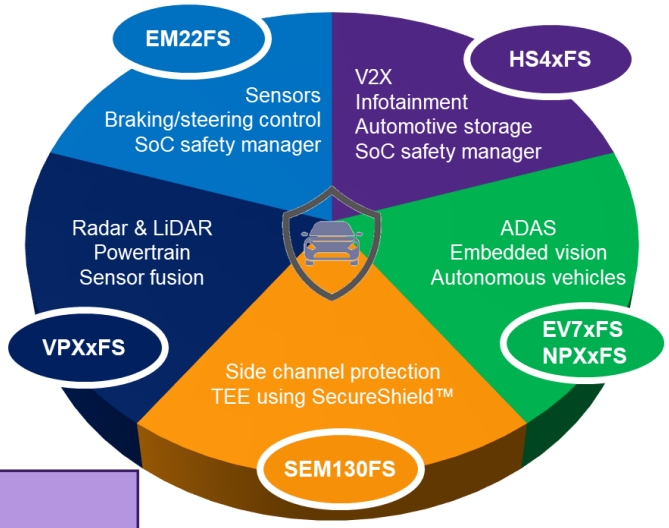
- Scalable cryptography: custom instructions (CryptoPack) to crypto cores with side channel protection
- NIST-compliant TRNG
- Secure Instruction Controller with side channel protection for secure external memory access
- Software: secure applications SDK, crypto library, device drivers & reference designs




**HS4xFS Safety Enhanced RT Controller**

- Dual-issue, 10-stage pipeline processor, configurable in lockstep or hybrid modes
- Single-core and quad-core options (DCLS)
- Industry leading integrated H/W safety features
- 20% higher single core performance than Cortex-R52

# SoC Zonal Architecture Overview




**Safe Compute Cluster**

HS4xFS

Real-Time Cores

Shared L2 Memories

**Performance Compute Cluster**

HS4xFS

App Cores

Shared L2 Memories

**Processing Accelerators**

DSP

NN

VPXxFS

EV7xFS
NPXxFS

Interconnect Fabric

System Peripherals

Safety/Security Islands

EM22FS

SEM130FS

Central SRAM

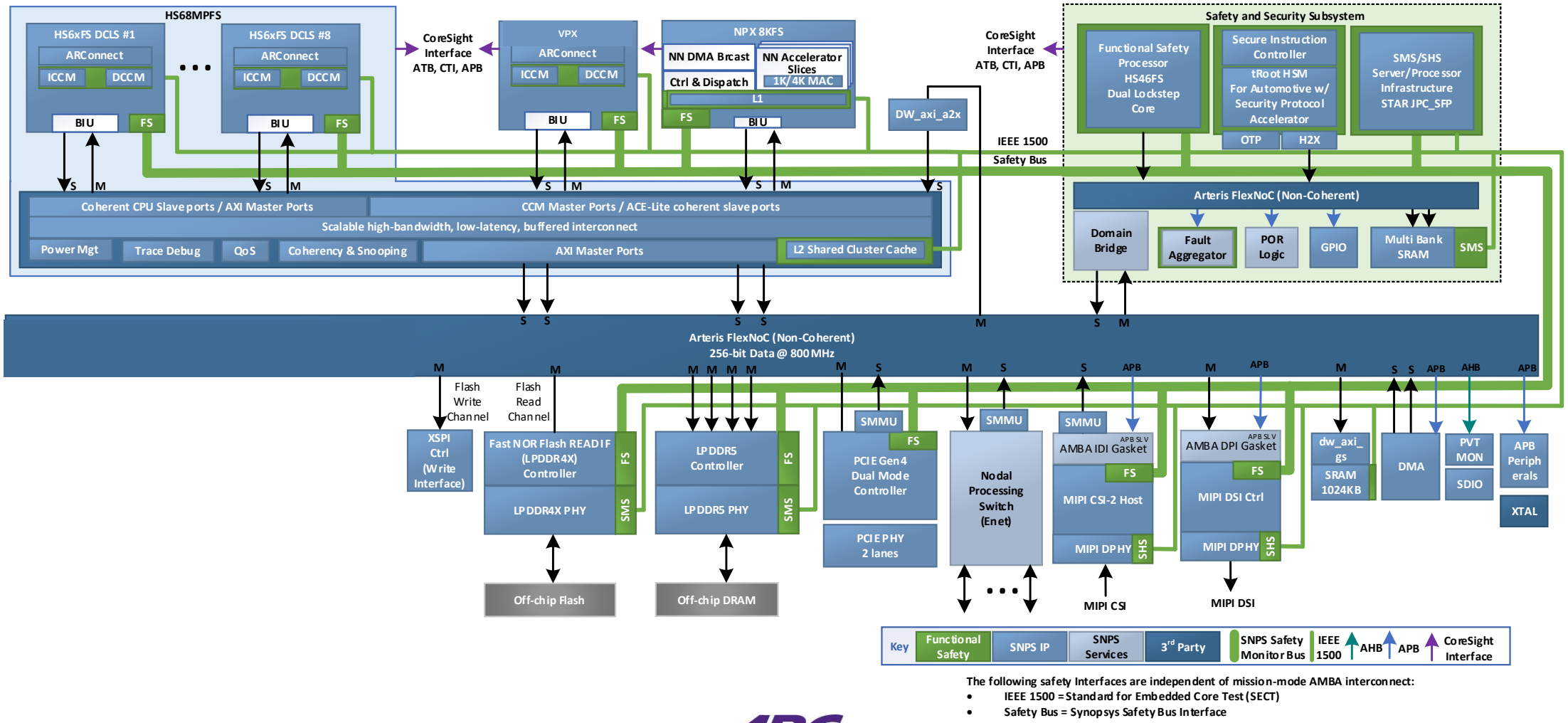
Communication

PCIe

Ethernet

USB

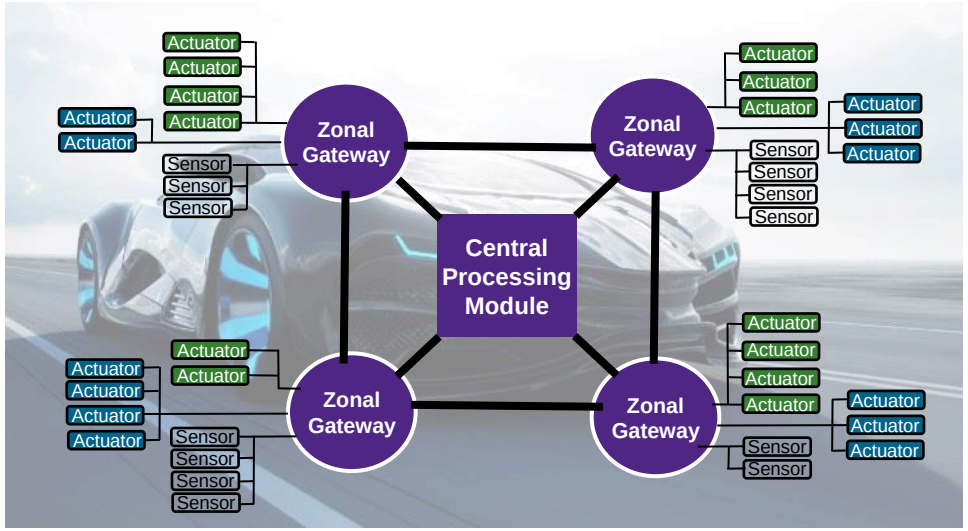
# Zonal Architecture Reference Design




The following safety Interfaces are independent of mission-mode AMBA interconnect:

- IEEE 1500 = Standard for Embedded Core Test (SECT)
- Safety Bus = Synopsys Safety Bus Interface

# AI aware Zonal Architecture is Reshaping Automotive SoCs

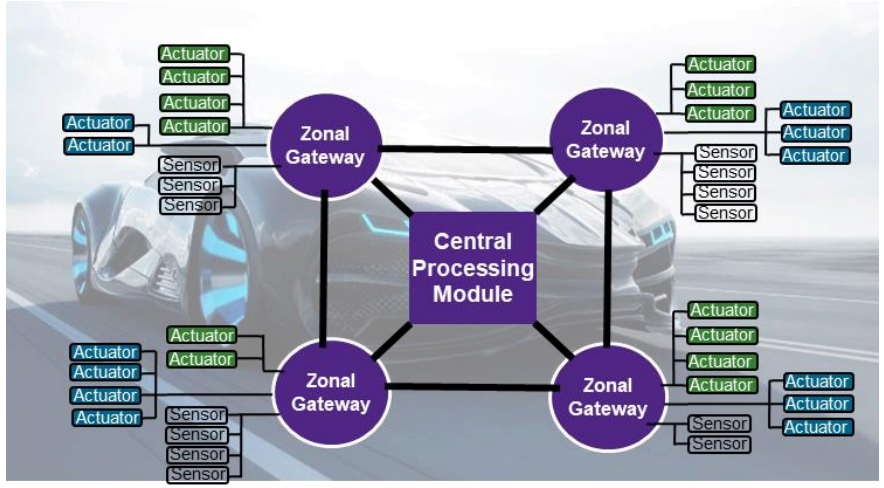



**OEMS**

**T1/T2s**

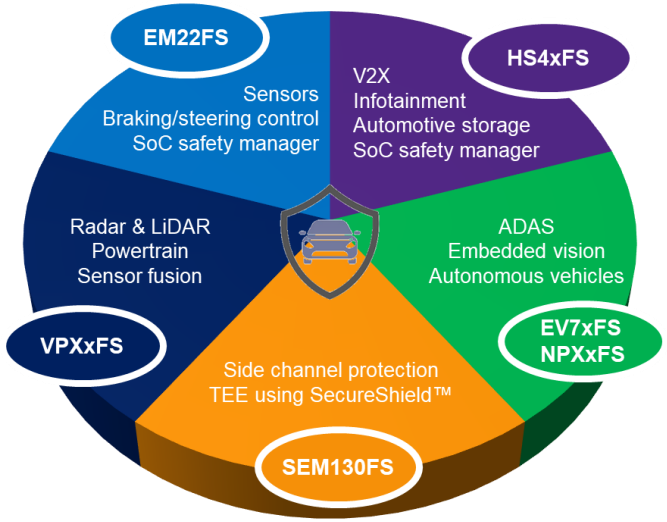
**OEMs/Tier 1s/Tier 2s publicly announced investigation into Zonal / Central Architecture**

# Summary




- **Why** virtualization of AI Accelerators?
  - Required for safety & security isolation between virtual machines
  - Ease of independent software development and partitioning
    - Independent OS-choice
    - Independent certification
  - **Improved hardware utilization and Silicon cost efficiencies** by sharing hardware resources
  - Virtualization as a **safety mechanism** and **requirement for spatial and temporal isolation**

- **When** will we see Virtualization of AI accelerators in our vehicles:
  - SoCs already Taping Out for T1, OEM and partner software development



SYNOPSYS®
Silicon to Software™

# Thank You