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

Michał Siwiński
EVP and CMO

RISC-V DAY
TOKYO 2025 SPRING

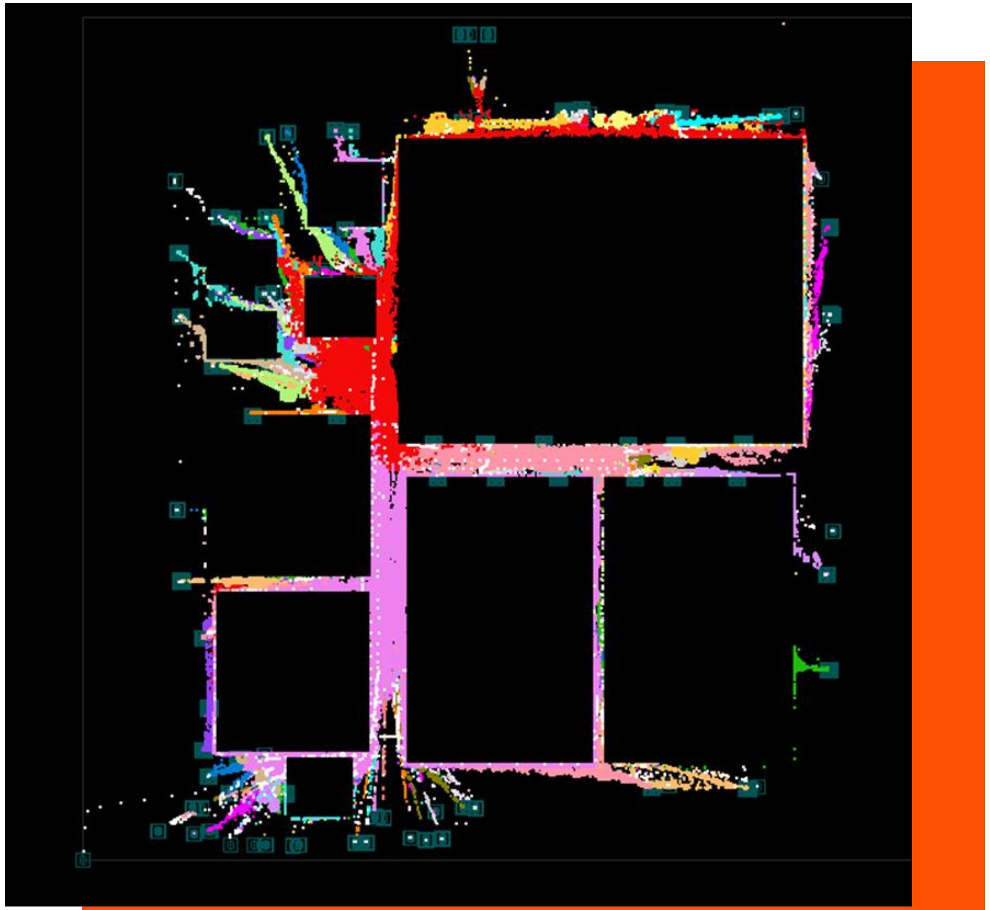
February 27, 2025
RISC-V Day Tokyo 2025

ARTERIS IP

Arteris – Connecting Innovation

- Arteris specializes in semiconductor System IP
  - **On-chip communications**
  - **SoC integration technologies**

- From architecture to IP assembly
  - to sub-system and chip integration
  - all with physical awareness



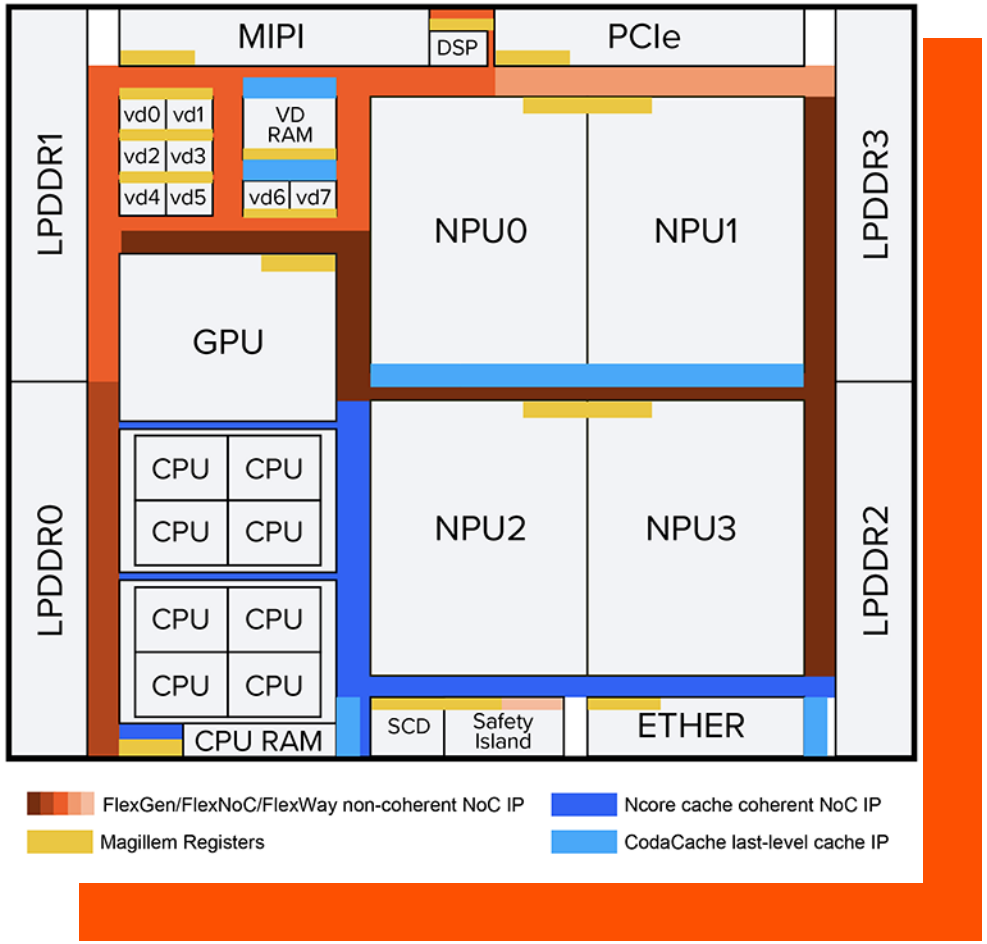
Arteris – Connecting SoCs

Arteris technology is **pervasive in SoCs**

- **5-20 NoCs** (Network-on-Chip) on chip or chiplet
- NoCs represent **10-13% of silicon**
- Registers represent **3-20% of silicon**

With **90+ patents**, Arteris addresses critical SoC design challenges

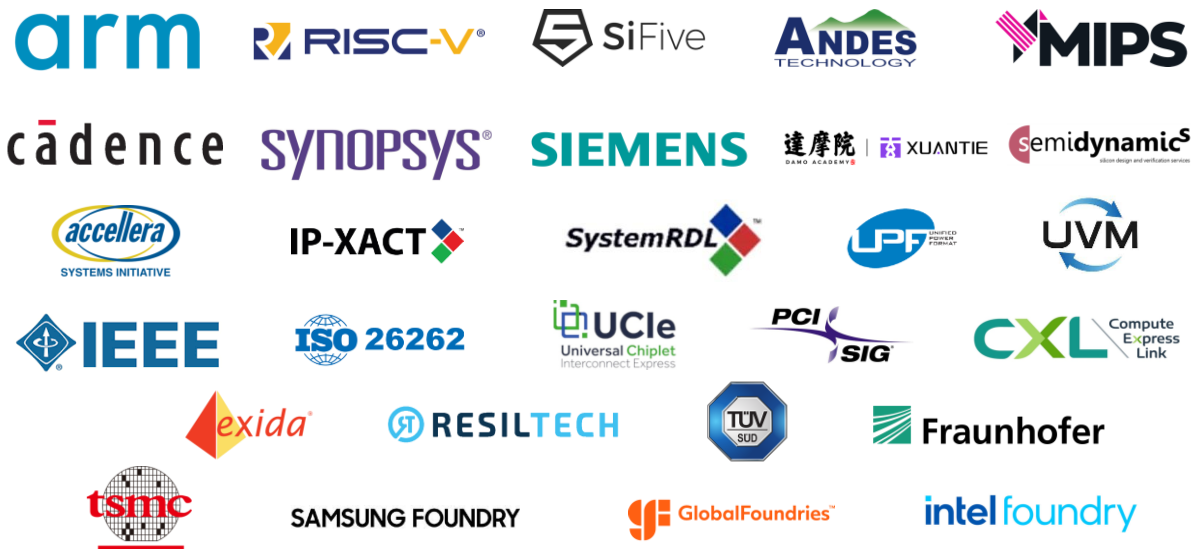
- Integrating AI
- Functional safety
- Power efficiency
- Performance
- Die area optimization

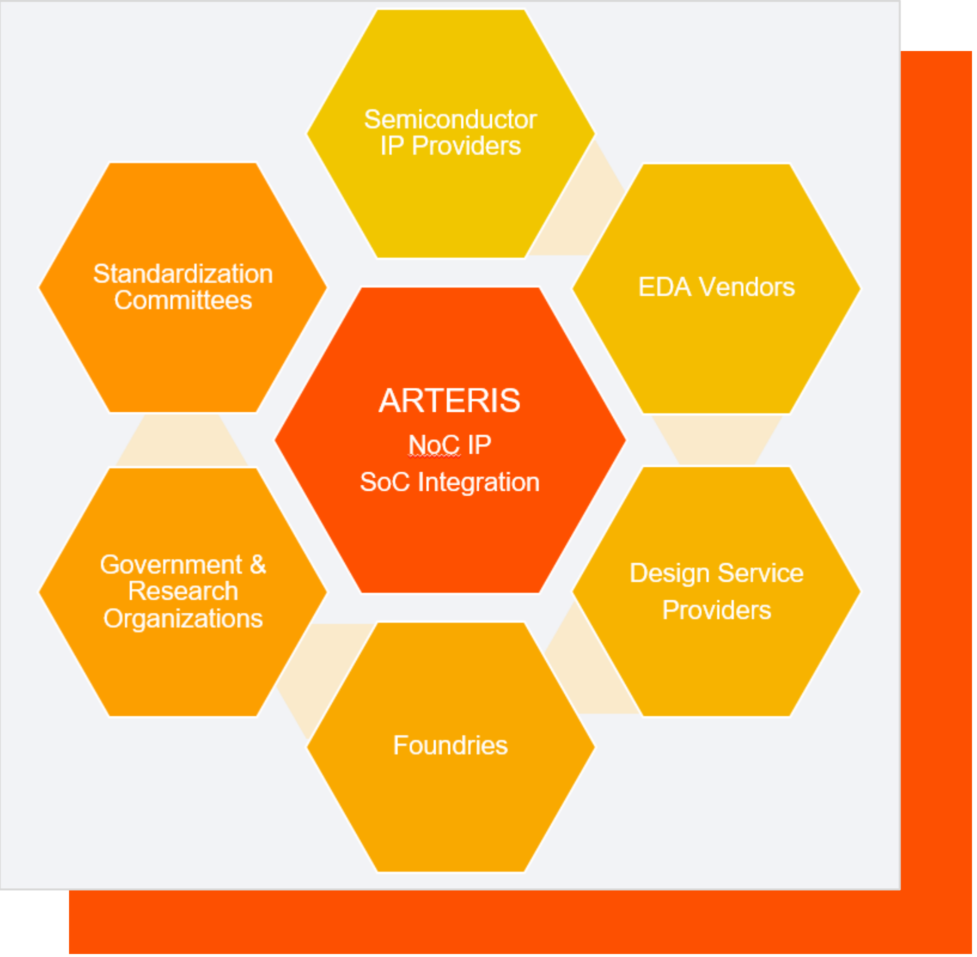


Arteris – Connecting Ecosystem

Arteris technology – vendor-**agnostic**

- Connecting across the entire **semiconductor ecosystem**
- Simplifying & derisking SoC designs




Semiconductor IP Providers
EDA Vendors
Standardization Committees
ARTERIS
NoC IP
SoC Integration
Government & Research Organizations
Design Service Providers
Foundries

Arteris – Connecting Technology

- Arteris technology proliferation
  - Over 850 SoC designs
  - Shipped in >3.7 billion SoCs globally
  - Touching every industry

Automotive

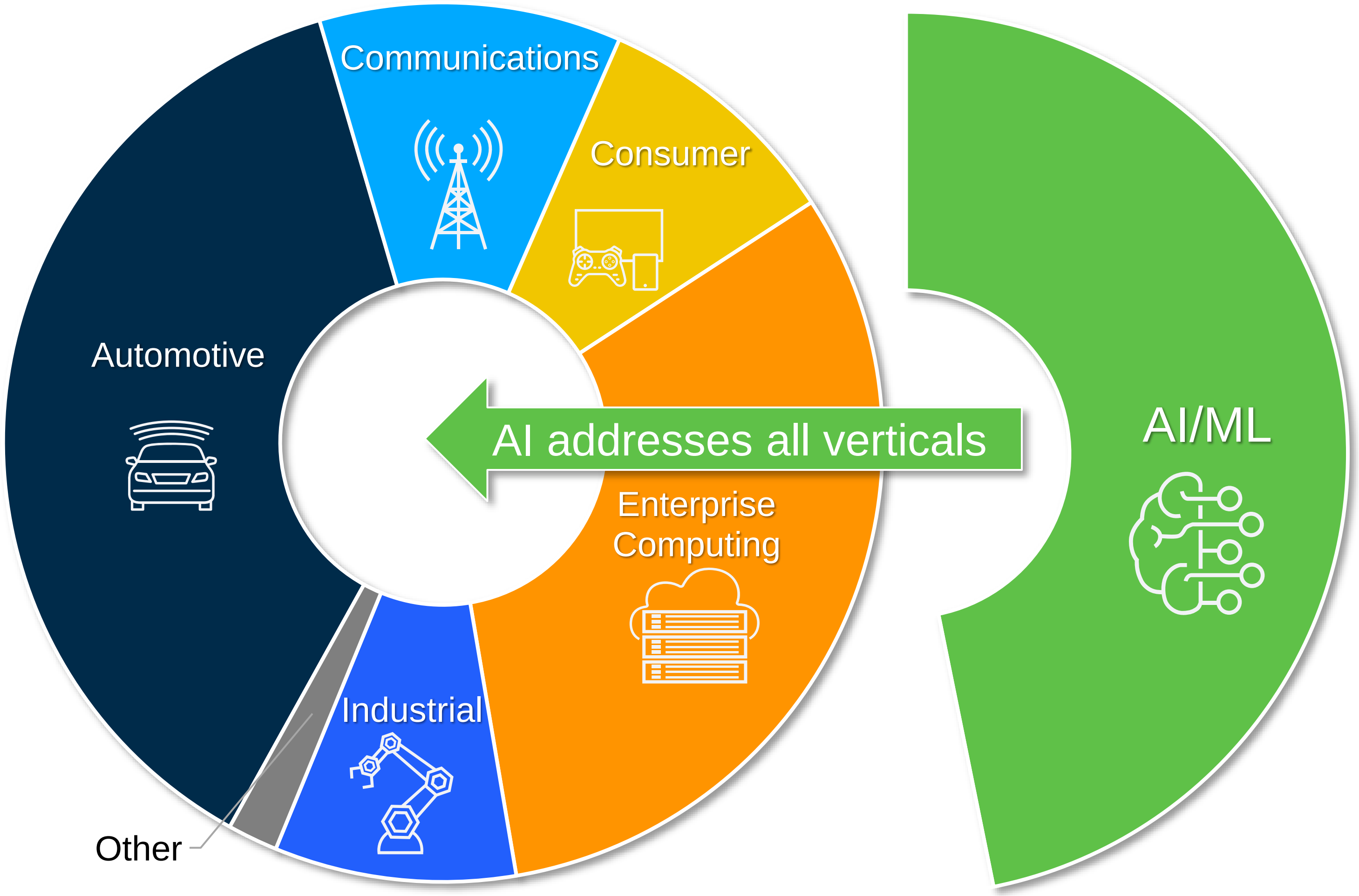
Communications

Consumer Electronics

Enterprise Computing

Industrial

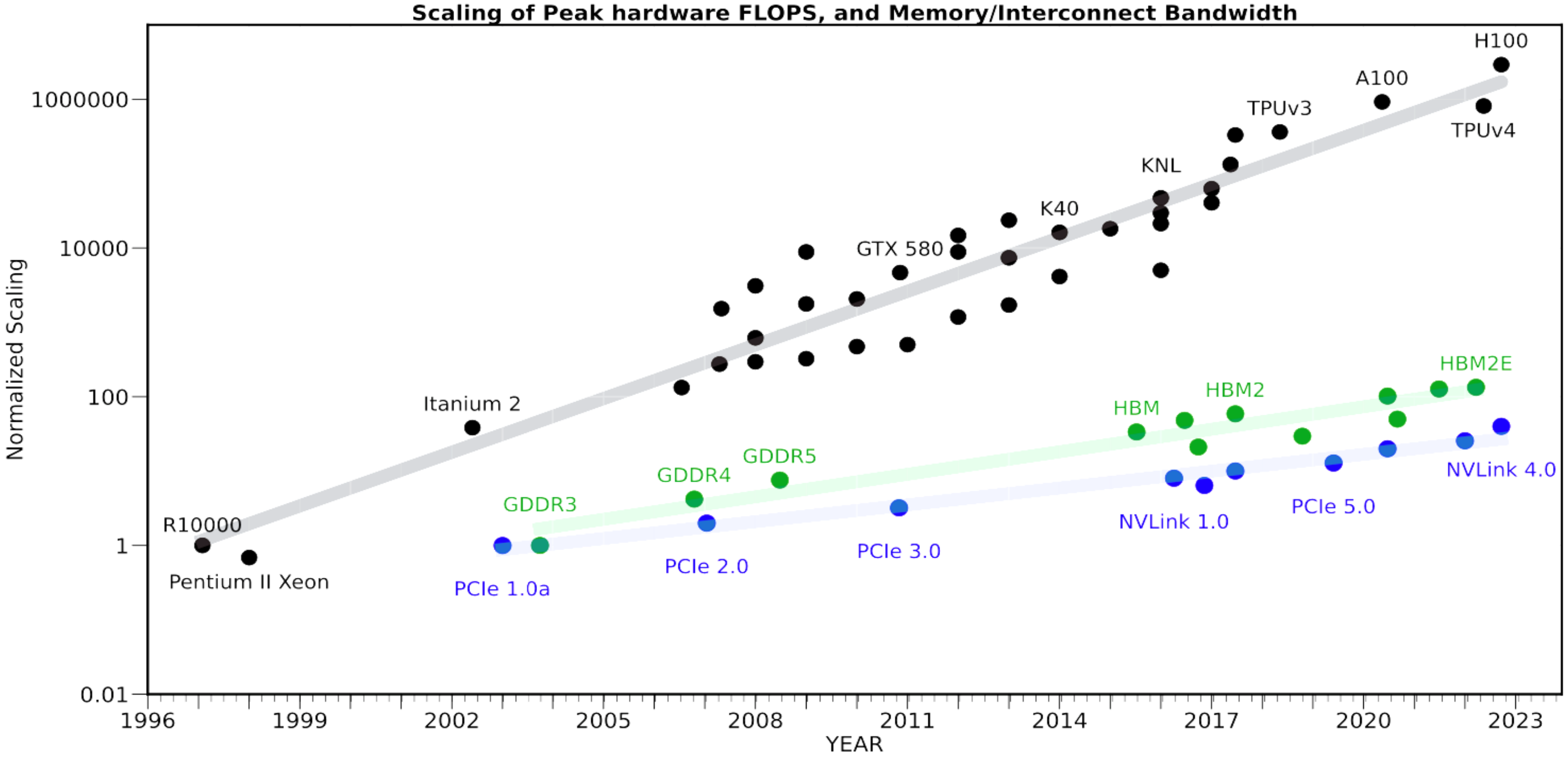
Arteris – Enabling Increasing Numbers of AI/ML Chiplets and SoCs



- Training
- Inference
- Generative AI
- Vehicle endpoints
- Robotics
- Datacenter
- Infrastructure

The Opportunities and Challenges

Interconnect is the enabler for high performance & complex SoCs



Over the past 20 years:

- Peak compute increased: **60,000x**
- DRAM bandwidth increased: **100x**
- Interconnect bandwidth increased: **30x**

Hennessy & Patterson predicted a few years ago, ***‘A new golden age for computer architecture’*** driven by:

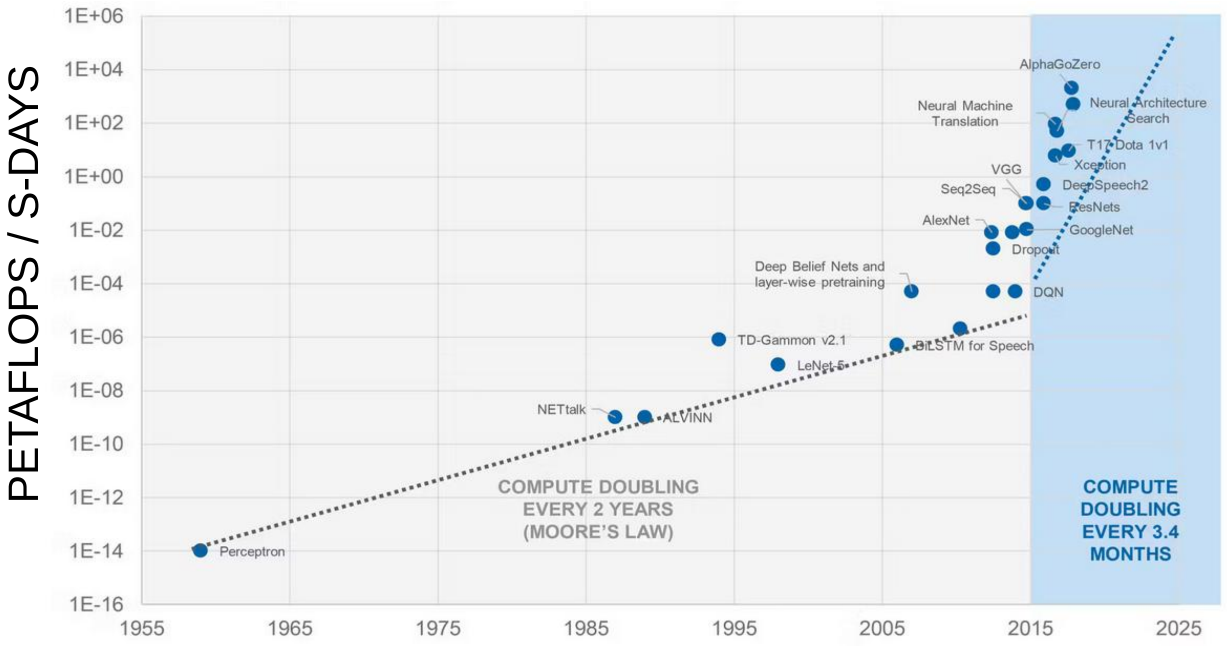
- Open Instruction Sets
- Domain-Specific Architecture
- Agile Chip Development

Source: John L. Hennessy and David A. Patterson. 2019. A new golden age for computer architecture. Commun. ACM 62, 2 (February 2019), 48–60. https://doi.org/10.1145/3282307
A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney and K. Keutzer, "AI and Memory Wall," in IEEE Micro, vol. 44, no. 3, pp. 33-39, May-June 2024, doi: 10.1109/MM.2024.3373763

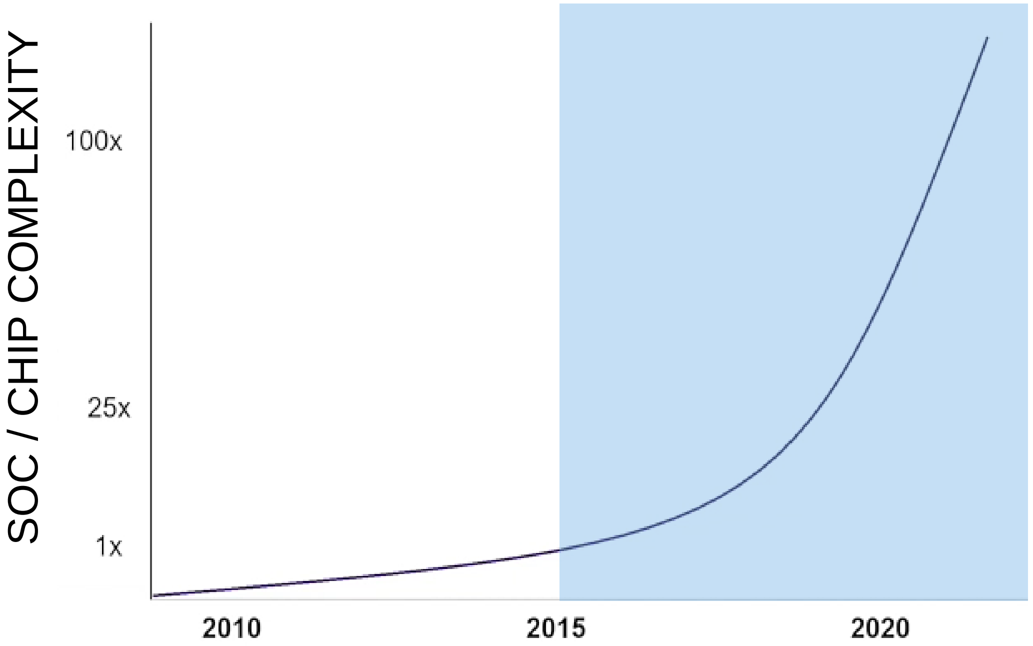
Expanding AI Compute Accelerating the Rise in Chip Complexity

More logic and design constraints further complicating on-chip connectivity (“the glue”)

AI Era of Compute: 7x Acceleration



Driving Exponential Chip Complexity



*Complexity defined by number of elements and corners

Modern SoCs are connected by **billions of wires**, ever-expanding, with growing complexity

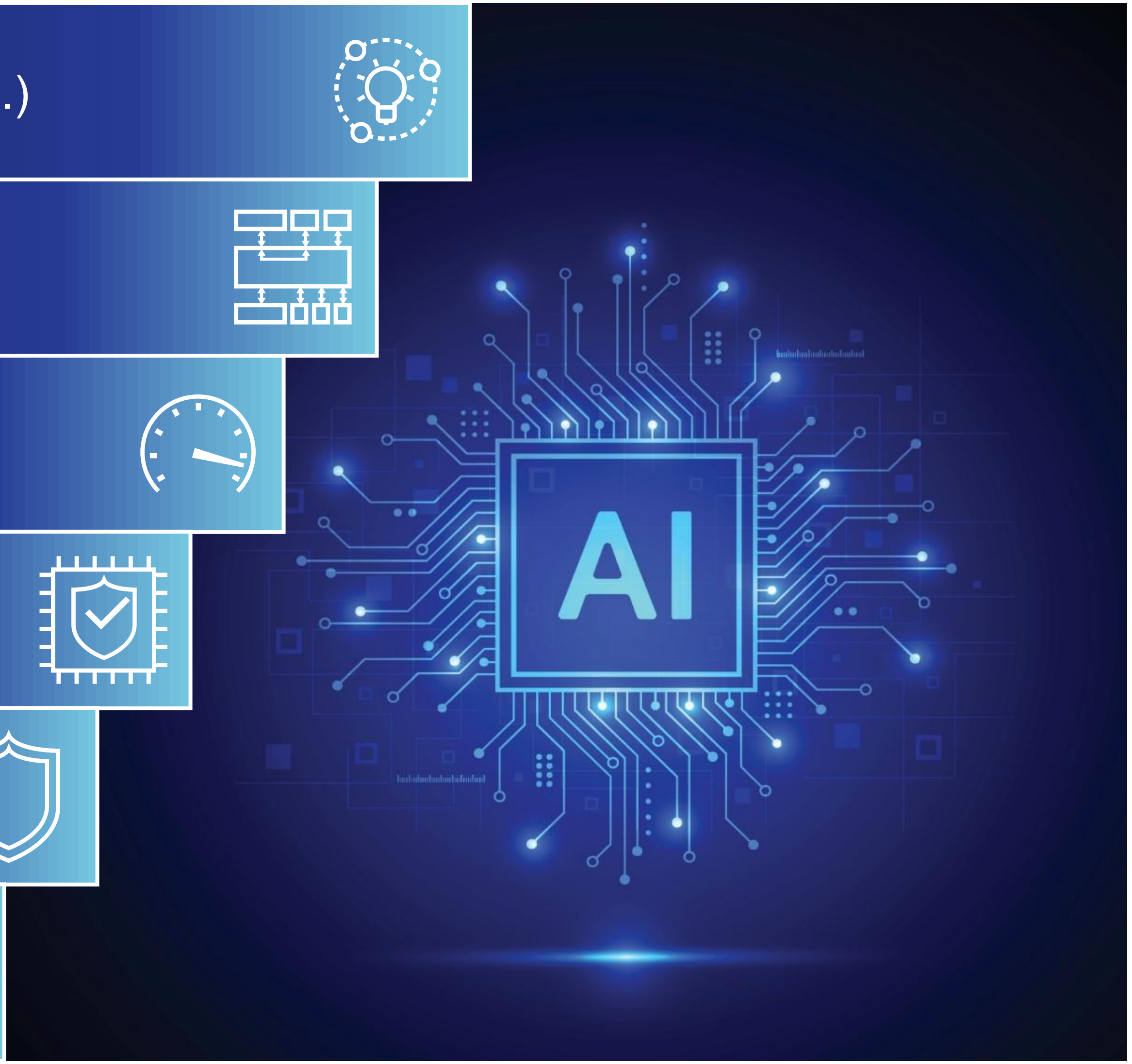
→ More & bigger networks-on-chip (NoCs)

→ Requiring more complex architecture and topology

Sources: OpenAI: AI and Compute Research Report, Semi Engineering, Synopsys

Challenges of RISC-V SoCs – Interconnect Is the Problem to Address

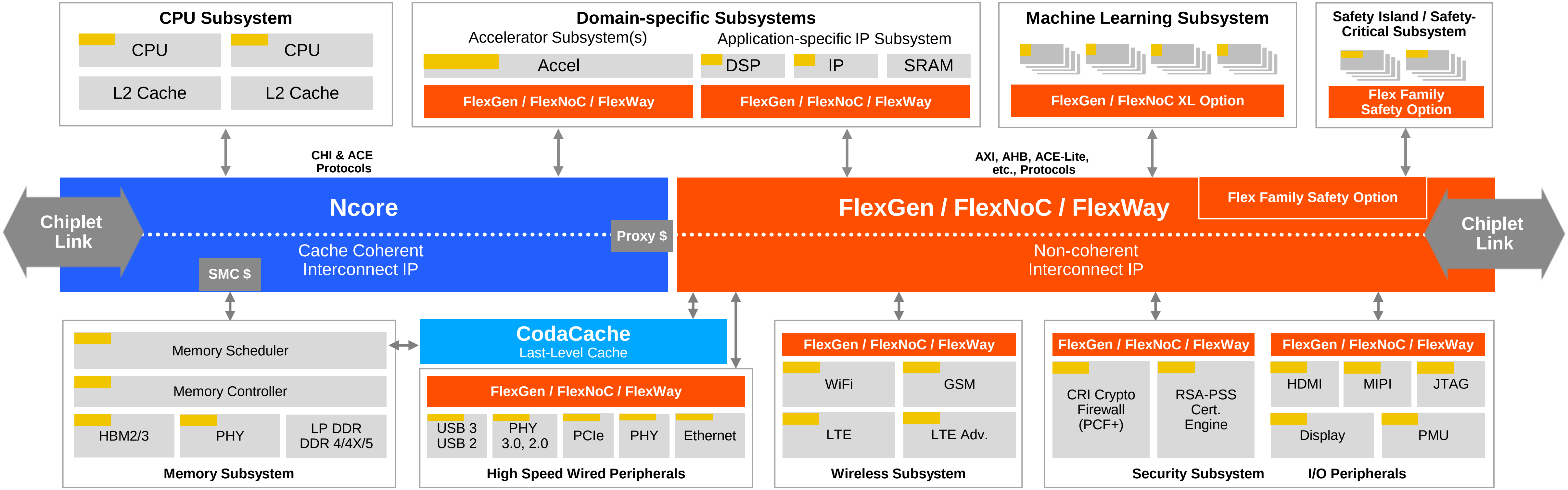
- Diverse interface protocols (CHI, ACE, ACE-Lite, AXI….)
- Various coherency models (MESI, MOESI)
- ‘Memory wall’ - Massive memory bandwidth of AI/ML
- Interoperability / Verification / Performance models
- Functional safety standards for automotive
- Physical Design (PD)



Arteris System IP and Network-on-Chip (NoC) for RISC-V AI SoCs

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



CPU Subsystem: CPU, CPU, L2 Cache, L2 Cache

Domain-specific Subsystems: Accelerator Subsystem(s) – Accel – FlexGen / FlexNoC / FlexWay; Application-specific IP Subsystem – DSP, IP, SRAM – FlexGen / FlexNoC / FlexWay

Machine Learning Subsystem: FlexGen / FlexNoC XL Option

Safety Island / Safety-Critical Subsystem: Flex Family Safety Option

CHI & ACE Protocols

AXI, AHB, ACE-Lite, etc., Protocols

Chiplet Link

Ncore – Cache Coherent Interconnect IP – SMC $ – Proxy $

FlexGen / FlexNoC / FlexWay – Flex Family Safety Option – Non-coherent Interconnect IP

Chiplet Link

Memory Subsystem: Memory Scheduler, Memory Controller, HBM2/3, PHY, LP DDR DDR 4/4X/5

CodaCache – Last-Level Cache

High Speed Wired Peripherals: FlexGen / FlexNoC / FlexWay – USB 3 USB 2, PHY 3.0, 2.0, PCIe, PHY, Ethernet

Wireless Subsystem: FlexGen / FlexNoC / FlexWay – WiFi, GSM, LTE, LTE Adv.

Security Subsystem: FlexGen / FlexNoC / FlexWay – CRI Crypto Firewall (PCF+), RSA-PSS Cert. Engine

I/O Peripherals: FlexGen / FlexNoC / FlexWay – HDMI, MIPI, JTAG, Display, PMU

🟦 Arteris Ncore™
Cache coherent interconnect IP

🟧 Arteris FlexGen™, FlexNoC®, FlexWay™
Non-coherent interconnect IP

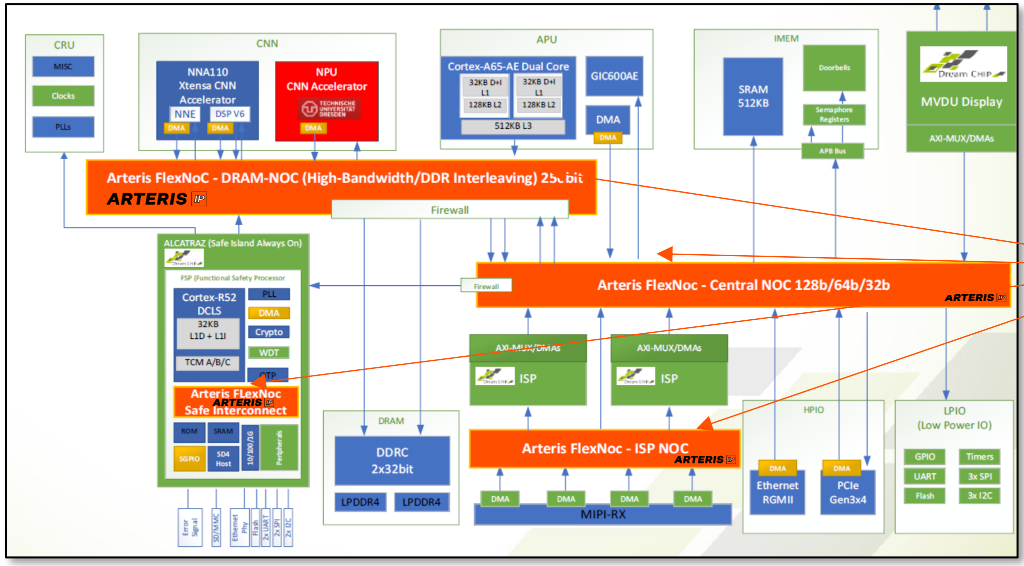
🟦 Arteris CodaCache®
Last-level cache IP

🟨 Arteris Magillem® Registers
Hardware/Software interface

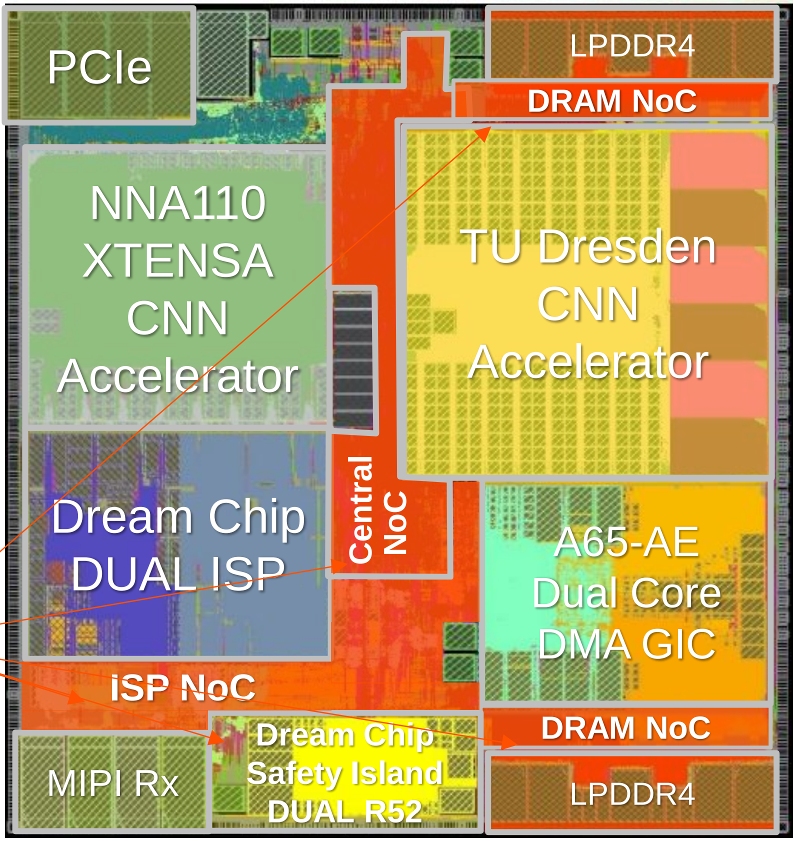
Mixed Architecture AI SoCs are Commonplace

DreamChip example of ADAS Level 2+ SoC with 2x CNN accelerator sections

- Custom AI Accelerator – 6-24 TOPS
- 2x AI accelerators – 10 TOPS
  - **RISC-V** CPU in NPU with 768 Processing Elements (PE)
  - **Cadence® Tensilica®** AI Max coupled with Vision DSP V6
- Dual-core **Arm® Cortex®-A65AE** processor cluster
- **Arm Cortex-R52** functional safety processor
- 2x DreamChip Image Signal Processors (ISPs)
- ISO26262 safety certified up to ASIL-D (TÜV SÜD)



Multiple Arteris NoC IP Instances

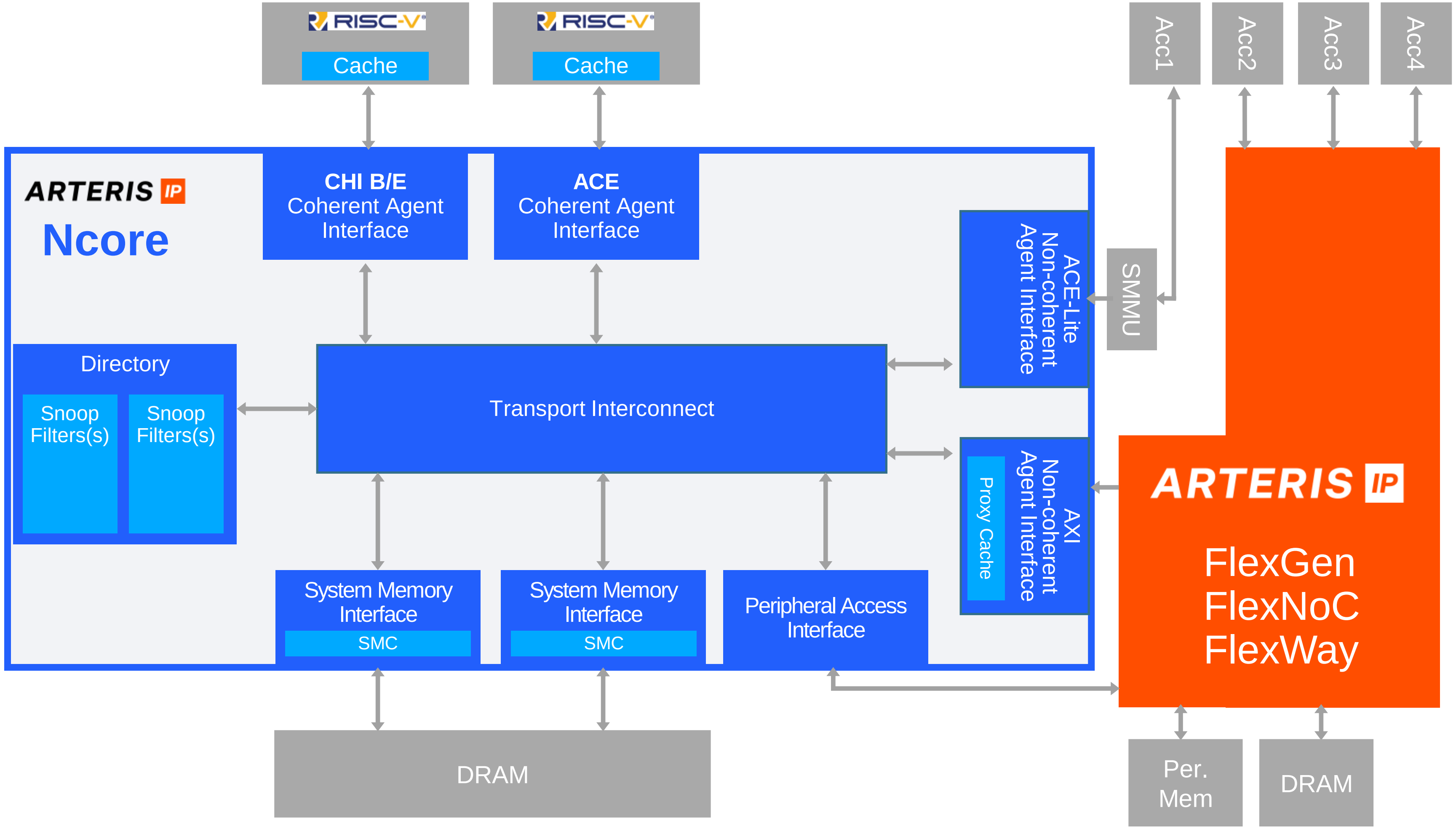


NoCs: ~15% of SoC

Sources: DreamChip, https://www.eenewseurope.com/en/china-owned-dream-chip-tapes-out-10tops-soc/ , Case Study

Why Do Designs Use AMBA CHI and ACE in the Same System?

Adapt to RISC-V dynamic ecosystem

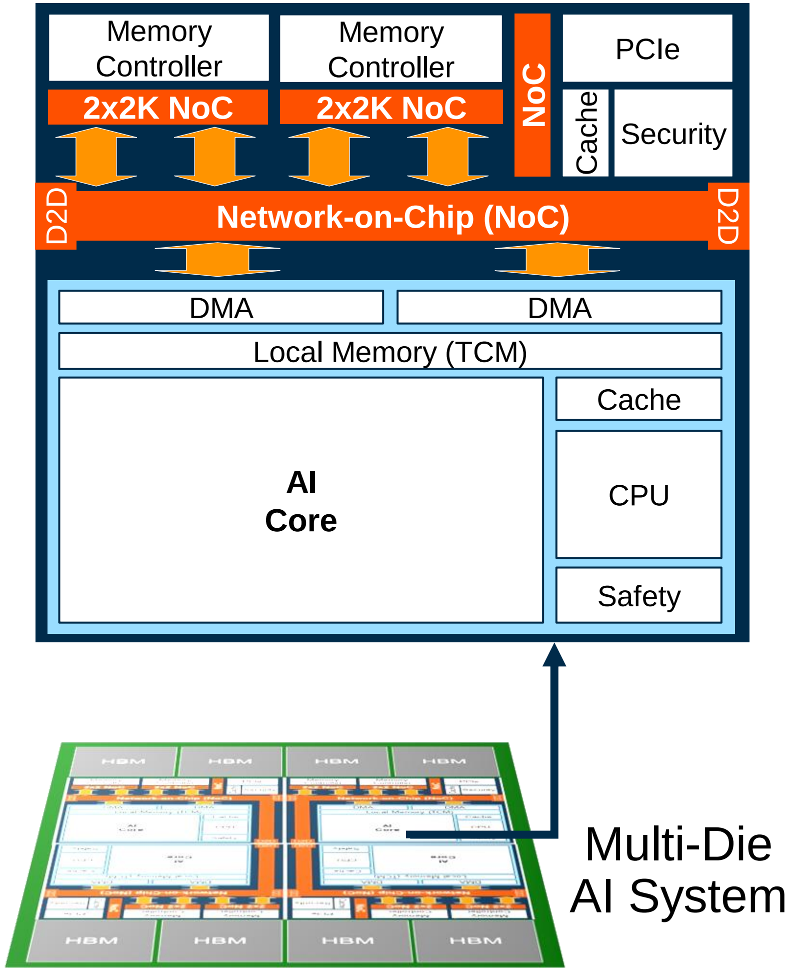


- **RISC-V is a diverse and evolving ecosystem**
- **Mixed ACE/CHI** can **ease integration** of new and legacy processors
  - Coherent Interfaces: CHI-B, CHI-E or ACE, **interoperable**
  - **Mix the latest** high-performance RISC-V clusters using CHI with **older** RISC-V CPUs using ACE
  - **Mix of RISC-V and Arm**
  - Leverage investment in ACE IP
- Proxy caches ease integration of non-coherent accelerators into the coherent domain
- Provide PCIe connectivity for storage and data center applications

Generative AI and LLM Data Transport Architecture

Ultra-wide Data to Support the Rapidly Growing Number of Parameters

- A tremendous amount of data has to be moved in Gen AI chips
- Maximizing Throughput to Handle Trillion+ Parameter AI
- Reduce data movement by bringing compute and memory
- Performance Scaling for Multi-TFLOP/GHz
- Support of Arm, RISC-V, and mixed architectures
- Support of Multi-Die / Chiplet architectures



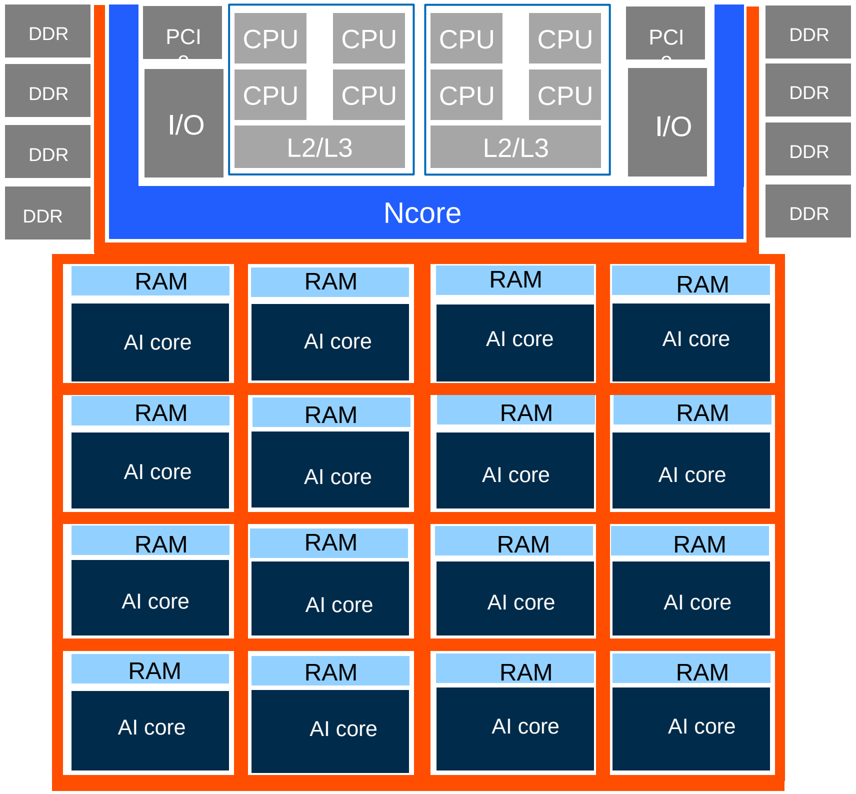
Efficient and Performant **AI/ML Data Transport Architecture**

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

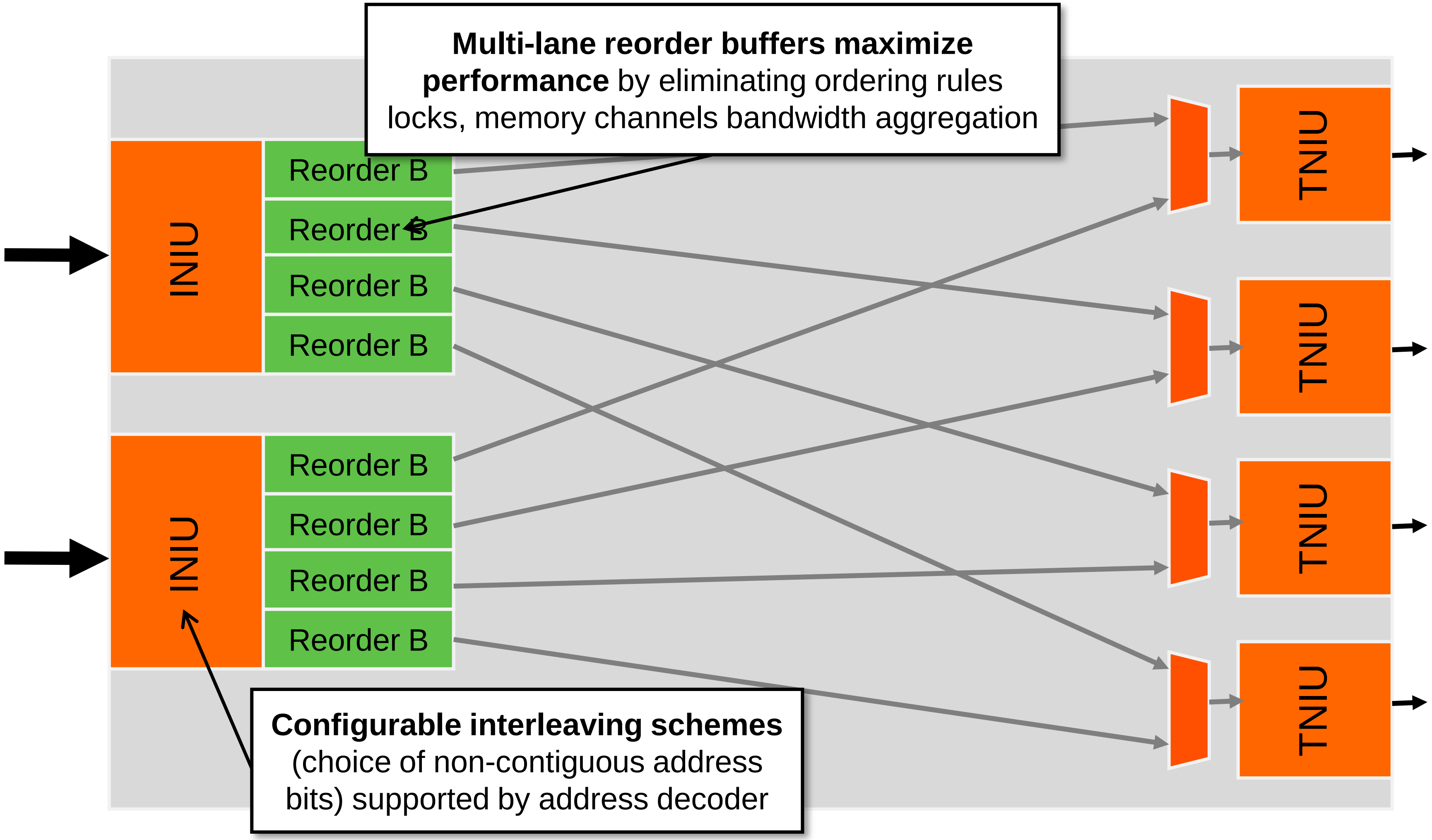
FlexGen / FlexNoC 5 connect the AI core accelerator units



AI/ML Accelerator SoC

High Memory Bandwidth from Interleaving Channels

- Up to 8 or 16 channels interleave
- Read-reorder buffers
- Traffic aggregation / data width conversions
- Up to 2048 bits wide connections



Multi-lane reorder buffers maximize performance by eliminating ordering rules locks, memory channels bandwidth aggregation

Configurable interleaving schemes (choice of non-contiguous address bits) supported by address decoder

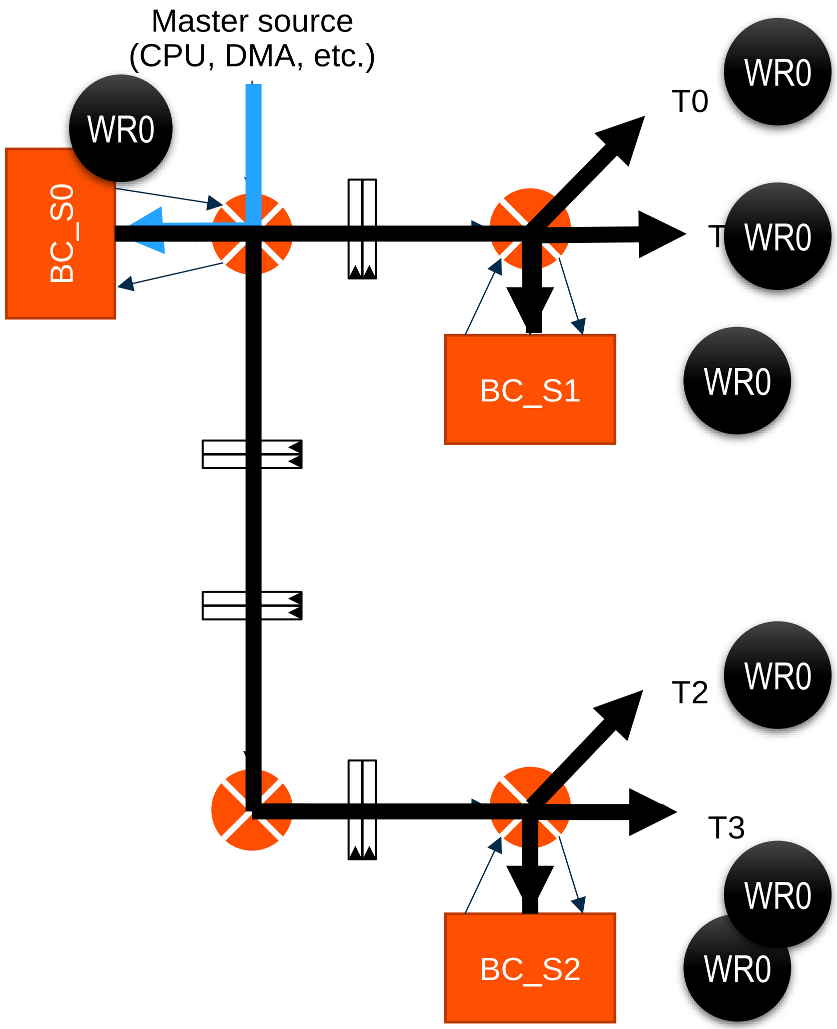
INIU: Initiator Network Interface Unit
TNIU: Target Network Interface Unit

Intelligent Multicast Write

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

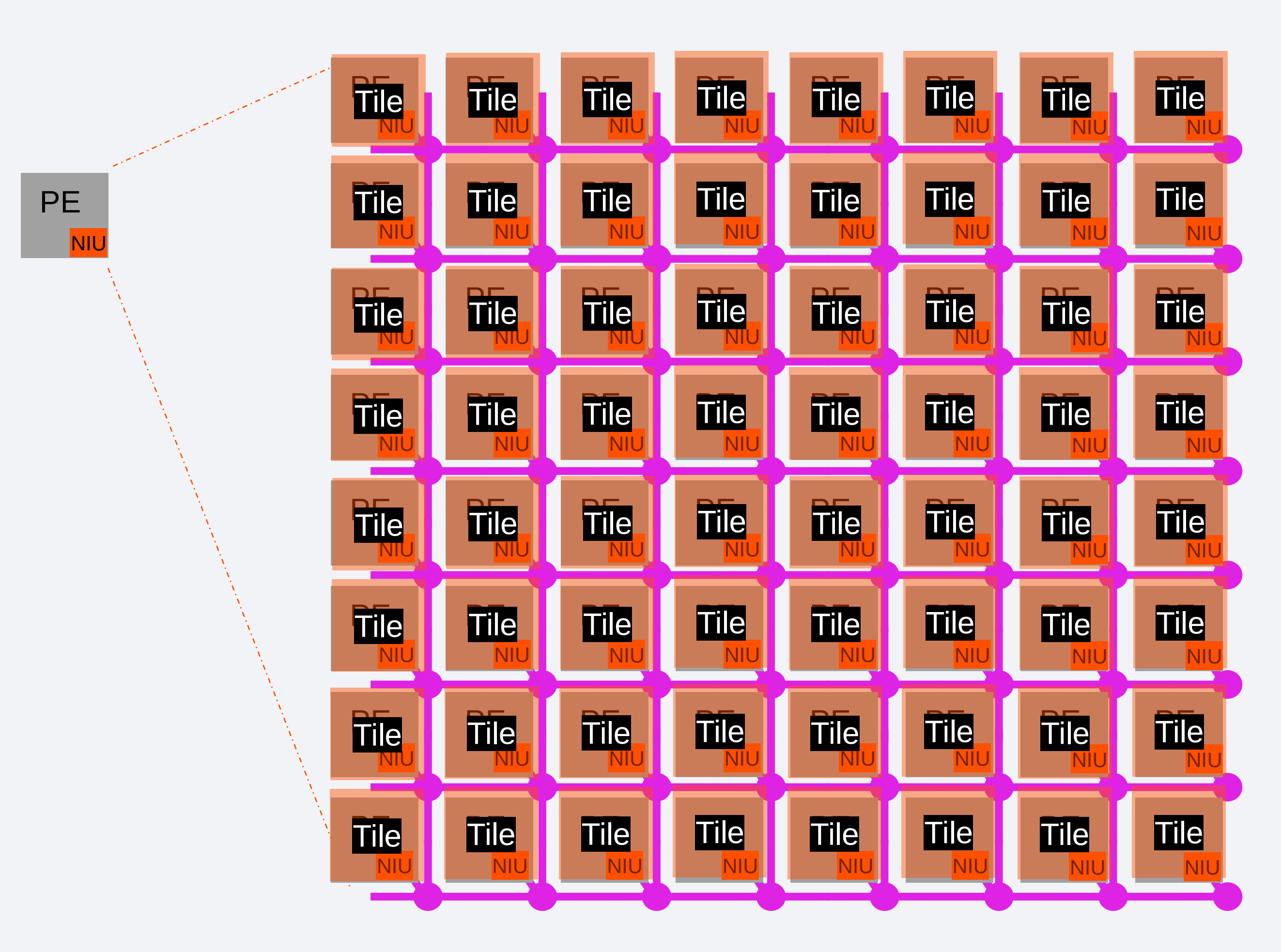
- **Used in AI for Deep Neural Network (DNN) weight and image map updates**



AI Tiling: >10x Scalable Performance with Mesh Connected Tiles

Meeting AI’s massive demand for faster and more powerful computing

- NoC tiling allows **AI chips to boost their processing power** by more than **10x without changing the basic design**
- **The effort** to implement the NIUs, the most logically intense elements in the NoC, **is drastically reduced**
- **NIUs can be implemented once, then tiled** using external tie-offs for IDs
- **AI Workloads:** Vision, ML, DL, NLP including LLMs and Generative AI



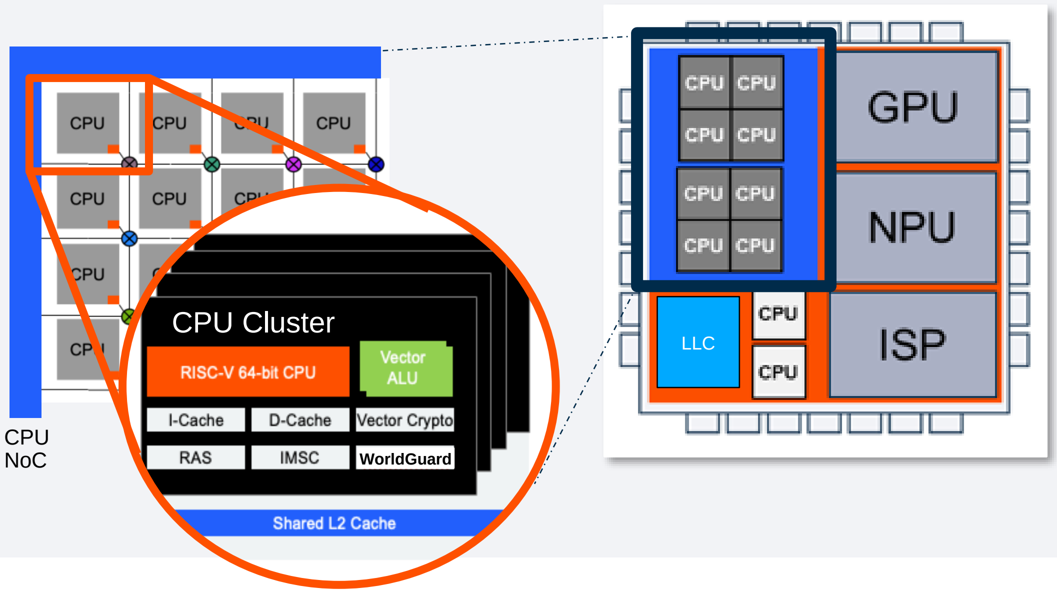
Arteris estimations based on NoC tiling customer use cases for AI/ML

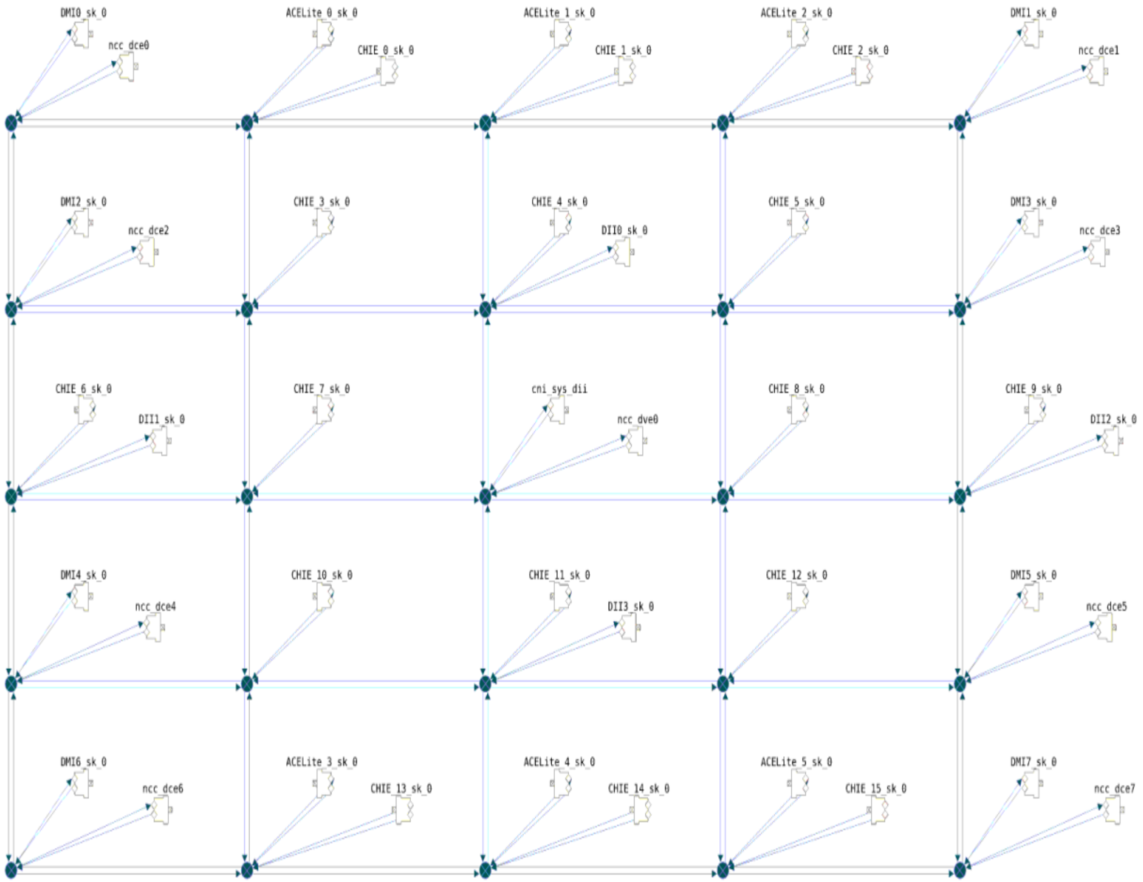
Large Compute Support with CPU Tiling and Mesh Topology

Cache-coherent Ncore IP with flexible and highly scalable support up to 512 CPUs in clusters

CPU Tiling with Mesh Topology Example

Coherent mesh NoC with tiled CPU clusters, each containing up to 32 CPUs. A 5x5 mesh configuration allows 16 CPU clusters access to maximum memory bandwidth. The remaining mesh sockets are used for caches and service networks.



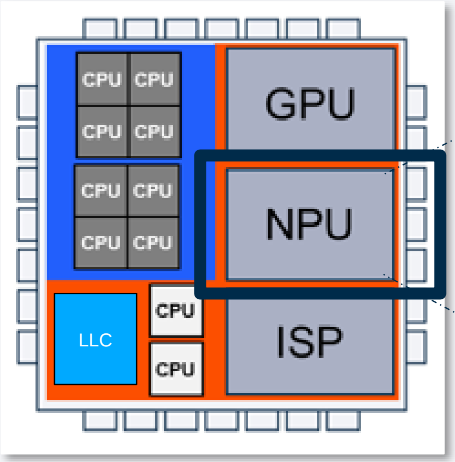


NoC Tiling with Mesh Topology for NPUs, GPUs, TPUs

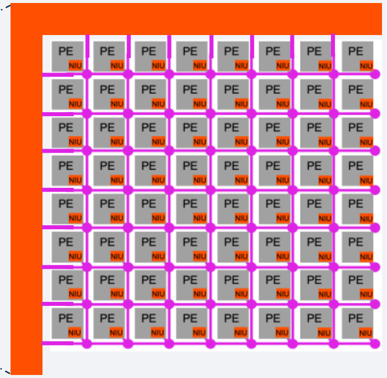
Flexible, highly scalable tiling supported by mesh, up to 1024 tiles, in FlexNoC IP

NPU Tiling with Mesh Topology Example

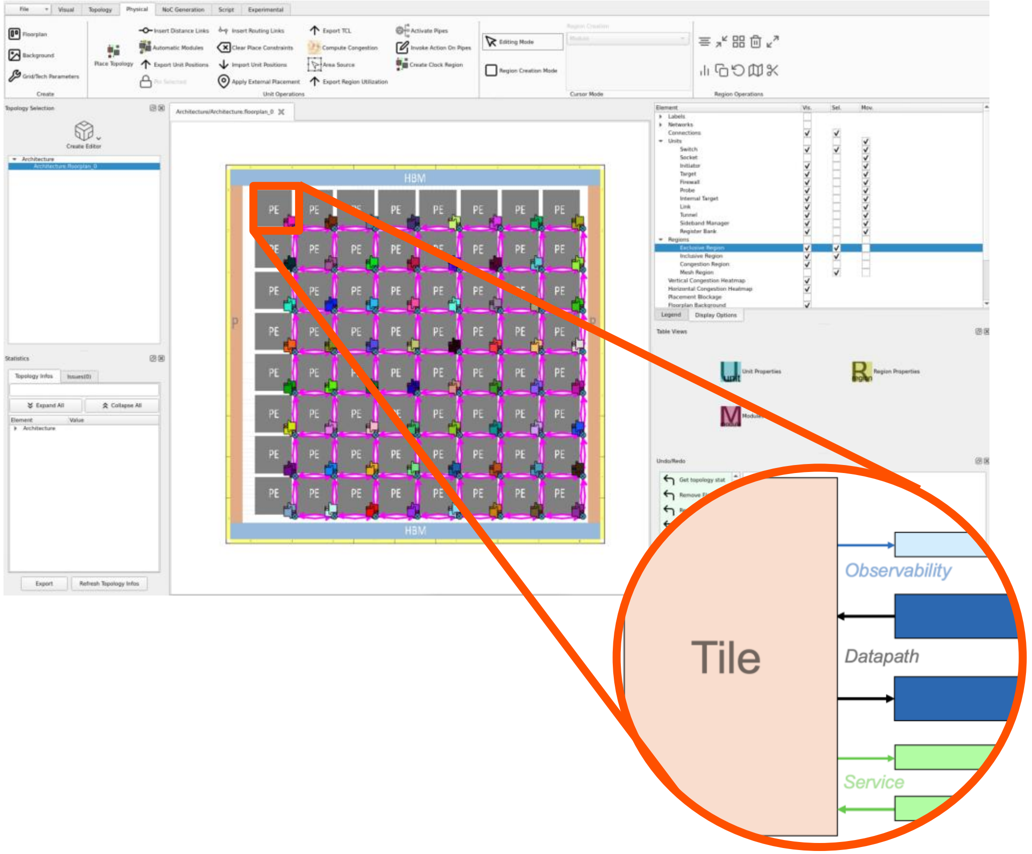
NoC tiling divides the design into modular, repeatable units called “tiles”, allowing for significant scalability, power efficiency, reduced latency, and faster development without redesigning the entire NoC architecture.



NPU on the same die and made from tiles and each PE comprises a NoC network interface unit (NIU)

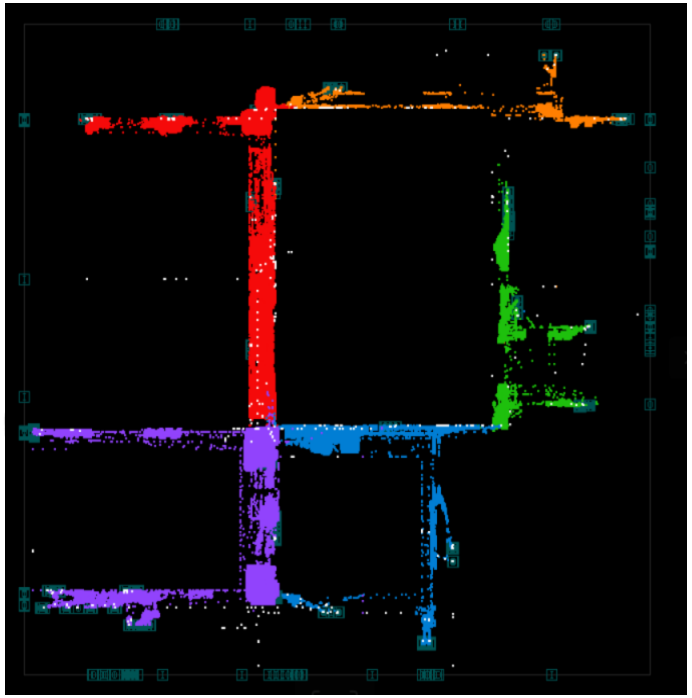


Main NoC



Improving TAT and PPA for Complex AI Chips and Chiplets
Sample FlexGen Smart NoC IP Results for Automotive AI SoC (ADAS)

FlexNoC (Manual)



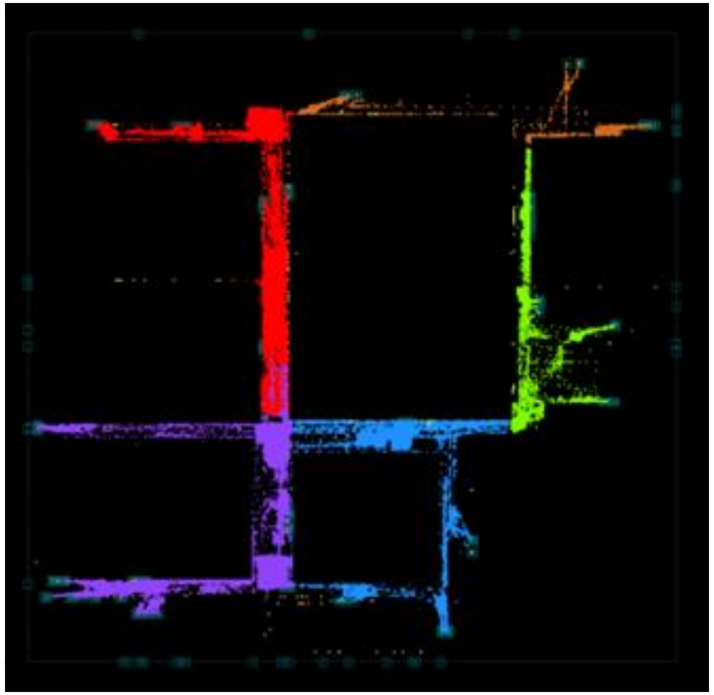
Total Wire Length	138,709 mm
Length of Longest Wire	904 mm
Number of Switches	258
Number of Links	313
No. of Clock Adapters	152
No. of Packet Adapters	157
Latency	65.18 ns
Maximum Latency	1005.67 ns
Main NoC area	3.64 mm^2

10x productivity
-26% wire length
-28% longest wire

-5% latency
-51% max latency
-3% area

DREAM CHIP

FlexGen (Automated)



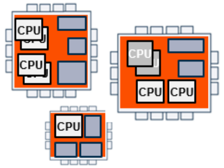
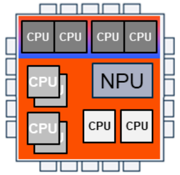
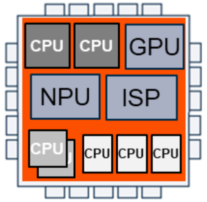
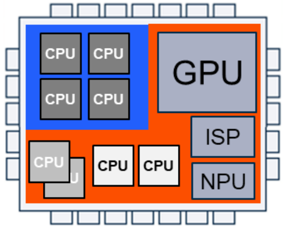
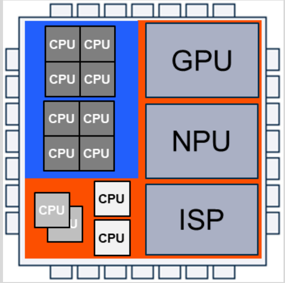
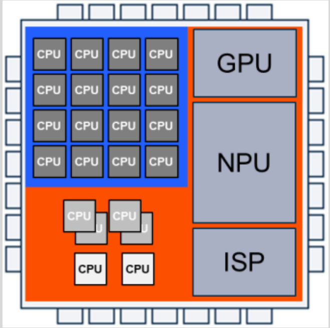
Total Wire Length	102,587 mm
Length of Longest Wire	650 mm
Number of Switches	282
Number of Links	420
No. of Clock Adapters	141
No. of Packet Adapters	210
Latency	62.08 ns
Maximum Latency	491.67 ns
Main NoC area	3.51 mm^2

Source: Dream Chip Technologies

TAT: Turn-Around Time
PPA: Power, Performance, Area

Automotive Domains and Their Complexity

Cache coherency is required in safety-critical systems

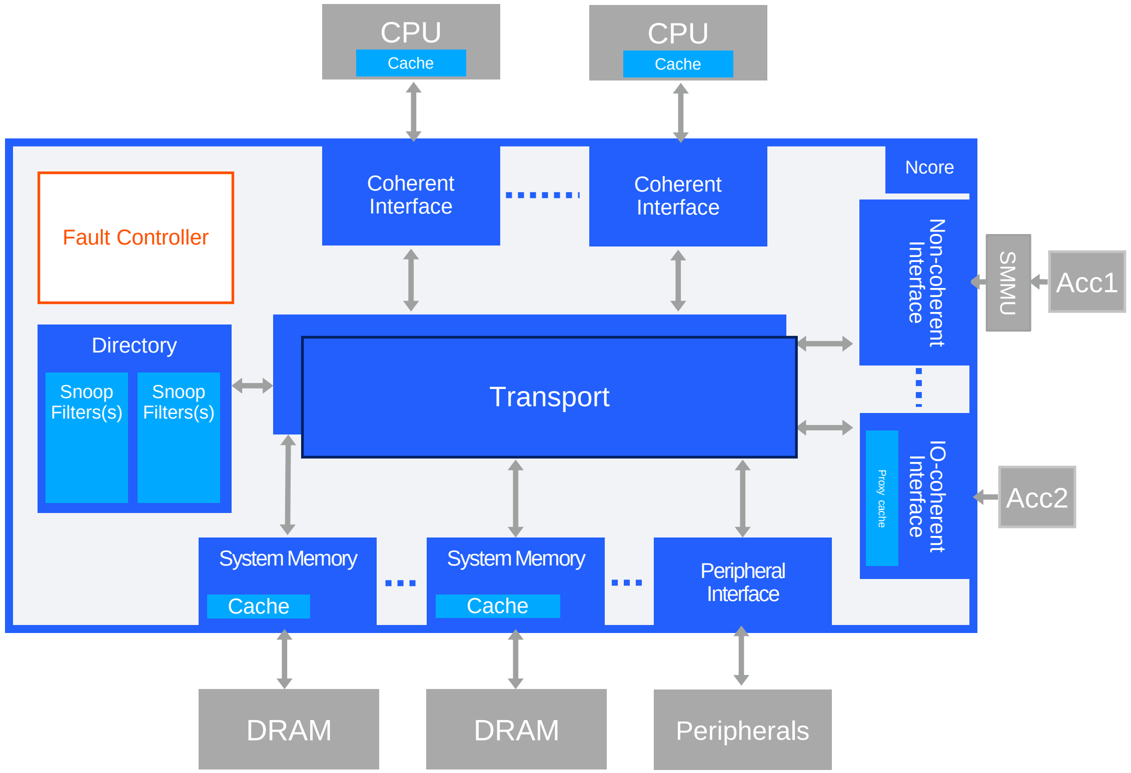
Auto MCU	Zonal Controller	Vision	Cockpit	L2+ ADAS	Autonomous
Non-coherent	Coherent or non-coherent	Mostly non-coherent	Mix of coherent and non-coherent	Mix of coherent and non-coherent, some with large coherent meshes	Mix of coherent and non-coherent, large coherent meshes
					
Monolithic	Monolithic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

Chiplets increasingly important in future

CPU Application   CPU Real Time   CPU Microcontroller   Cache coherent NoC   Non-coherent NoCs

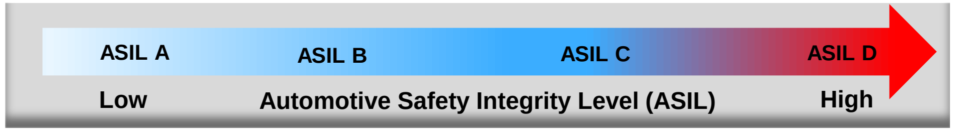
Challenge of **Safety-certification** for Coherent Systems

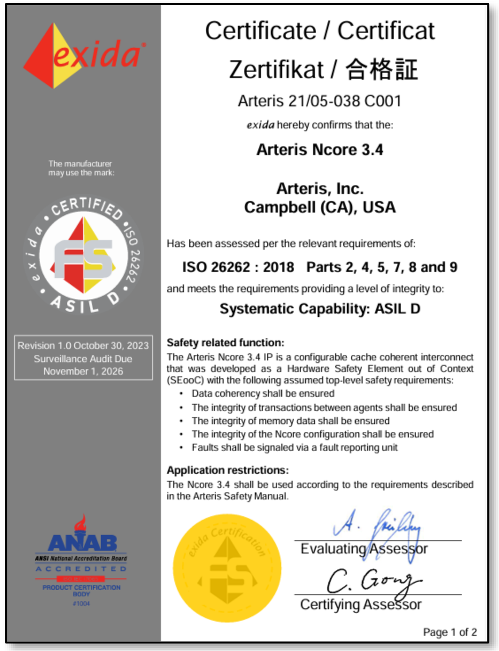
Automotive ADAS/autonomous driving is a key application of AI/ML



Ncore 3.4 is ISO 26262 ASIL D certified

- 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





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 (SEooC) 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.

The manufacturer may use the mark:

CERTIFIED • ISO 26262 • ASIL D • exida •

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

ANAB
ANSI National Accreditation Board
ACCREDITED
ISO/IEC 17065
PRODUCT CERTIFICATION BODY
#1004

exida Certification

A. [illegible]
Evaluating Assessor

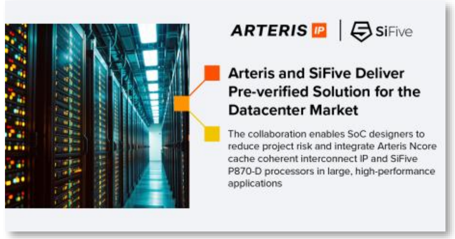
C. Gong
Certifying Assessor

Page 1 of 2

RISC-V Ecosystem Collaboration and Interoperability

Broad and Expanding Ecosystem

SiFive

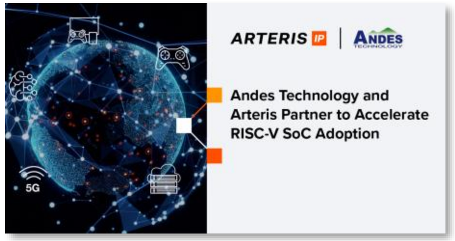


ARTERIS IP | SiFive

Arteris and SiFive Deliver Pre-verified Solution for the Datacenter Market

The collaboration enables SoC designers to reduce project risk and integrate Arteris Ncore cache coherent interconnect IP and SiFive P870-D processors in large, high-performance applications

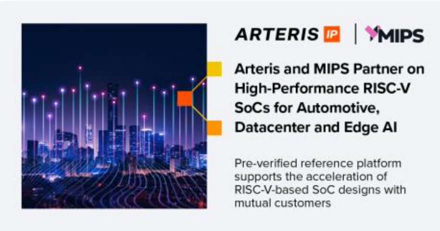
ANDES TECHNOLOGY



ARTERIS IP | ANDES

Andes Technology and Arteris Partner to Accelerate RISC-V SoC Adoption

MIPS

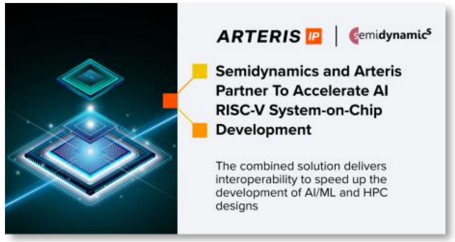


ARTERIS IP | MIPS

Arteris and MIPS Partner on High-Performance RISC-V SoCs for Automotive, Datacenter and Edge AI

Pre-verified reference platform supports the acceleration of RISC-V-based SoC designs with mutual customers

semidynamics

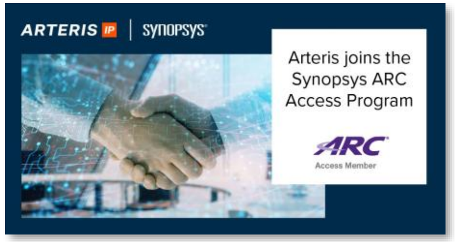


ARTERIS IP | semidynamics

Semidynamics and Arteris Partner To Accelerate AI RISC-V System-on-Chip Development

The combined solution delivers interoperability to speed up the development of AI/ML and HPC designs

SYNOPSYS



ARTERIS IP | SYNOPSYS

Arteris joins the Synopsys ARC Access Program

ARC Access Member

達摩院 DAMO ACADEMY | XUANTIE

. . .

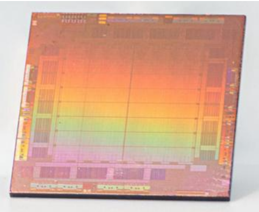
Silicon-Proven Examples

mobileye + MIPS

Mobileye EyeQ ULTRA
with MIPS eVocore P8700
L4 Autonomous Vehicle

Meta + ANDES TECHNOLOGY

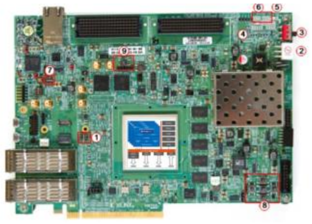
Meta MTIA v2
with Andes Technology QiLai
Large-Scale AI/ML

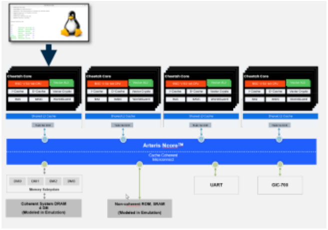


. . .

Tested Interoperability

Test Boards, Simulation, and Emulation

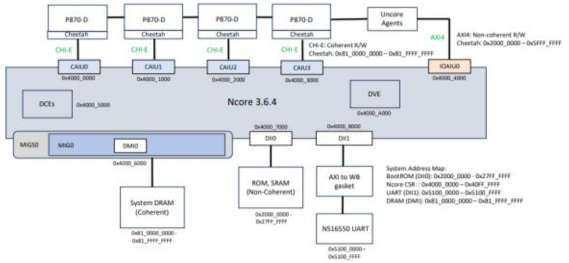




Cheetah Core
Shared L2 Cache
Arteris Ncore™
Cache Coherent Interconnect
Memory Subsystem
Coherent System DRAM 4 GB (Modeled in Emulation)
Non-coherent ROM, SRAM (Modeled in Emulation)
UART
GIC-700

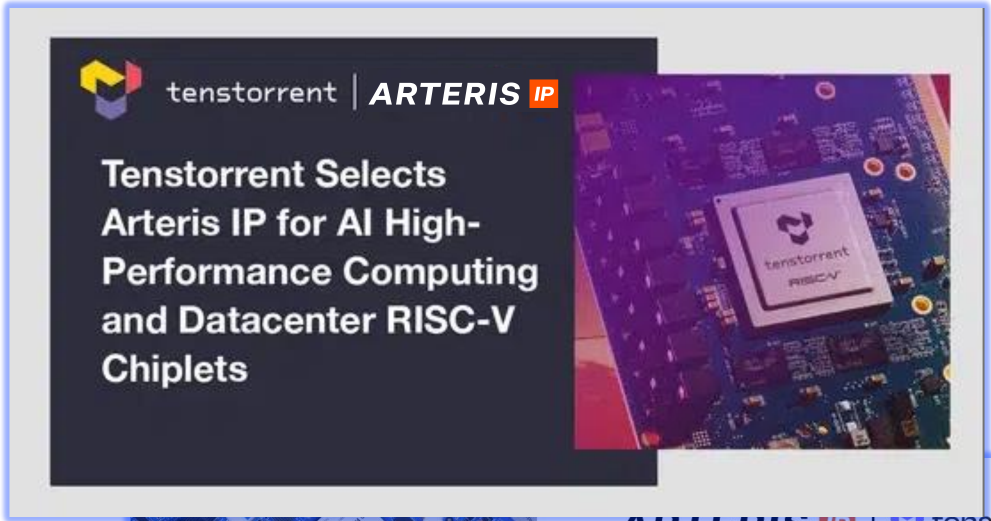
System Configurability

Software (Magillem) driven IP
Configurability of RISC-V CPUs and NoCs



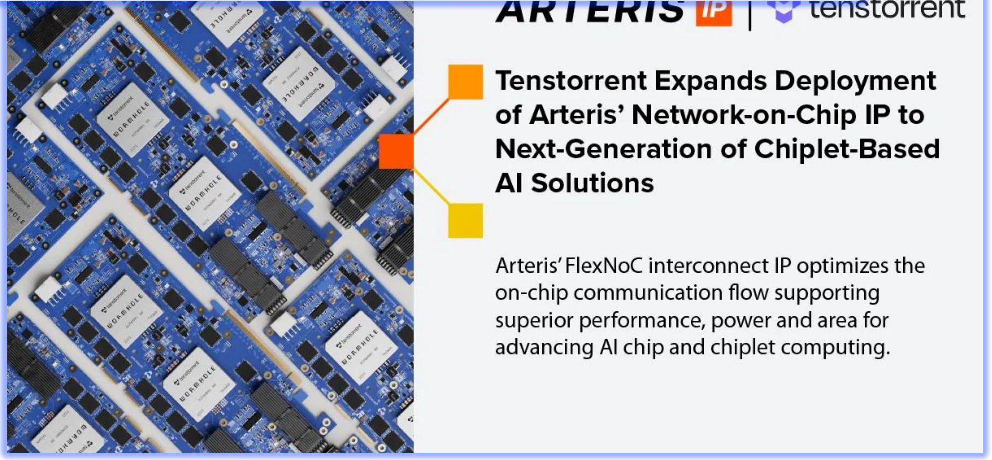
Sources: SiFive, Andes, MIPS, Semidynamics, Synopsys, Damo Academy, Hot Chips, SMYG Limited

Tenstorrent – RISC-V AI High Performance Computing



tenstorrent | ARTERIS IP

Tenstorrent Selects Arteris IP for AI High-Performance Computing and Datacenter RISC-V Chiplets



ARTERIS IP | tenstorrent

Tenstorrent Expands Deployment of Arteris' Network-on-Chip IP to Next-Generation of Chiplet-Based AI Solutions

Arteris' FlexNoC interconnect IP optimizes the on-chip communication flow supporting superior performance, power and area for advancing AI chip and chiplet computing.

> “We are happy to share that we are partnering with Arteris to use Ncore and FlexNoC IP in our next-generation product, The combination of performance and features made it a great choice for both our AI chips and our high-performance RISC-V CPUs. The Arteris team and IP solved our on-chip network problems so we can focus on building our next-generation AI and RISC-V CPU products.”

tenstorrent

Jim Keller, CEO of Tenstorrent

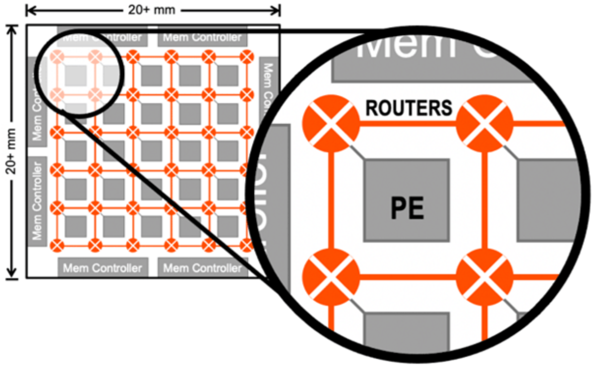
> “We continue to leverage Arteris' network-on-chip IP products in our designs as we drive the next wave of advancements in AI computing. Arteris is a proven technology partner — their FlexNoC IP provides superior performance for our next-generation AI compute.

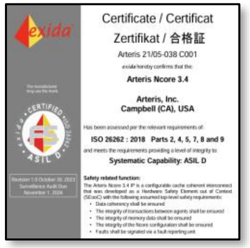
tenstorrent

David Bennett, CCO of Tenstorrent

Summary: RISC-V for AI/ML Is Here, and Growing in Adoption

- **The proliferation of AI/ML hardware is increasing, including with RISC-V Compute**
  - Strong traction for Edge AI Inference, ADAS, Accelerators, and increasingly for training compute
  - Growing deployment of AI chiplets, per modularity, scaling of systems, and cost reductions
- **Expanded bandwidth need for rising AI compute needs an optimized SoC dataflow**
  - Wide buses for massive AI bandwidths
  - Mesh topology for large regular structures that align with physical layout
  - Broadcast writes bandwidth savings in deep neural networks
- **Flexibility, configurability, and smart automation is key to scaling**
- **Mission-critical applications require certifications such as ISO 26262 up to ASIL D**
- **Protocol and ecosystem interoperability is key to pragmatic RISC-V use for AI,…**
  - Interoperable support of mix of standards: CHI-E, CHI-B, ACE, ACE-Lite, AXI, UCIe, …
  - Interoperability testing and silicon proof points




exida

The manufacturer may use the mark

CERTIFIED · exida · ASIL D

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

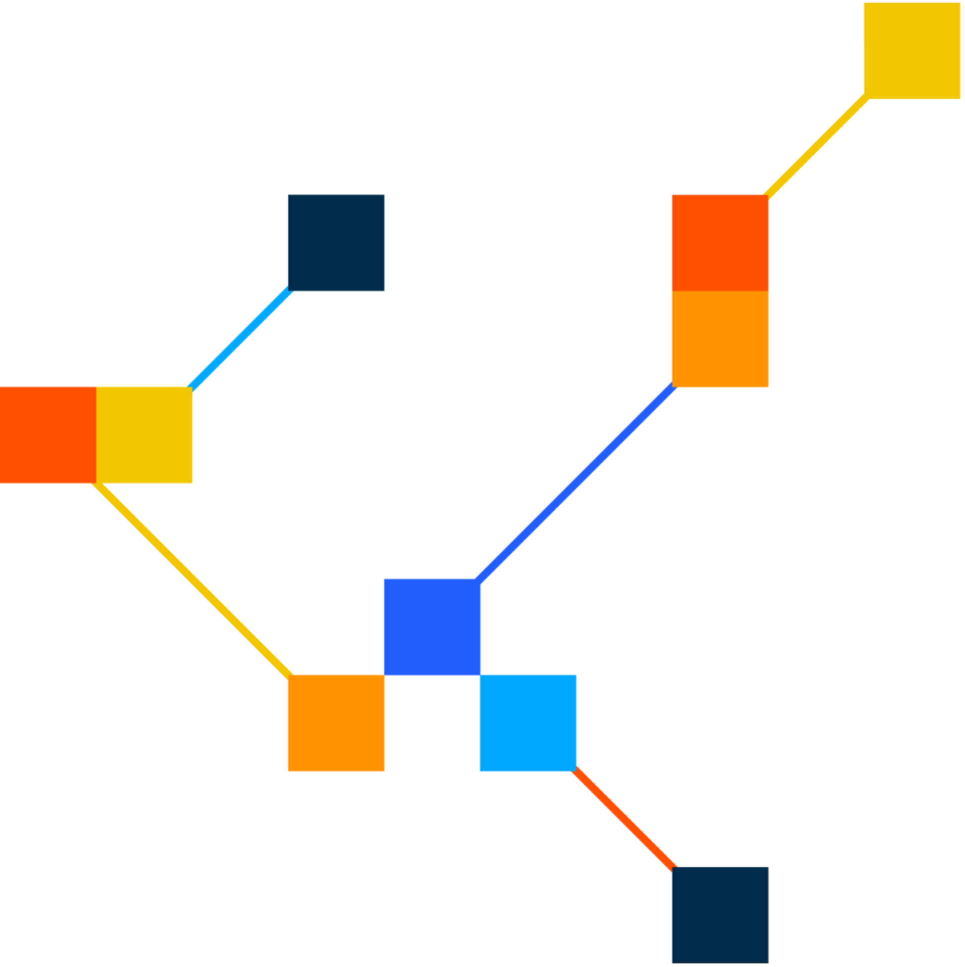
Has been assessed per the relevant requirements of:

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 (SEooC) 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



ARTERIS IP

Thank you