

# Optimizing Data Transport Architectures in RISC-V SoCs for AI/ML Applications

RISC-V Summit Europe 2024

Ashley Stevens  
Director of Product Management & Marketing

June 2024

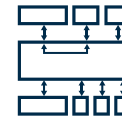
**ARTERIS** 

# Challenges of RISC-V Based AI/ML SoCs

Diverse interface protocols (ACE, CHI, ACE-Lite, AXI....)



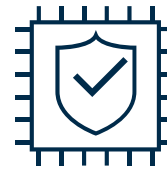
Varying coherency models (MESI, MOESI)



'Memory wall' - Massive memory bandwidth of AI/ML



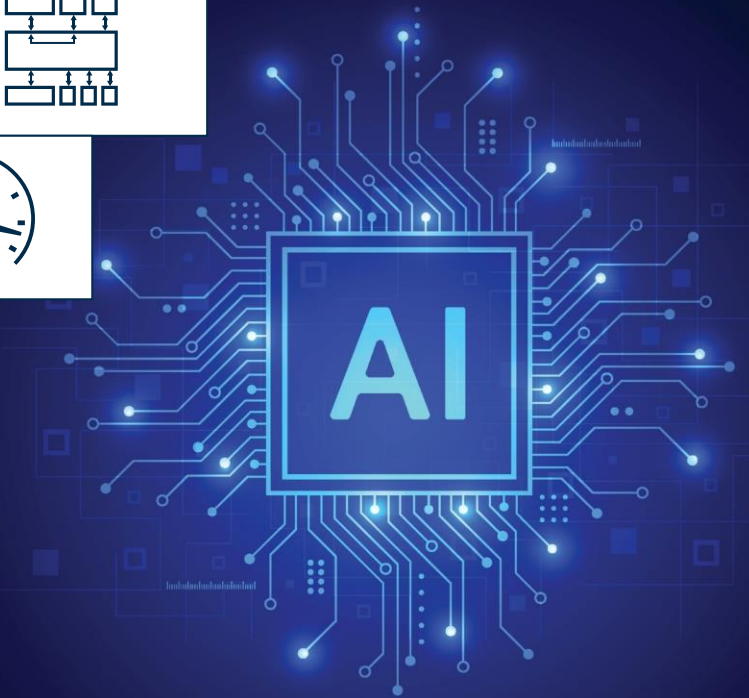
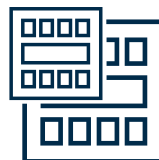
Safety standards for automotive functional safety



Verification / Performance models / FPGAs



Physical implementation (PD)



# System IP and Network-on-Chip (NoC) SoC Interconnect IPs

Networking techniques for improved on-chip communication & data flow



Smaller Die Area



Lower Power Consumption



Faster Frequency Lower Latency



Shorter, Predictable Schedules



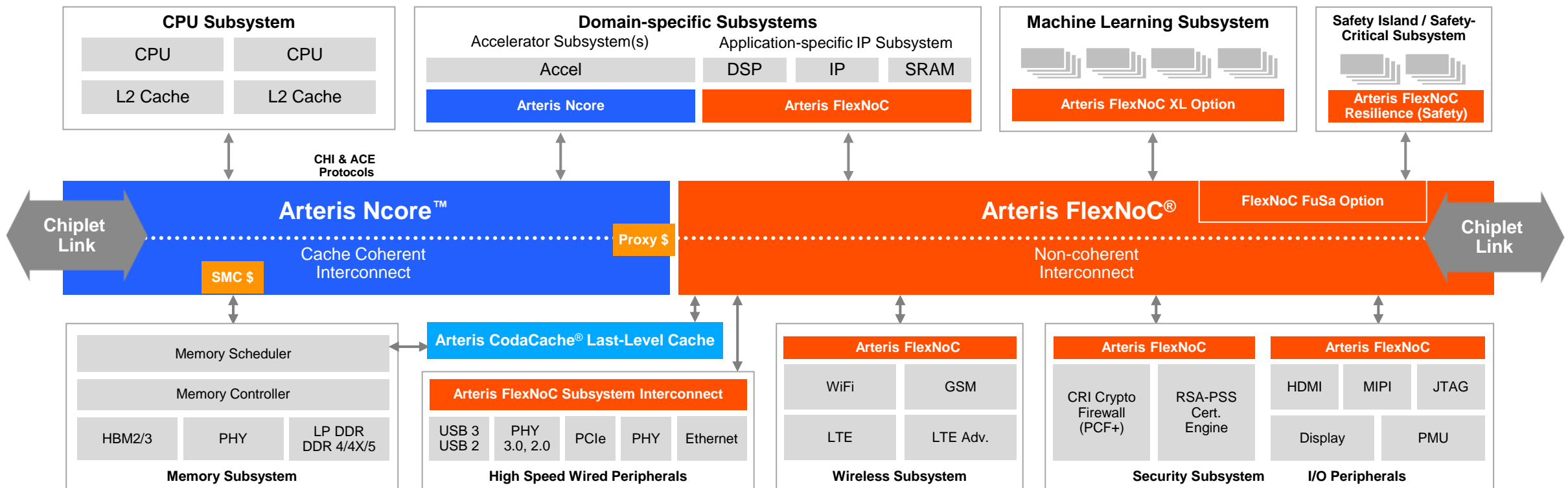
Rapid Timing Closure Estimation



Automated Verification



Easy Configuration



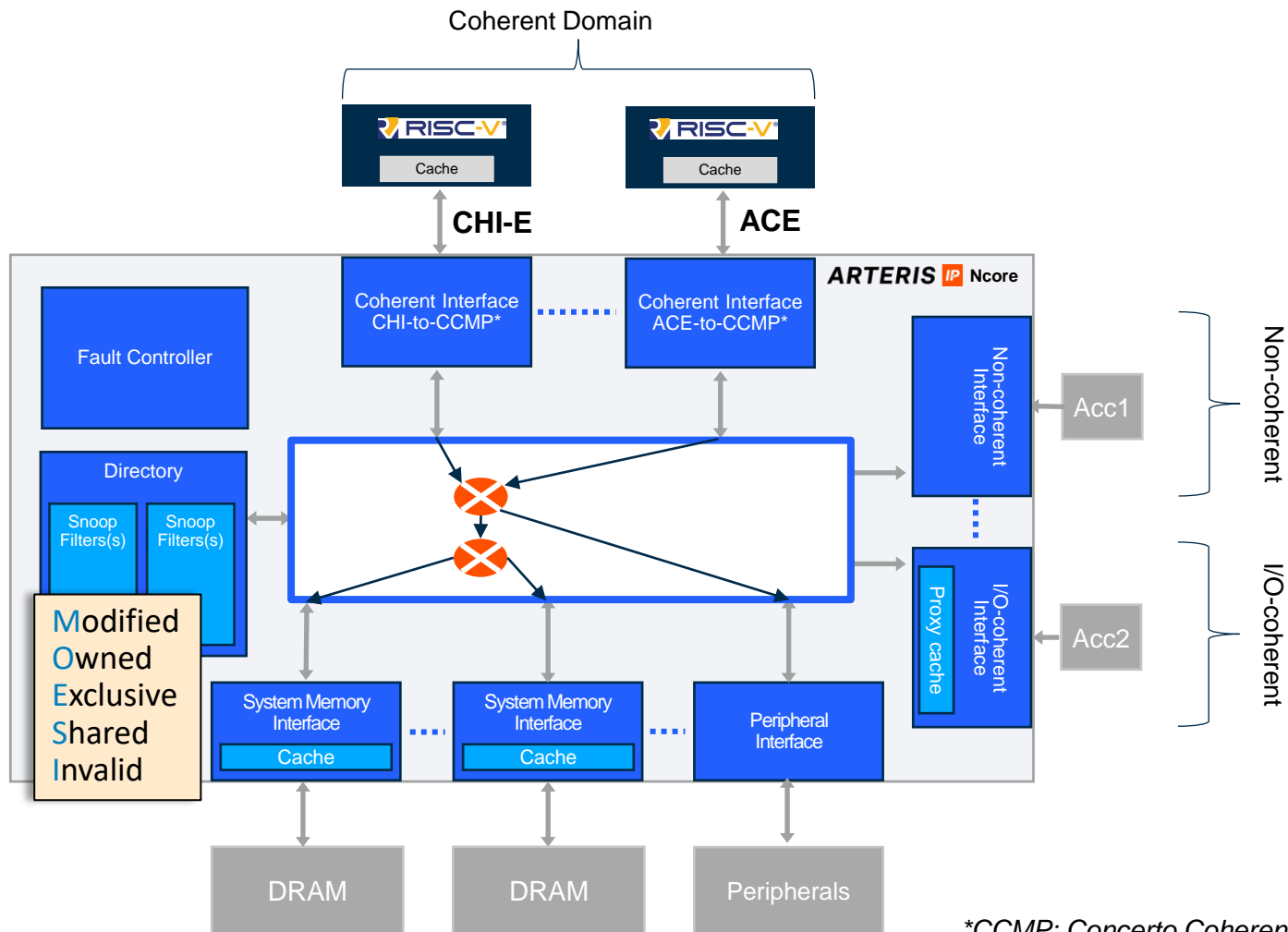
Arteris Ncore™ cache coherent interconnect IP

Arteris FlexNoC® non-coherent interconnect IP

Arteris CodaCache® last-level cache

# Arteris Ncore Configurable Coherent Interconnect

## Multi-protocol coherent interconnect

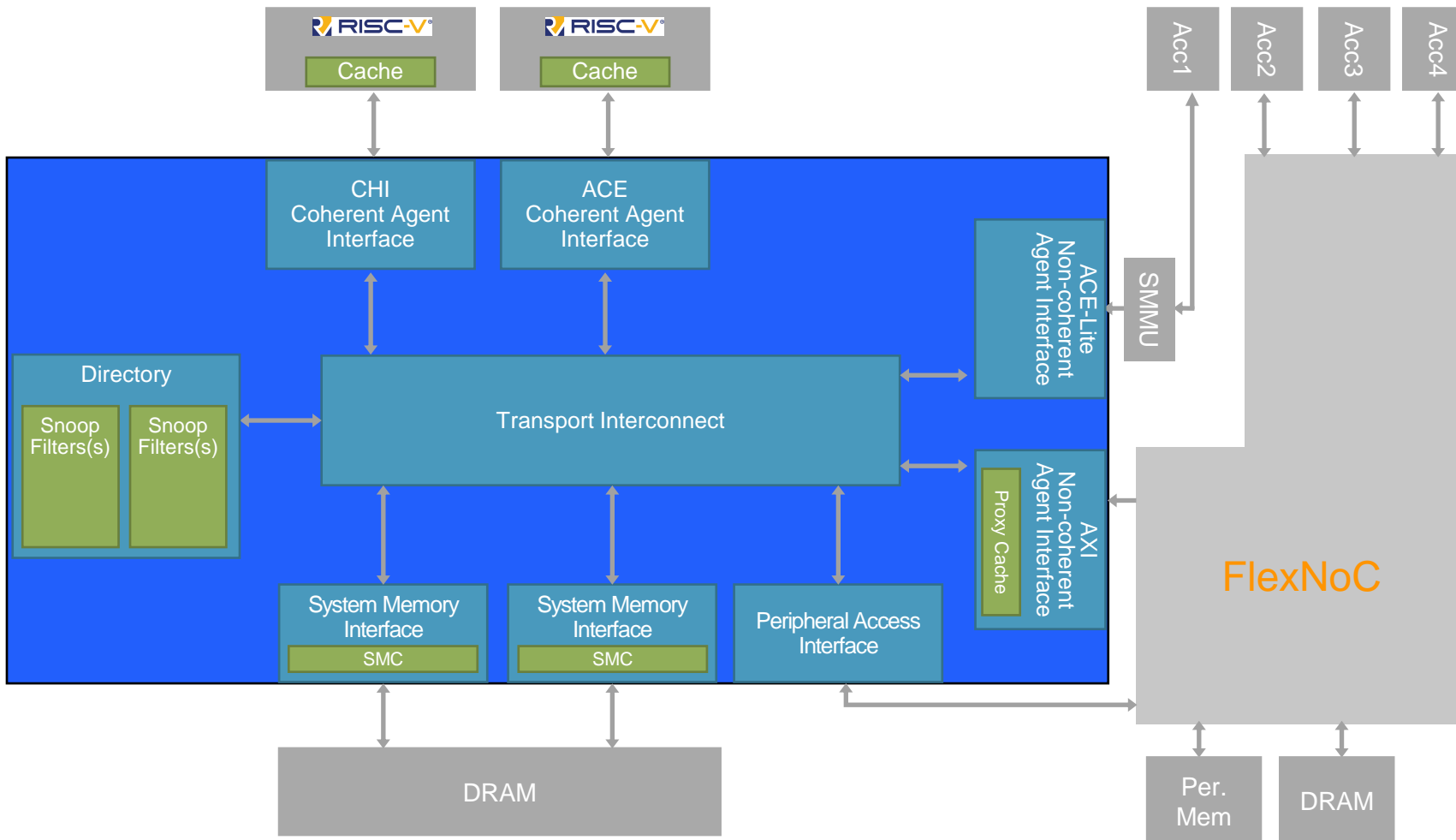


\*CCMP: Concerto Coherent Messaging Protocol

- AMBA protocols converted to internal Arteris CCMP protocol
  - Coherent Interfaces: CHI-B, CHI-E or ACE, **interoperable**
  - Arteris internal protocol supports MESI and MOESI coherency models
- I/O Interfaces:
  - ACE-Lite, AXI
  - Optional Proxy Cache participates in coherency domain as fully coherent cache
- Memory interface with optional system memory cache
- Peripheral Interface for I/O targets
- Directory with snoop filters
- Fault controller for functional safety option
- Transport created from switches

# Why Use AMBA CHI and ACE in the Same System?

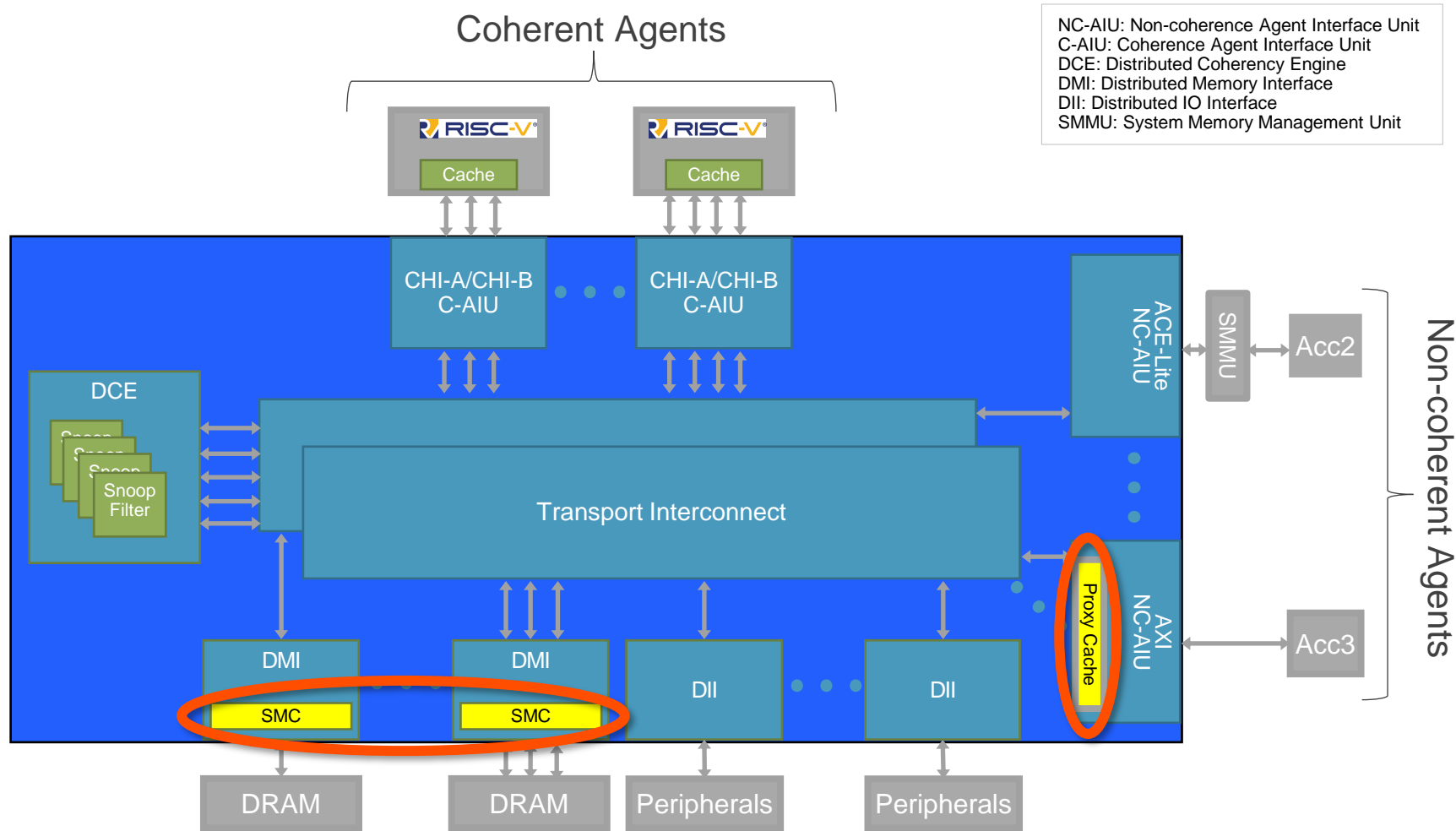
RISC-V dynamic ecosystem diversity



- RISC-V is a diverse and evolving ecosystem
- Mixed ACE/CHI can ease integration of new and legacy processors
  - Mix latest high-performance RISC-V clusters using CHI with older RISC-V CPUs using ACE
  - Leverage investment in ACE IP
- Proxy caches ease integration of non-coherent accelerators into the coherent domain

Ncore is verified with SiFive X280 RISC-V CPUs

# Proxy Cache & System Memory Cache (SMC)



## Proxy cache

- Configurable up to 8MB, 1-16 ways
- Cache for non-coherent or I/O coherent accelerators
- Fully coherent with caches and memories in the system
- Reduces accelerator traffic into coherency system
- Smooths accelerator traffic with varying bursts into 64B coherency granules

## SMC

- System Memory Cache per distributed memory interface
- Configurable up to 8MB, 1-16 ways per DMI
- Scratchpad, partitioning, atomics
- Cache Maintenance Operations

Both can be configured with parity or ECC for FuSa systems

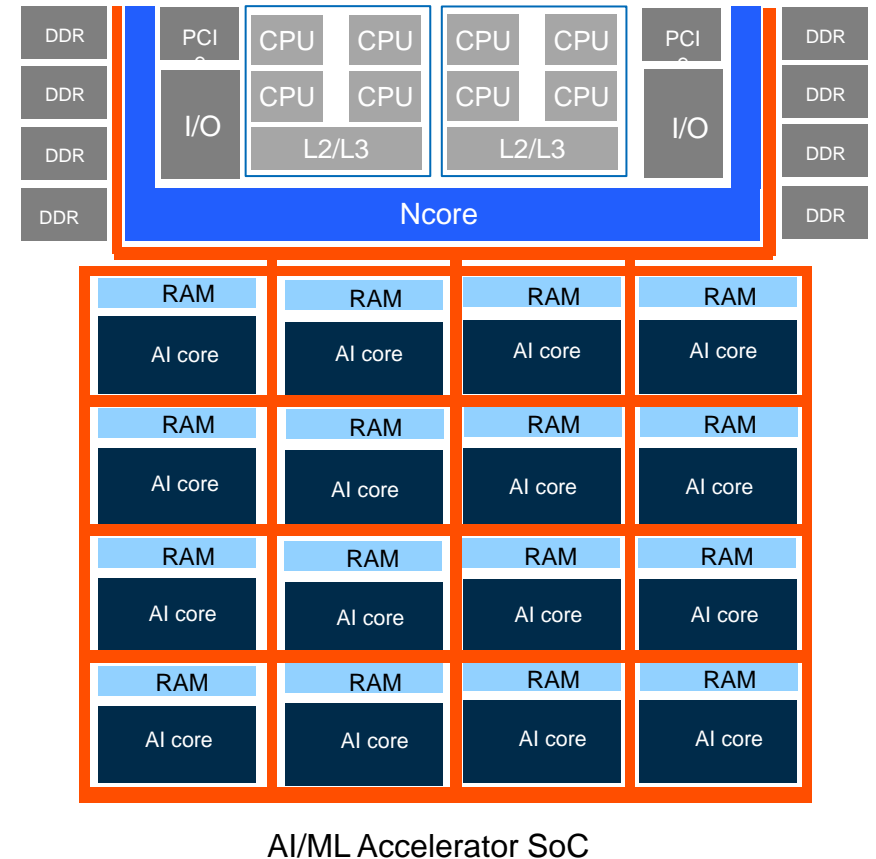
# Efficient and Performant AI/ML Data Transport Architecture

Optimal solutions combine coherent and non-coherent NoCs

- Coherent NoCs required for data shared with cached CPUs
  - Coherent systems work on 64B coherency granules (512b cache line)
- Extreme bandwidths in AI/ML devices
  - Local memories may reduce traffic to external memory
  - Separate shared and non-shared memory traffic
- Provide a fast and wide path to memory for non-shared traffic
- Combine coherent and I/O-coherent NoCs for optimal performance
  - Coherent hub close to the cached CPUs with narrower buses
  - Wide NoC connects the rest of the SoC including AI core array
  - Mesh topology can be appropriate for AI applications

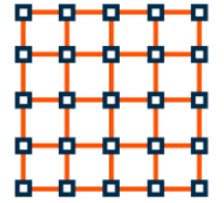
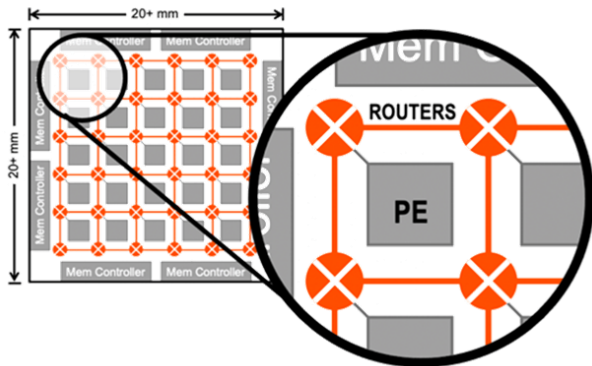
**Ncore 3** coherent interconnect provides the coherent hub

**FlexNoC 5** connects the AI core accelerator units



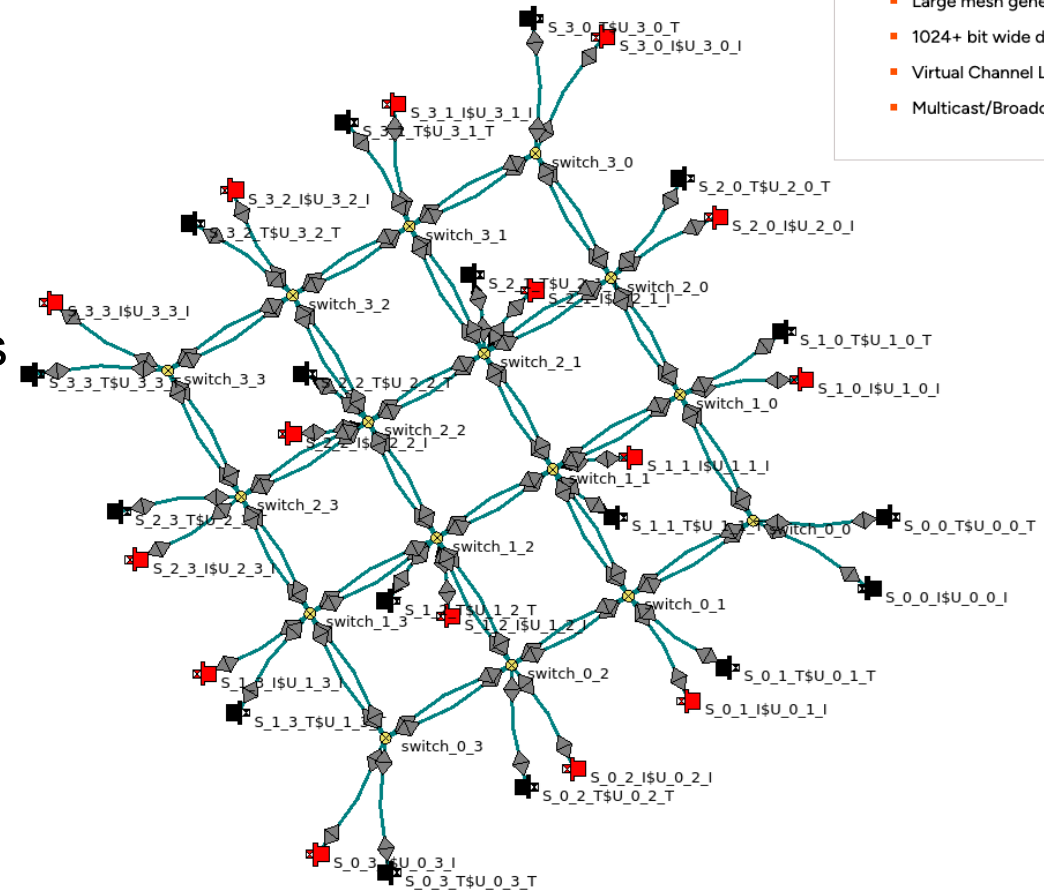
# AI Bandwidth Demands Met with FlexNoC 5 XL

- FlexNoC<sup>®</sup> 5 XL addresses non-coherent bandwidth requirements of AI/ML systems
  - Large capacity mesh generator
  - Up to 2048-bit wide connections
  - Up to 200 Network Interface Units (NIUs)
  - Up to 512 Pending transactions
- Quality of Service ensured by virtual channels
- Multi-Cast/Broadcast Stations
  - Broadcast to multiple units to reduce bandwidth



## Large Systems: XL Option

- Large mesh generator
- 1024+ bit wide data connections
- Virtual Channel Links
- Multicast/Broadcast Stations

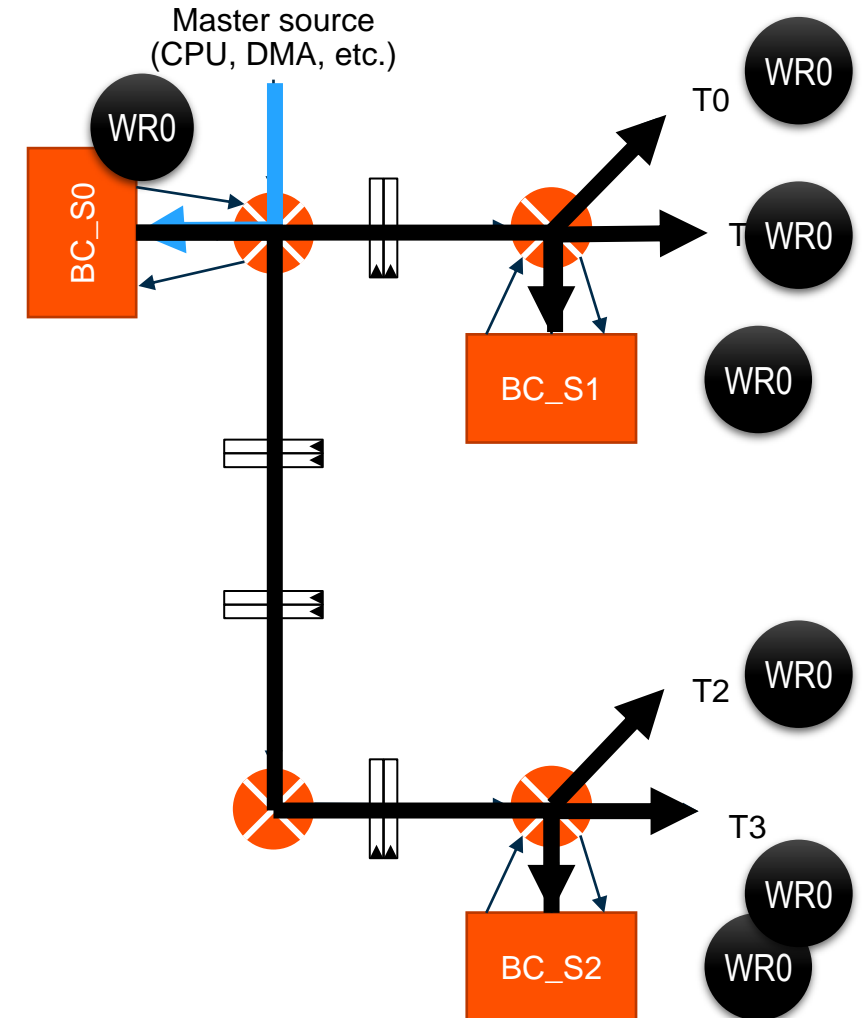




# FlexNoC Intelligent Multicast Write for AI/ML

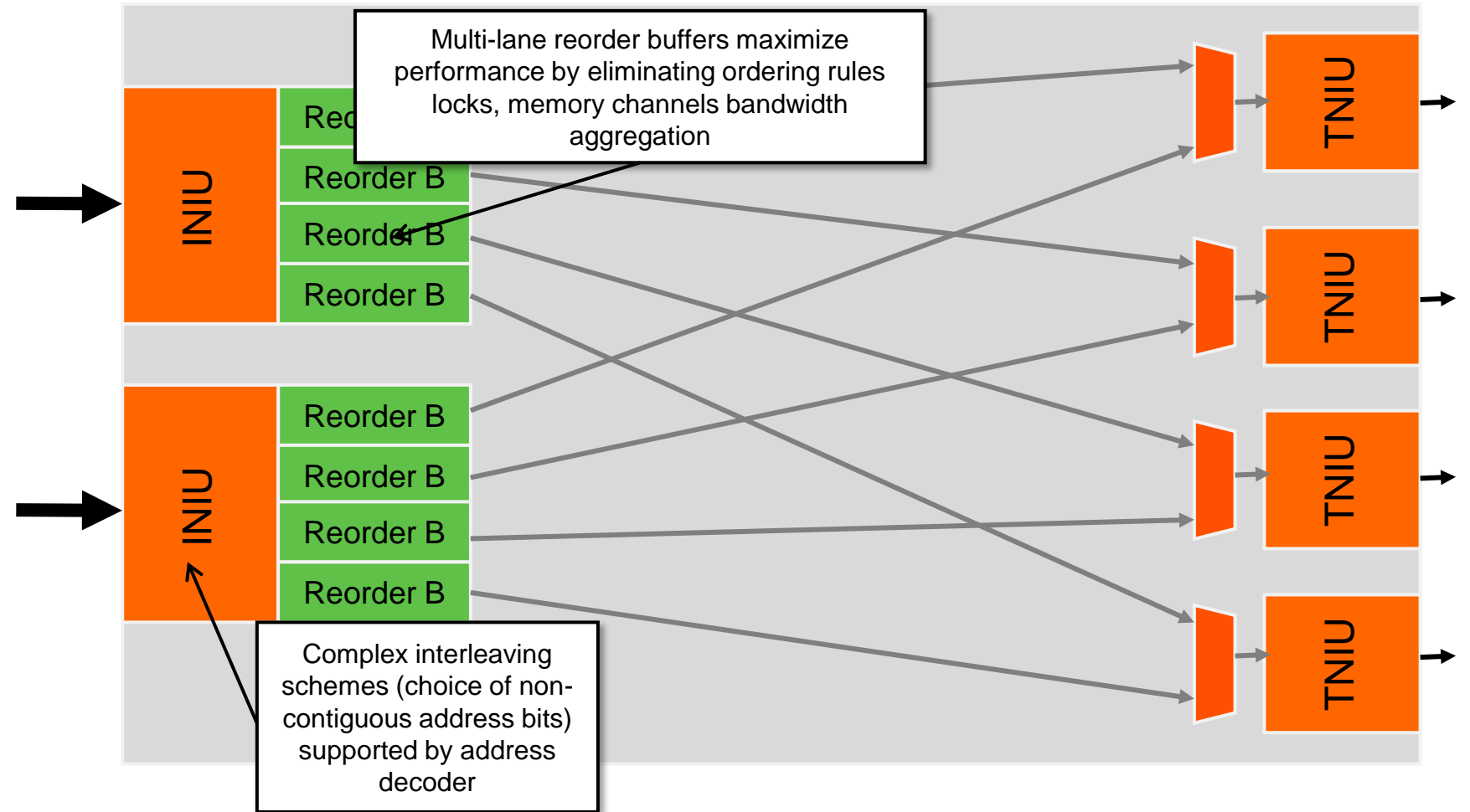
Efficient multicast – bandwidth saving

- Broadcast station optimizes use of NoC bandwidth
  - Broadcasts performed as close as possible to the destination
  - Any number of broadcast stations in a FlexNoC
- Writing to broadcast station will cause it to send posted writes to multiple destinations
- Used in AI for DNN weight and image map updates



# High Memory Bandwidth from Interleaving Channels

- Up to 8 or 16 channels interleave
- Read-reorder buffers
- Traffic aggregation / data width conversions
- Up to 1024 bits wide connections
- Non power-of-2 interleave supported



# Improved Productivity & Configurability

## Save time and resources with library reuse and automation

### Manual “from scratch” development

Quick estimates, simplified assumptions:  
Area, performance, timing closure, power

Manually-created topologies

Change requests:

- Add/remove interfaces
- User bits, QoS, address map, safety, buffering, service, probes, interrupts, modules, etc.
- ... **cause significant changes to RTL**

Floorplan & timing closure issues:

- Add/remove interfaces
- Interface location, blockages, fences
- Iterations for timing
- ... **causing even more changes**

### Chip Specification

- Power/clock domains
- Floorplan
- Subsystems
- Memory Maps

### Architecture

Topology  
Routing  
Architectural  
Choices

### Mapping

Automatic  
choice of library  
elements

### Refinement

Optimization  
Parameter Tuning  
Special Features

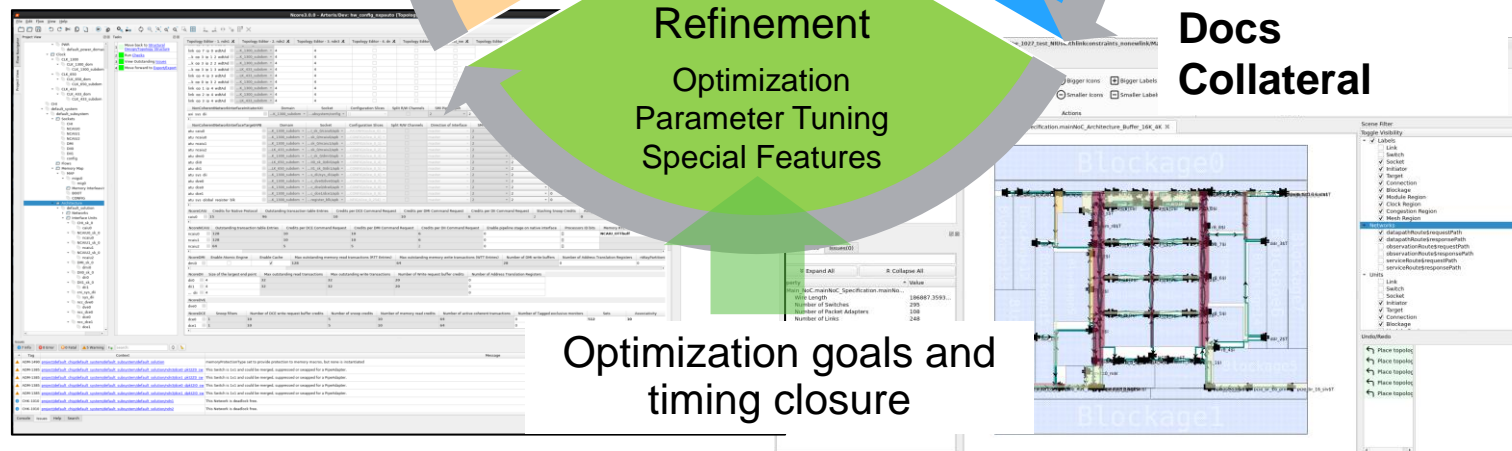
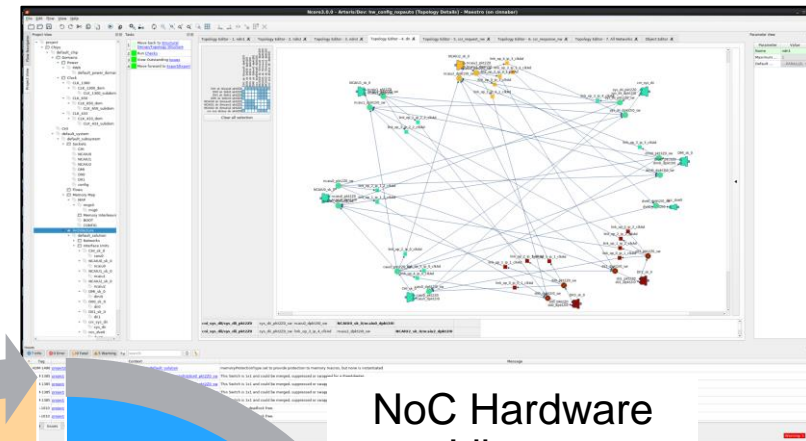
NoC Hardware  
Library

Inputs

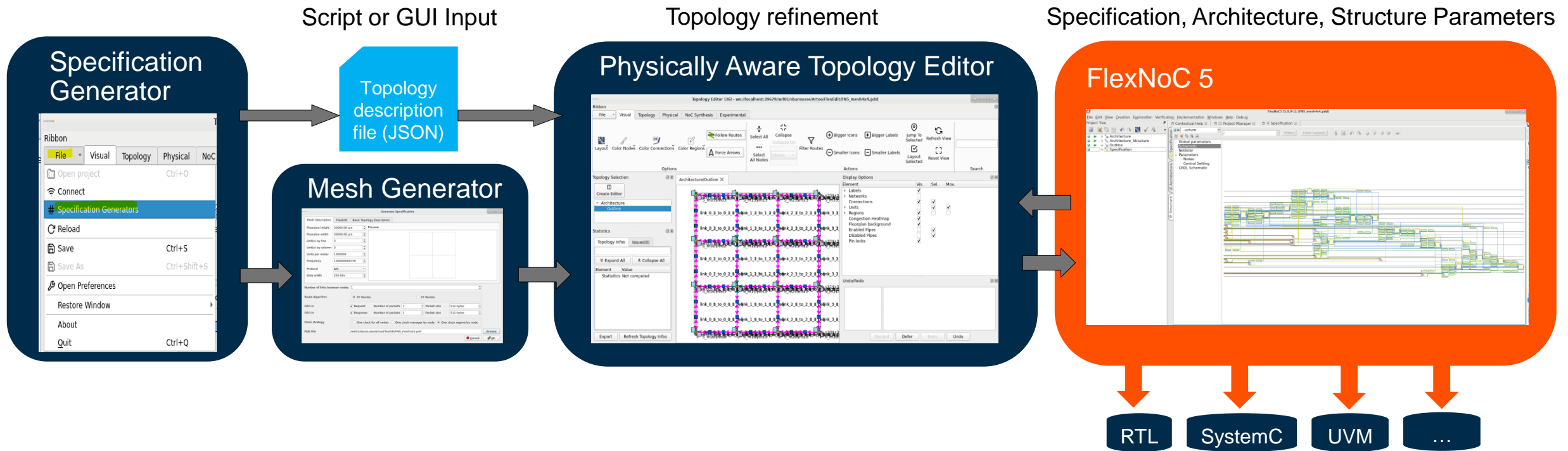
Outputs

RTL  
Scripts  
Docs  
Collateral

Optimization goals and  
timing closure



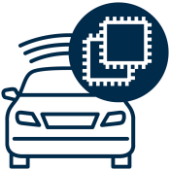




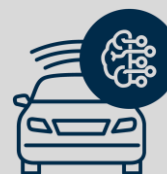
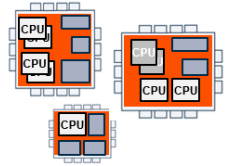
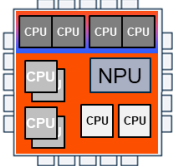
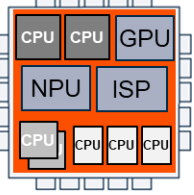
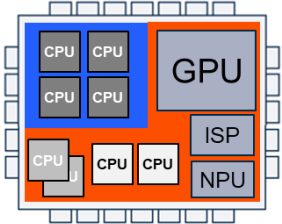
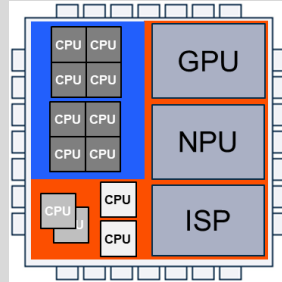
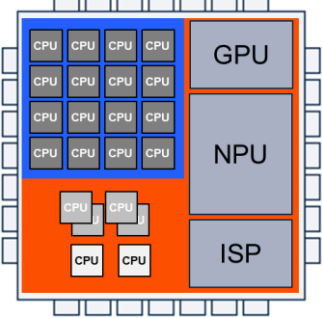
# FlexNoC Physically-Aware Mesh Topology Generator Flow



SystemC and UVM models enable system level simulation

# Automotive Domains and Their Complexity

Cache coherency is required in safety-critical systems

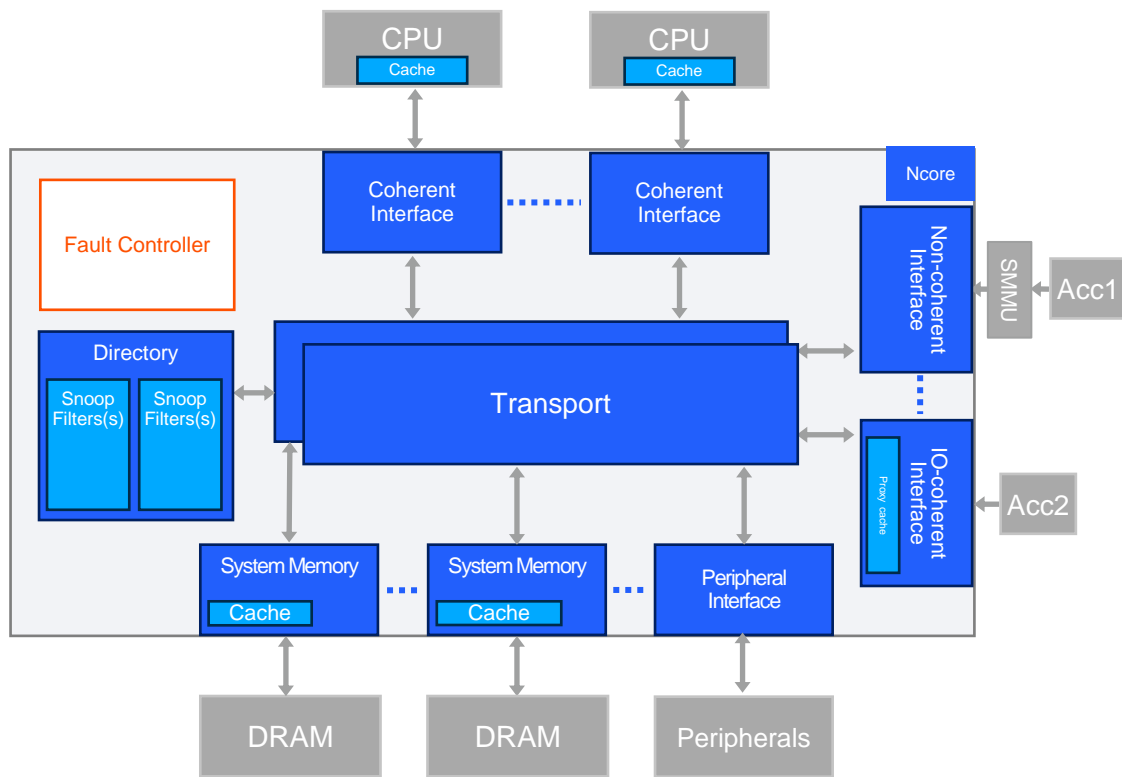
Auto MCU	Zonal Controller	Vision	Cockpit	L2+ ADAS	Autonomous
					
Non-coherent	Coherent or non-coherent	Mostly non-coherent	Mix of coherent and non-coherent	Mix of coherent and non-coherent, some with large coherent meshes	Mix of coherent and non-coherent, large coherent meshes
					
Monolithic	Monolithic	Maybe chiplet (Sensors!)	Monolithic or maybe chiplet	Monolithic, trending towards chiplet	Likely mostly chiplet in the future

Chiplets increasingly important in future



# Challenge of Safety-certification for Coherent Systems

Automotive ADAS/autonomous driving is a key application of AI/ML

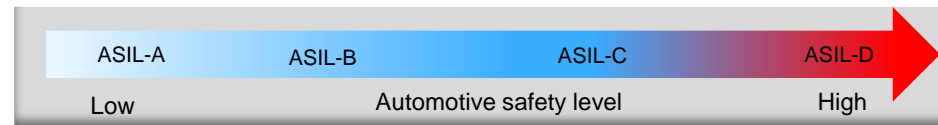


- The complexity of coherent systems makes safety certification especially challenging

- Ncore 3 safety/resilience capabilities:

- External ECC or parity
- Interface ECC or parity
- Interface duplication
- Cache/SF ECC or parity
- Transport link ECC or parity
- Directory duplication
- Fault controller/signaling

Ncore 3.4 is ISO26262 ASIL-D certified



The manufacturer may use the mark.

Revision 1.0 October 30, 2023  
Surveillance Audit Due November 1, 2026

Certificate / Certificat  
Zertifikat / 合格証

Arteris 21/05-038 C001

exida hereby confirms that the:

**Arteris Ncore 3.4**

**Arteris, Inc.  
Campbell (CA), USA**

Has been assessed per the relevant requirements of:

**ISO 26262 : 2018 Parts 2, 4, 5, 7, 8 and 9**

and meets the requirements providing a level of integrity to:

**Systematic Capability: ASIL D**

**Safety related function:**

The Arteris Ncore 3.4 IP is a configurable cache coherent interconnect that was developed as a Hardware Safety Element out of Context (SfEoC) with the following assumed top-level safety requirements:

- Data coherency shall be ensured
- The integrity of transactions between agents shall be ensured
- The integrity of memory data shall be ensured
- The integrity of the Ncore configuration shall be ensured
- Faults shall be signaled via a fault reporting unit

**Application restrictions:**

The Ncore 3.4 shall be used according to the requirements described in the Arteris Safety Manual.

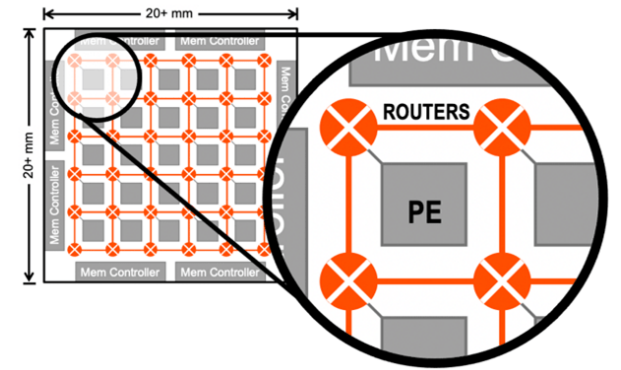
*A. Feilich*  
Evaluating Assessor

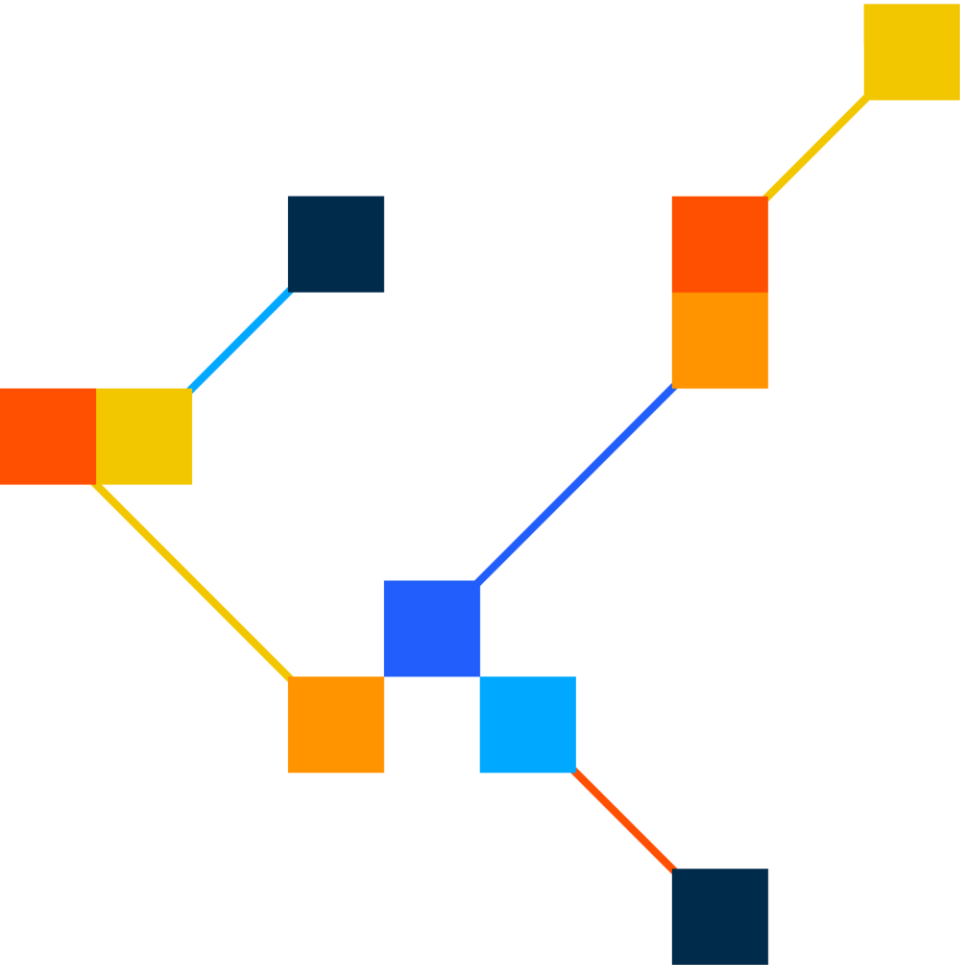
*C. Gong*  
Certifying Assessor

Page 1 of 2

# Summary

- Separate shared coherent traffic from high-bandwidth AI traffic where possible
- FlexNoC 5 Network-on-Chip XL option is suited to many AI designs
  - Mesh topology for large regular structures that align with physical layout
  - Wide buses for massive AI bandwidths
  - Broadcast writes for simultaneous updates of weights, map updates, and commands to AI units
- Tooling environments speed design iterations compared with point solutions
- Ncore is ISO 26262 certified to ASIL D and FlexNoC 5 is available with a safety package enabling safety for AI-enabled automotive
- Chiplets offer an additional optimization opportunity enabling modularity, scaling of systems, and cost reductions due to yield improvement from disaggregation across dies





**ARTERIS** **IP**

# Thank you

Arteris, Inc. All rights reserved worldwide. Arteris, Arteris IP, the Arteris IP logo, and the other Arteris marks found at <https://www.arteris.com/trademarks> are trademarks or registered trademarks of Arteris, Inc. or its subsidiaries. All other trademarks are the property of their respective owners.

Confidential © 2024 Arteris, Inc.