Introduction to AMD Embedded Heterogeneous Design and Versal Architecture

2024.1

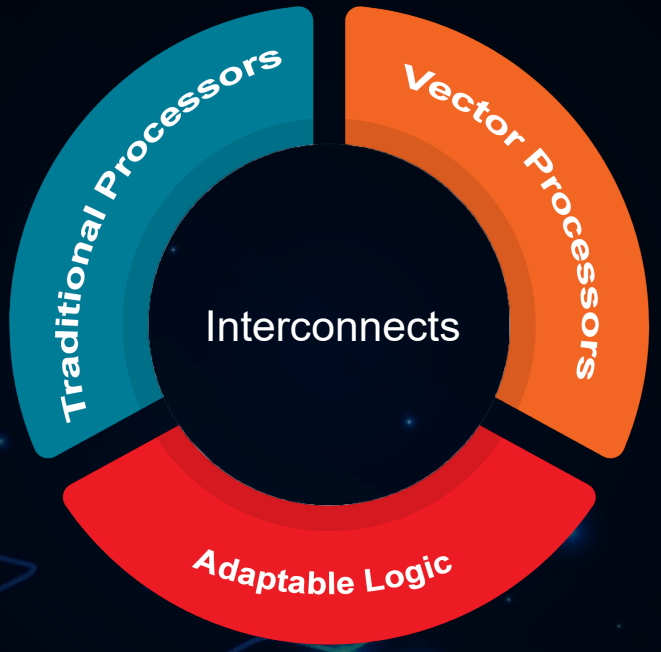
AMD

together we advance_

Defining the Embedded Heterogeneous Environment

Single-device solutions: Combining scalar processors, vector processors, and programmable logic

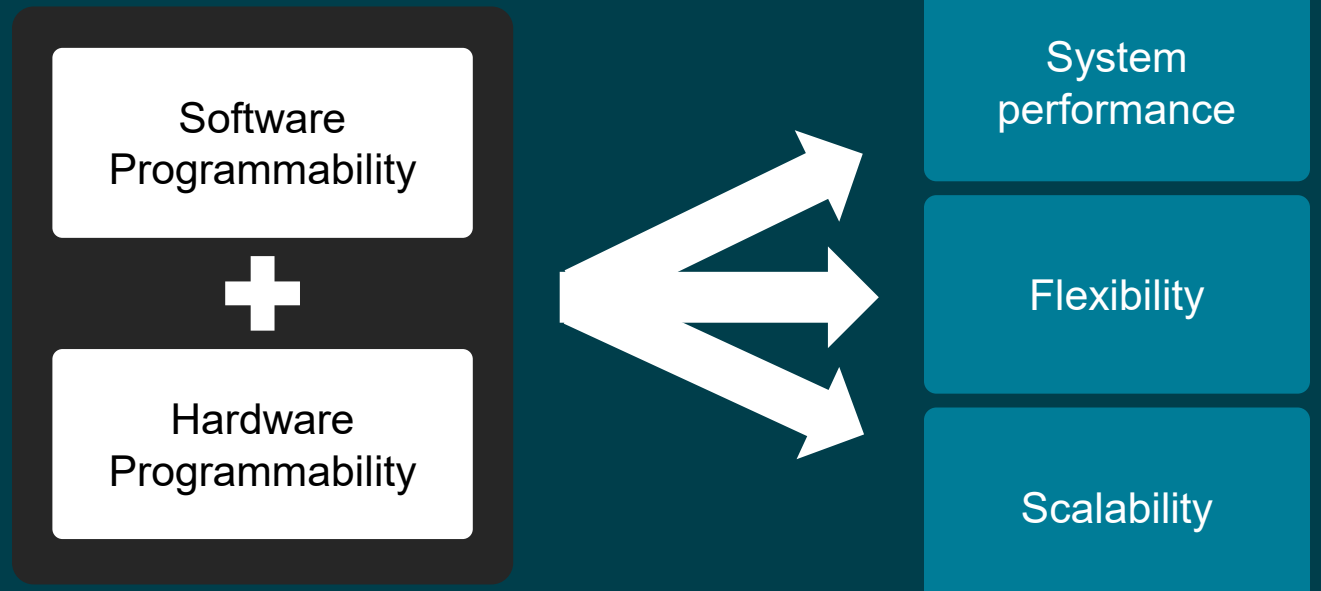
Flexible
Easy to program
Computationally slow



Optimized for vectors
Inefficient for other problems
Challenging to program?

Variable architectures
Challenging to develop

AMD Adaptive System on Chips (SoCs)



System performance

Flexibility

Scalability

Heterogeneous Processing

Multicore Arm® Processing Subsystems

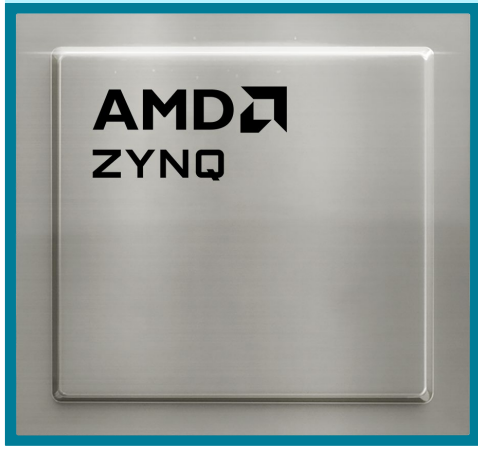
System Integration

Combines processors, fabric, and hardened blocks

AMD Adaptive System on Chips (SoCs)

Select the right SoC from AMD's scalable portfolio

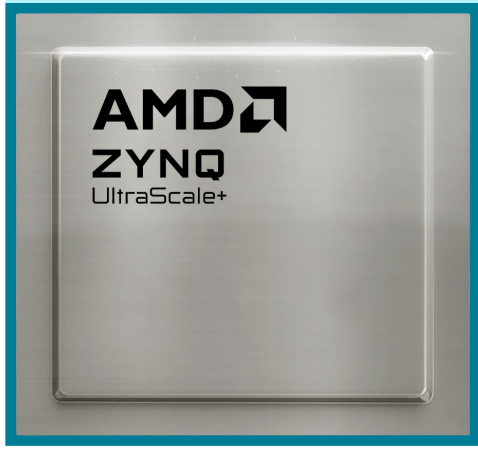
Zynq™ 7000 SoC



Cost-optimized scalable SoC platform

- Single or dual Arm® Cortex®-A9 cores
- 28nm 7 series programmable logic
- Up to 12.5G transceivers

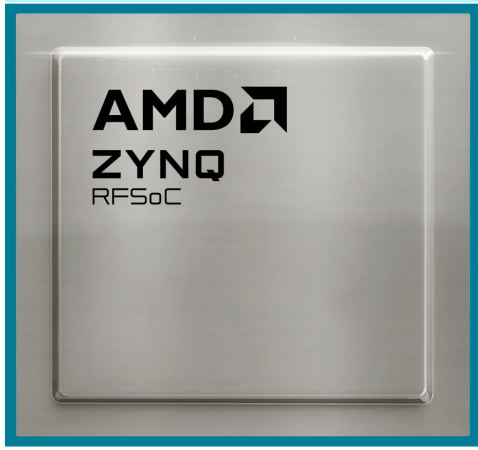
Zynq UltraScale+™ MPSoC



Industry's first heterogeneous adaptive SoC

- Dual or quad Arm Cortex-A53 cores
- Dual Arm Cortex-R5F core
- 16nm FinFET+ programmable logic
- Arm Mali™-400MP2 GPU
- H.264/H.265 video codec

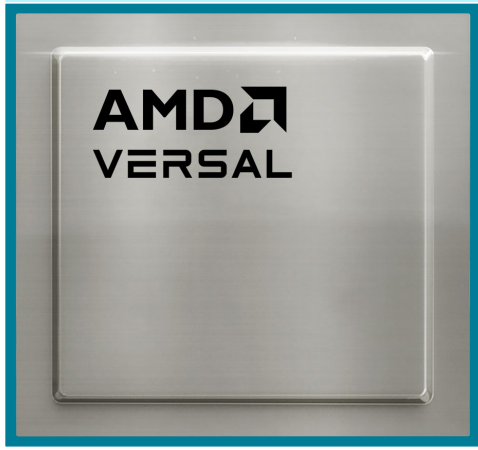
Zynq UltraScale+ RFSoC



Industry's first single-chip adaptive radio platform

- Quad Arm Cortex-A53 cores
- Dual Arm Cortex-R5F cores
- 16nm FinFET+ programmable logic
- Digital RF-ADC, RF-DAC, SD-FEC

Versal™ Adaptive SoC



Adaptive SoC

- Dual Arm Cortex-A72 cores
- Dual Arm Cortex-R5F cores
- 7nm programmable logic
- DSP and AI Engines
- Programmable network on chip (NoC)

Versal Adaptive SoC – Motivation

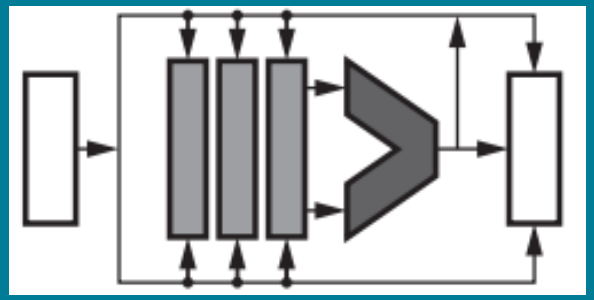
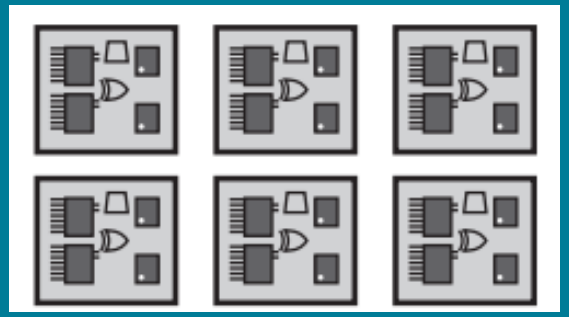
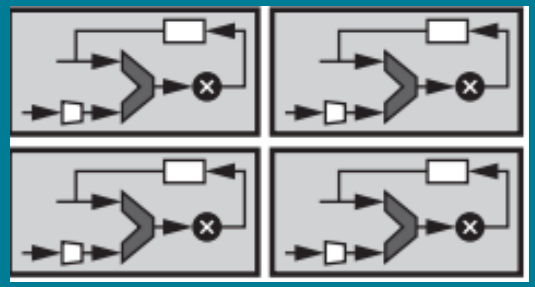
Recent technical challenges: Preventing scaling of the traditional “one-size-fits-all” CPU scalar compute domains

Alternative domain-specific architectures:

Vector-based Processing	Fully Parallel Programmable Hardware
DSPs and GPUs	FPGAs

Versal Adaptive SoC – Motivation

Different tasks benefit from different architectures

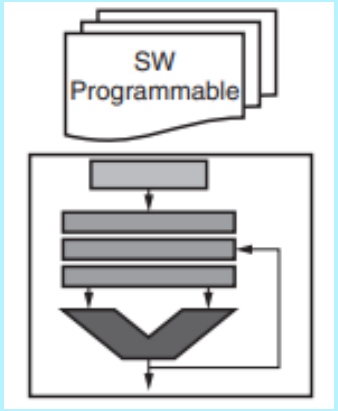
 **Scalar Processing**	 **Adaptable Hardware**	 **Vector Processing** (e.g., DSPs, GPUs)
Complex algorithms and decision making	Latency-critical workloads Automotive driver assistance	Domain-specific parallelism
—	Processing of irregular data structures Genomic sequencing	Signal processing Complex math, convolutions
—	Sensor fusion Pre-processing, programmable I/O	Video and image processing

Types of Compute Architecture

Versal Adaptive SoC – Motivation

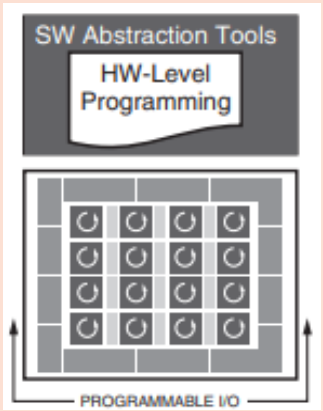
Versal™ adaptive SoC
✓ Delivers the best of all three

CPU



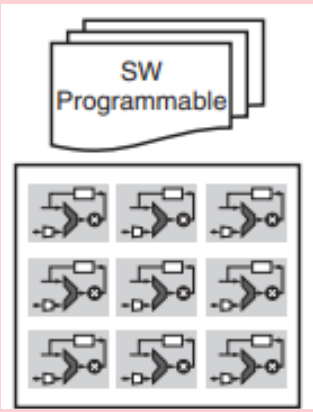
- Scalar, sequential processing
- Memory bandwidth limited
- Fixed pipeline, fixed I/O

FPGA



- Flexible parallel compute
- Fast local memory
- Custom I/O

Vector Processor



- Domain-specific parallelism
- High compute efficiency
- Fixed I/O and memory bandwidth

Processing System ⟷ Programmable Logic ⟷ AI Engines

Integrated Software Programmable Interface

Heterogeneous Software Programmable Interface

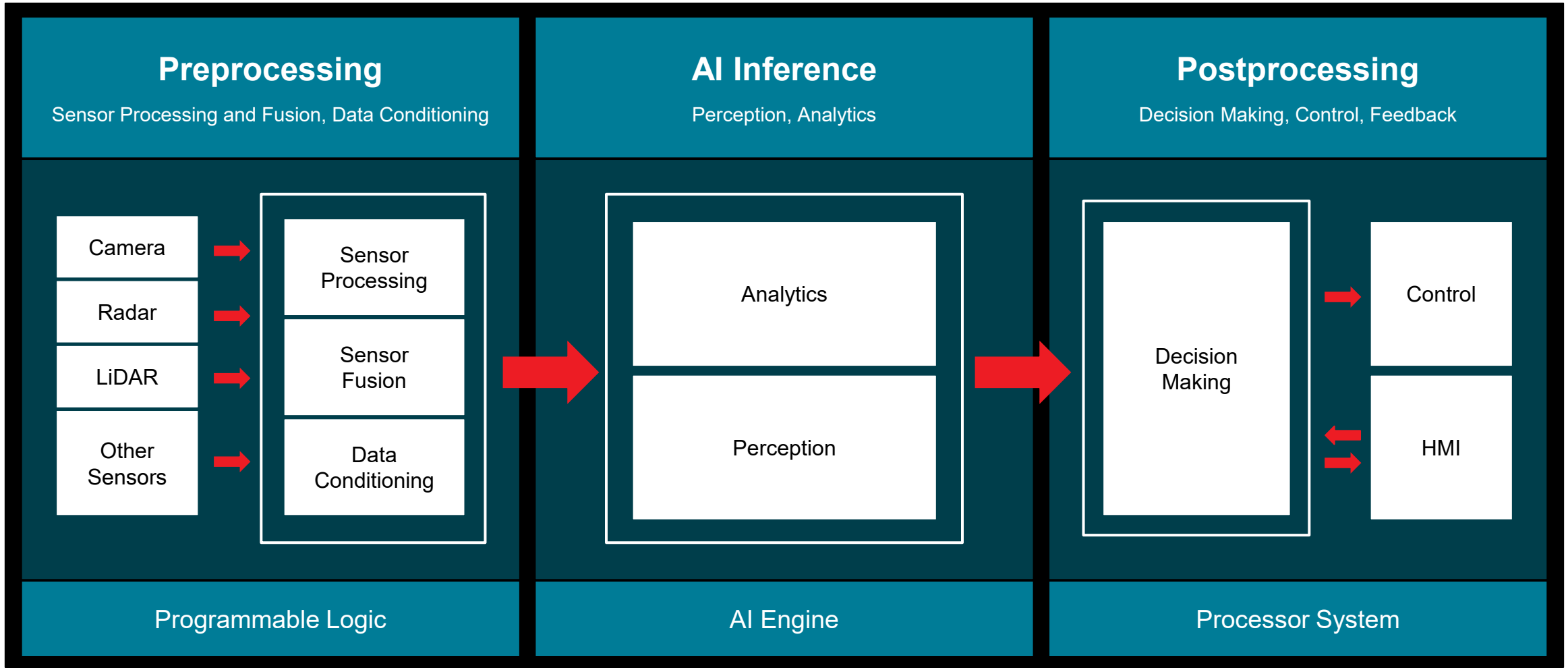
Representative Design

Data center applications accelerate software with hardware “kernels”

Edge applications operate on continuously flowing data requiring a hardware-centric flow

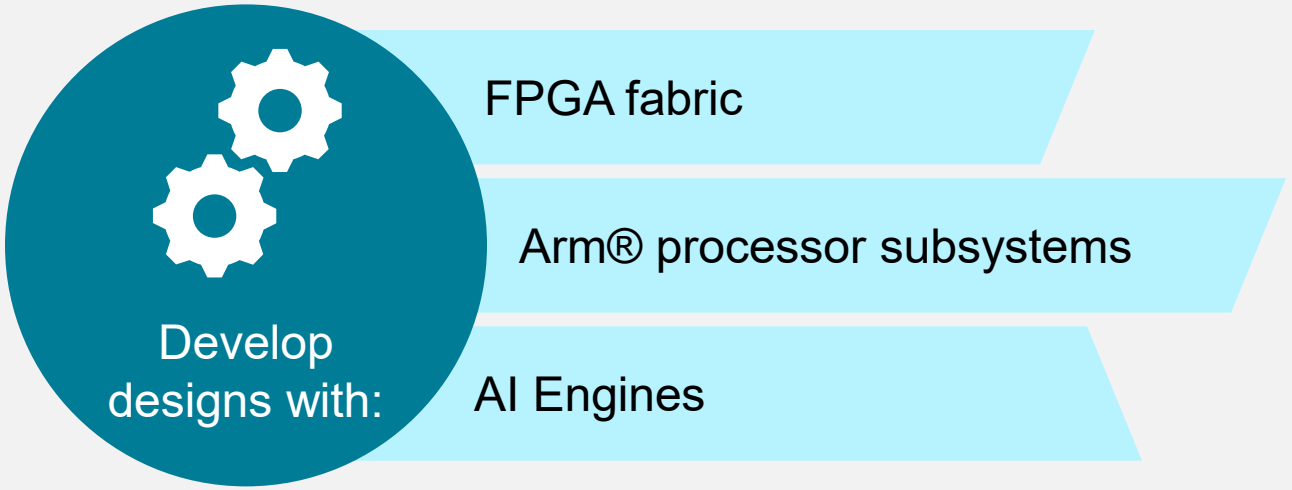
Role of CPUs in This Flow	
Manage: • User interface • Platform management • Housekeeping tasks	Monitor and tweak: • Hardware module parameters

Representative Design



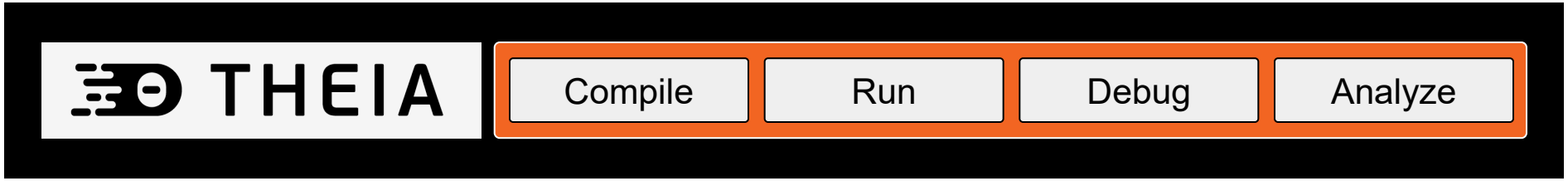
Vitis Software Platform Development Environment

Enables system design for hardware and software developers



Vitis tools work with the AMD Vivado™ ML Design Suite to provide high-level design development abstraction

Vitis™ Unified IDE supports GUI and command-line use



Versal Architecture: Overview

Programmable Logic
- High compute density
- Voltage scaling for perf/watt

Processing System
- Platform control
- Embedded edge compute

PCIe® Gen5 & CCIX
- High PCIe and DMA bandwidth
- Cache coherency

DDR4 Memory
- 3200-DDR4, 4266-LPDDR4
- High bandwidth/pin

Transceiver Leadership
- Broad range, 32G → 112G
- 58G in mainstream devices

AI Engines & DSP Engines
- AI compute
- Diverse DSP workloads

Programmable NoC
- Guaranteed bandwidth
- Efficiently moves data

Protocol Engines
- 400G/600G cores
- Power optimized

Programmable I/O
- Any interface or sensor
- Includes 3.2Gb/s MIPI

Processing Subsystem

Processing Subsystem for Platform Management

- Execute complex algorithms and decision making
- Provide safety processing and redundancy for mission- and safety-critical applications
- Manage the entire platform
- Load each aspect of the Versal™ device and monitor status
- Support capability extension
  - PL-instantiated MicroBlaze™ processor

Processing Subsystem

Application Processing Unit

Based on Arm® Cortex®-A72 dual-core processor

- System memory management unit (SMMU)
- Cache coherent interconnect (CCI) unit
- Interface channels
- System peripherals

SMMU and CCI provide shared memory environment with PS, PMC, and PL processors

Features:

- Up to 1.7 GHz speed
- Armv8 architecture
- Up in seconds
- Supports Linux® and bare-metal

Processing Subsystem

Real-time Processing Unit

Based on Arm® Cortex®-R5F dual-core processor

- L1 caches
- Tightly coupled memories (TCM)
- Configured into dual-processor mode or lock-step mode

Features:

- Functional safety
- Armv7 architecture
- Split mode or lock step
- Low latency, determinism, and real-time control
- ASIL/SIL certifiable

Platform Management Controller (PMC)

Primary Functions

Device Configuration

- Initialization of the device
- Boot and configuration from a supported boot device
- Configure the PL using CFI

Security & Device Integrity

Security core functions

- Encryption and decryption
- Authentication
- Key management

Testing & Debugging

- Provides test and debug infrastructure
- Supports boundary scan and Arm® CoreSight™ trace and debug technology

System Monitor

- Monitors system activity
- Responds to security and functional safety events
- Releases the PS from reset
- Provides system power and error management services

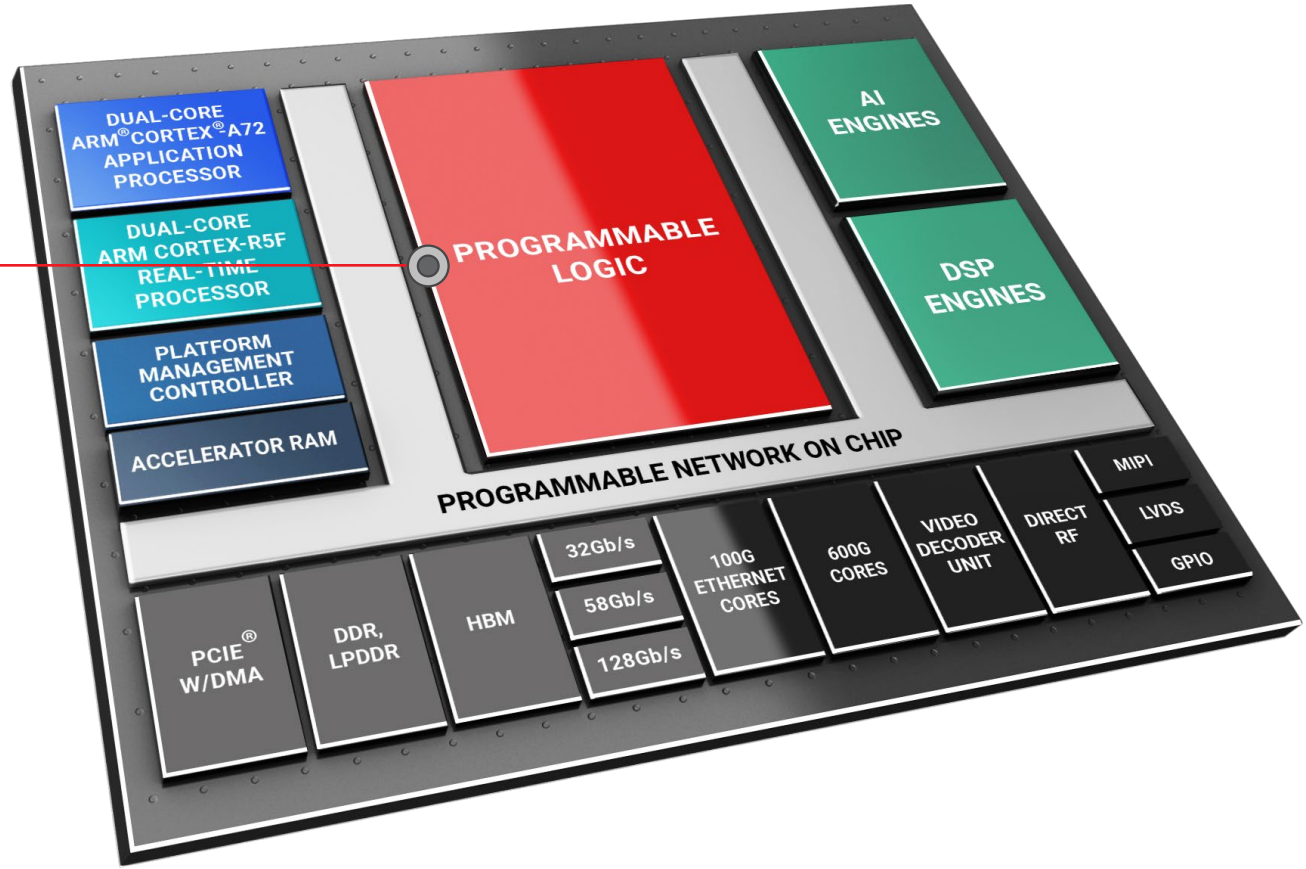
Programmable Logic

Programmable Logic

Millions of reconfigurable system logic cells

Support for parallel, pipelined and hybrid architectures

Wide variety of memory elements



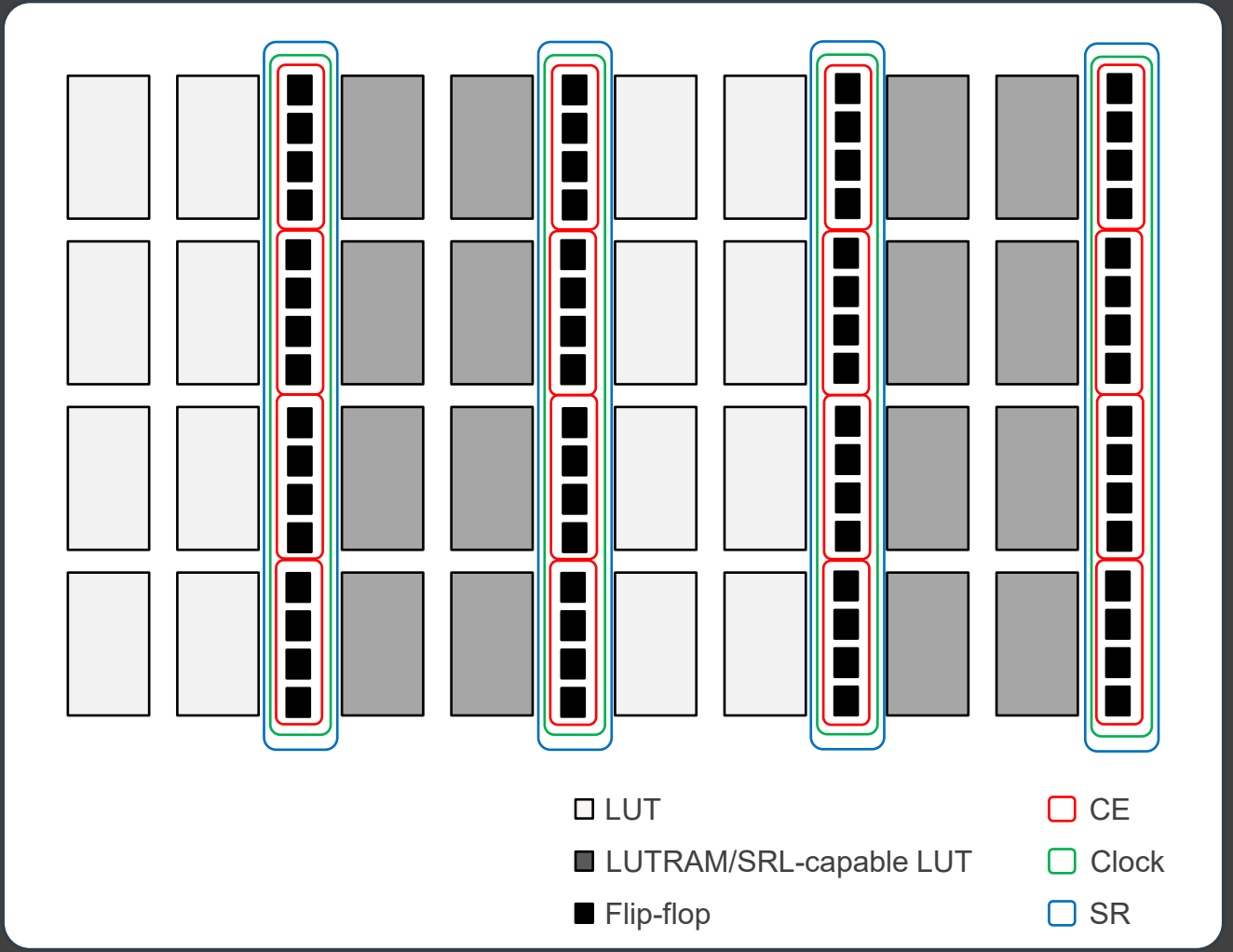
Configurable Logic Block

CLBs: Logic + look-up tables

- Configurable
  - Create special-purpose functions and processing units, etc.

CLB contains:

- 32 LUTs and 64 slice flip-flops
- Arithmetic carry logic and multiplexers
- Control signals
  - 4 clocks, 4 set/resets, 16 clock enables
- 50% of the LUTs form LUTRAM and shift registers
- Dedicated interconnect paths



Clocking

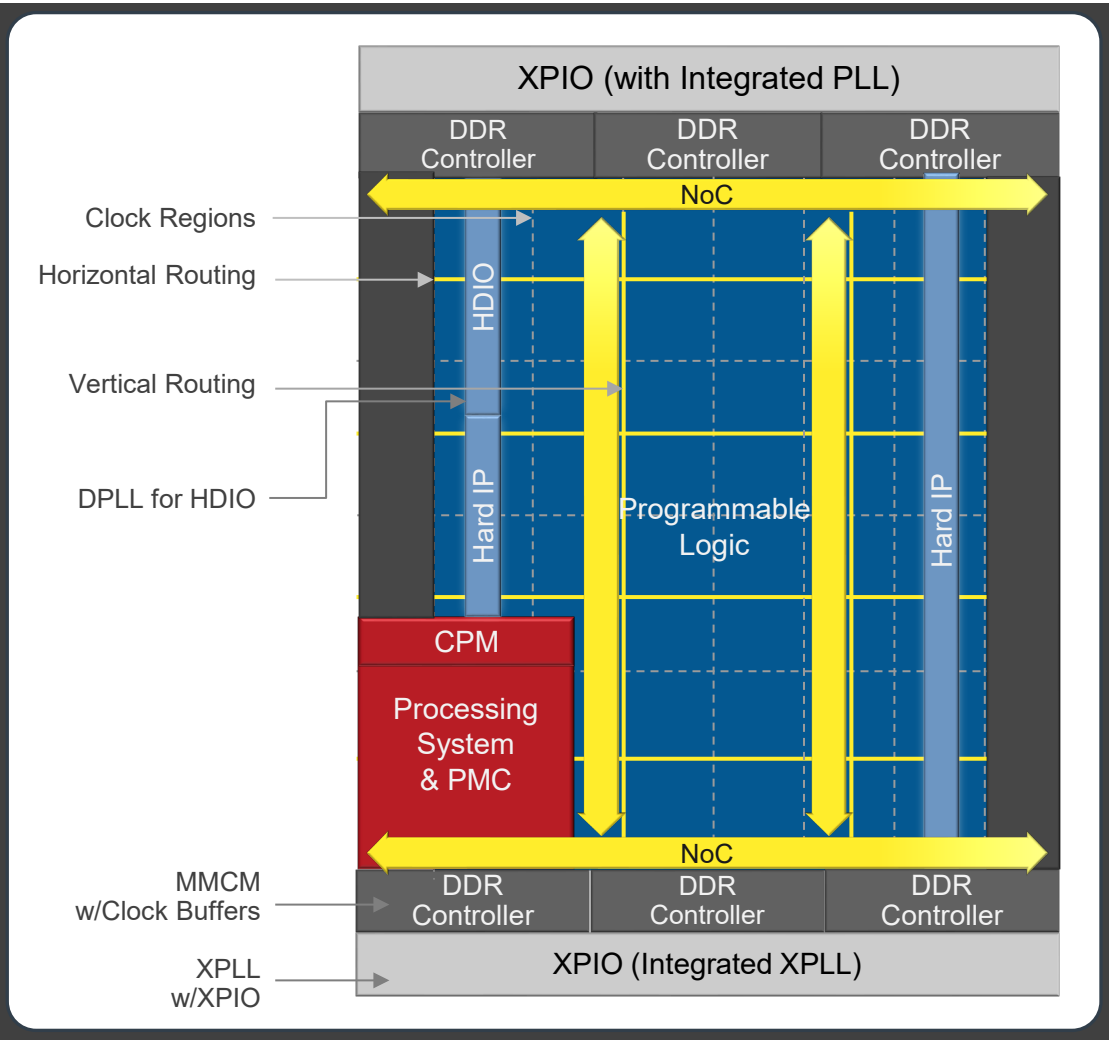
Dedicated global clocks

Bidirectional, segmented network of clock spines

PL subdivided into clock regions

Clock management functions provide:

- Clock frequency synthesis
- Deskew
- Jitter filtering



Programmable I/O

XPIO

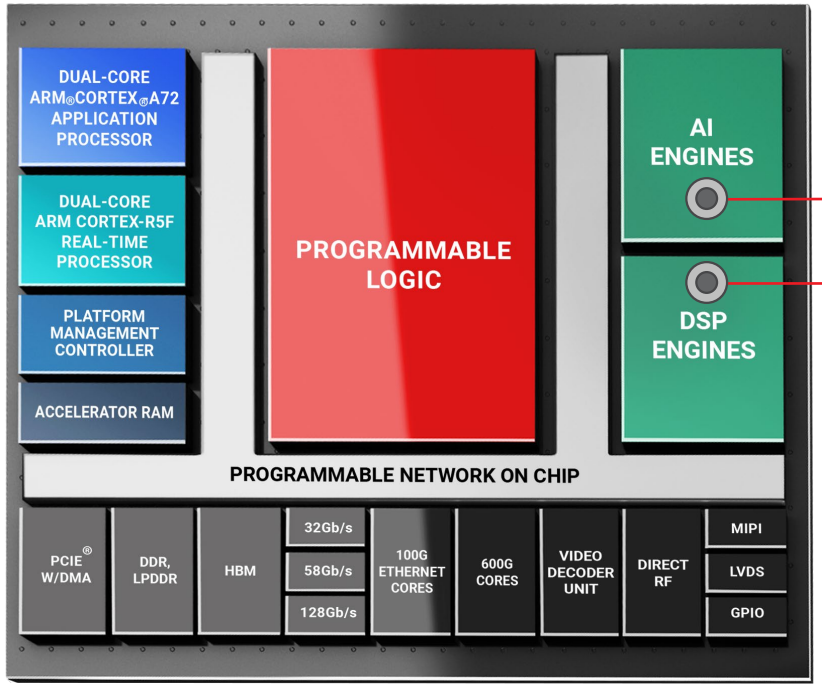
- High-performance XP I/O
- Dedicated logic
- Support for interfaces 1.0V – 1.5V
- Grouped into 54-pin banks
- Each XPIO can use the XPHY

HDIO

- High-density HD I/O
- Support for interfaces 1.8V – 3.3V
- Grouped into 22-pin banks

- No contradicting voltages or I/O standards per bank
- Support for low-speed SDR and DDR interfaces and coarse data alignment
- I/O buffers support many standards
- Processor peripherals routed to separate banks

AI Engines + DSP Engines



Wired communications, automotive, and consumer markets

AI Is Everywhere

For Diverse Compute

DSP Engines

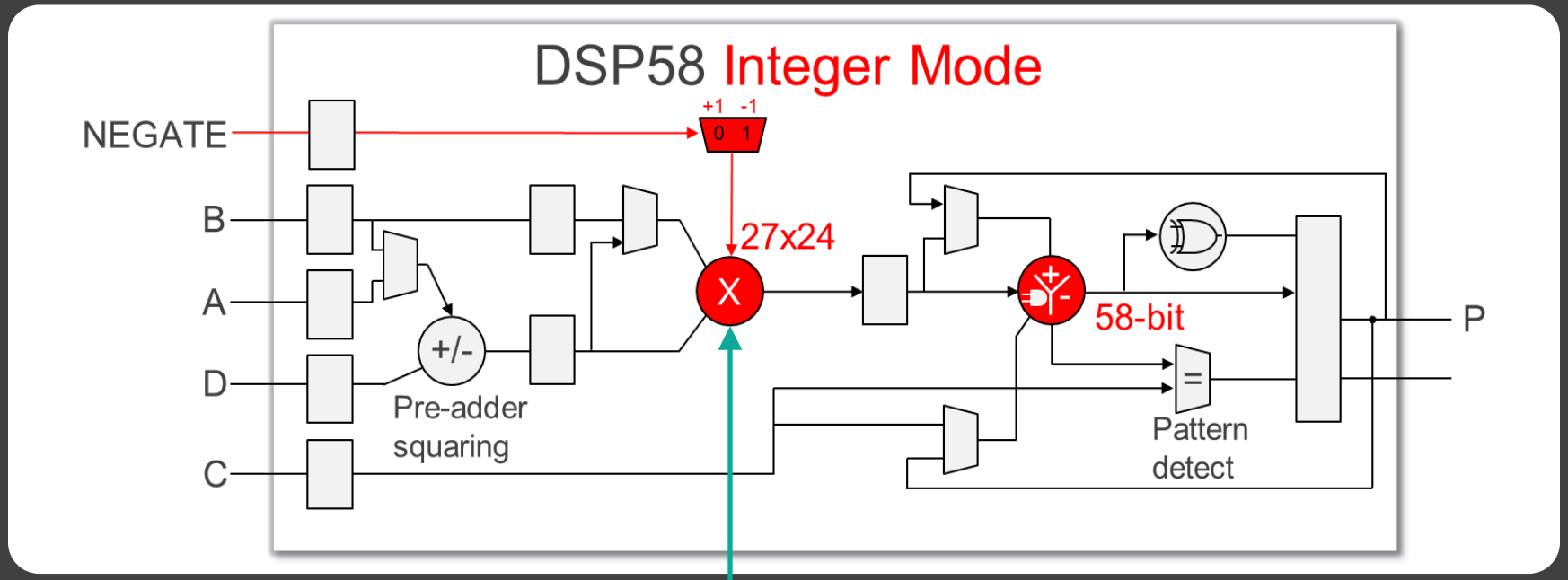
- High-precision, floating-point computation support
- Offload additional functions for acceleration

AI Engines

- High throughput, low latency, and power efficient
- Ideal for AI inference and advanced signal processing
- Array of VLIW processors with SIMD vector units
- Instruction-level and data-level parallelism

Digital Signal Processing Capability

High speed + small size offers system design flexibility



- Dynamically bypassable multiplier
- Two 58-bit inputs feed SIMD arithmetic unit
- Logic unit generates solution on the two operands

Supports dynamic negate post-multiplier

Digital Signal Processing Capability

DSP Engine Modes

- Fixed Point
- Single Precision Floating Point
- Int8 Dot Product
- 18x18 Complex Multiply and Accumulate

Programmable NoC: Bridging Blocks & Hard IP

High-bandwidth, terabit programmable NoC

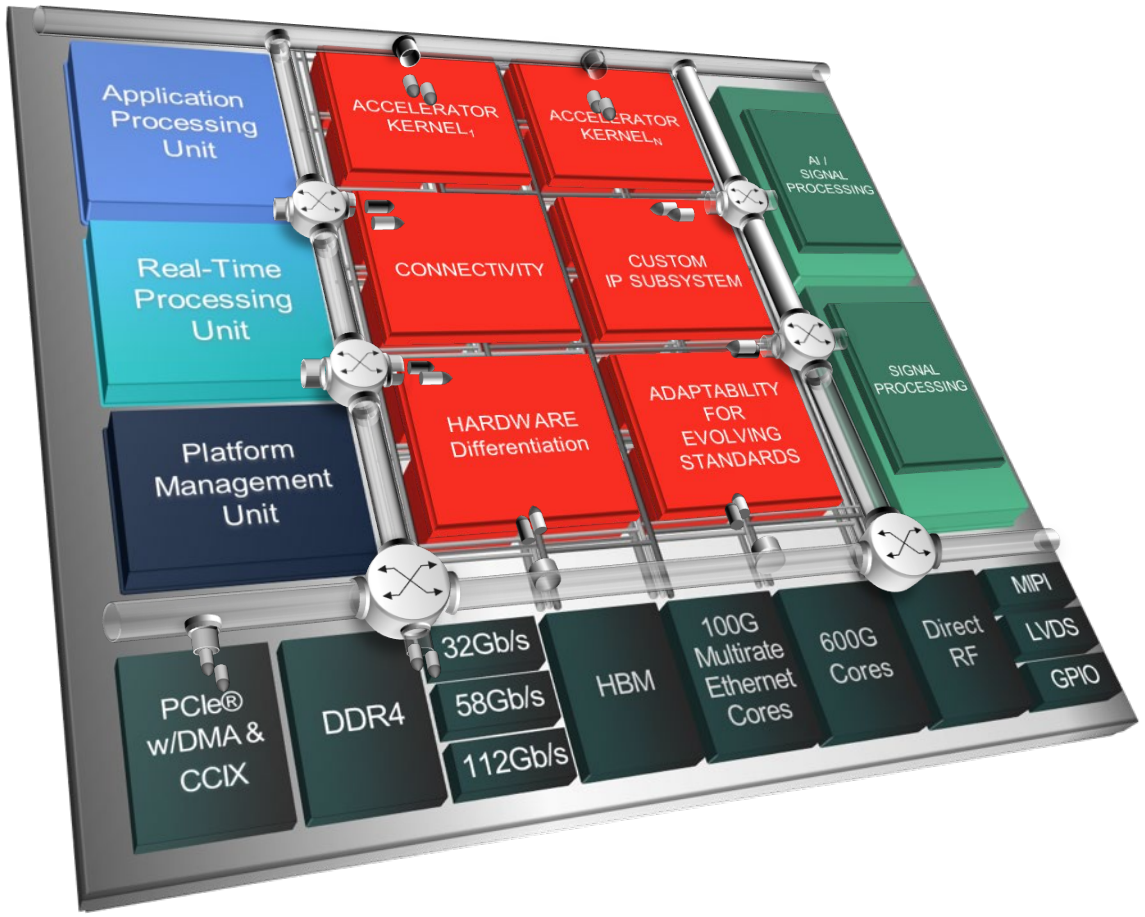
- Programmable among PS, high-speed I/Os, and engines
- Priority based on throughput, latency, and bit widths
- Guaranteed QoS

Eases IP and kernel placement

- Kernel compartmentalization with easy entry and exit points
- Any master to any slave connection
- AXI4-based network
- Hardened connectivity to integrated memory controllers and PCIe® core

Programming framework

- Memory-mapped access to all resources
- Built-in arbitration between compute resources and memory



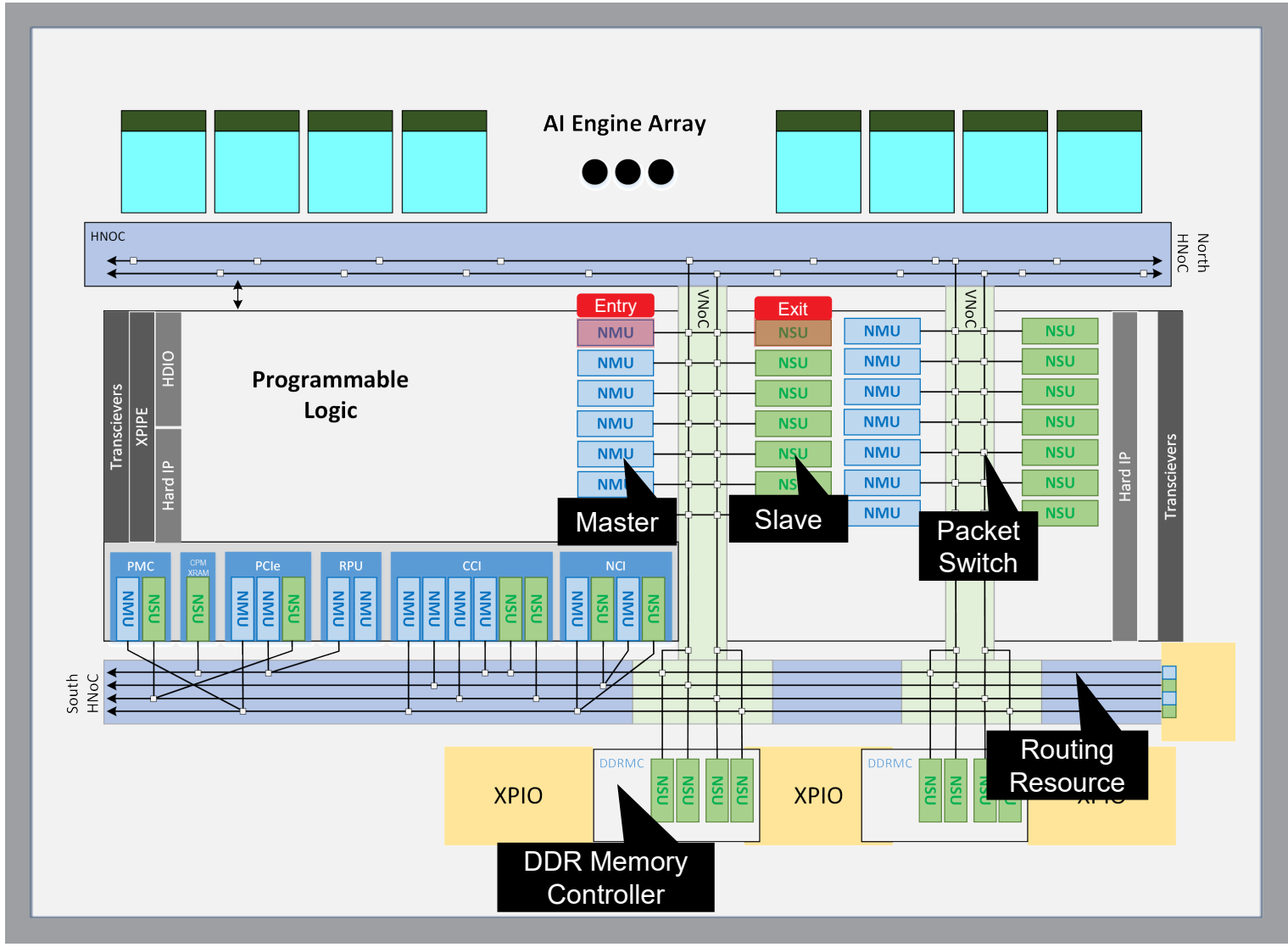
The NoC Dissected

NoC – network on chip subsystem

Entry/Exit Points

Entry: NMU

Exit: NSU



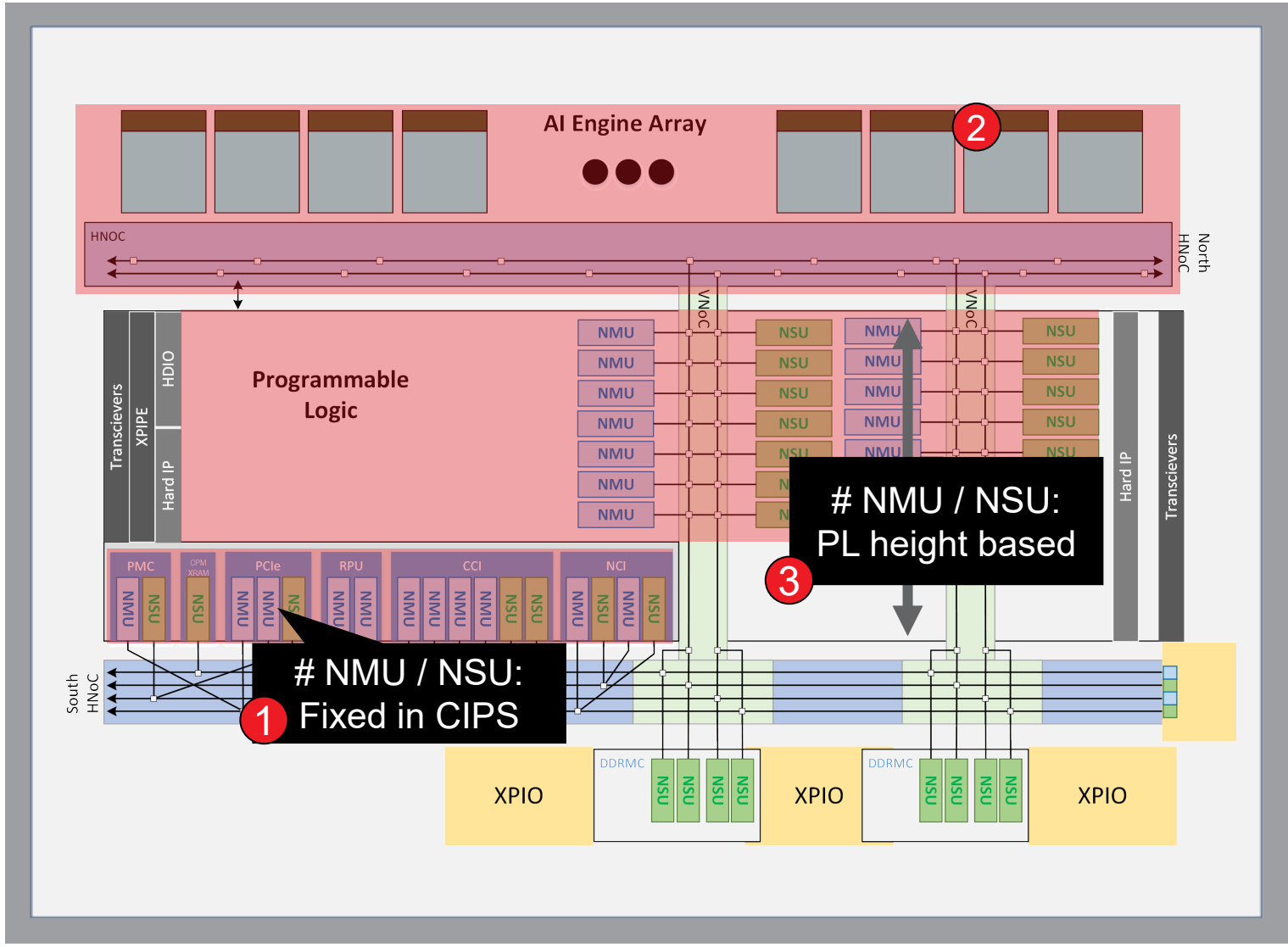
The NoC Dissected

NoC – network on chip subsystem

Entry/Exit Points

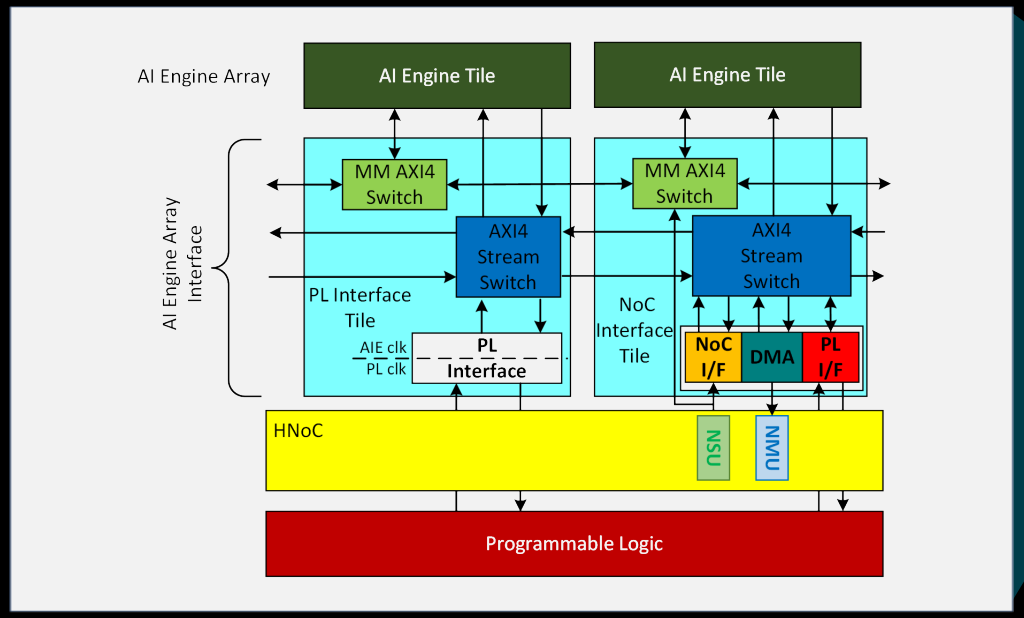
Entry: NMU

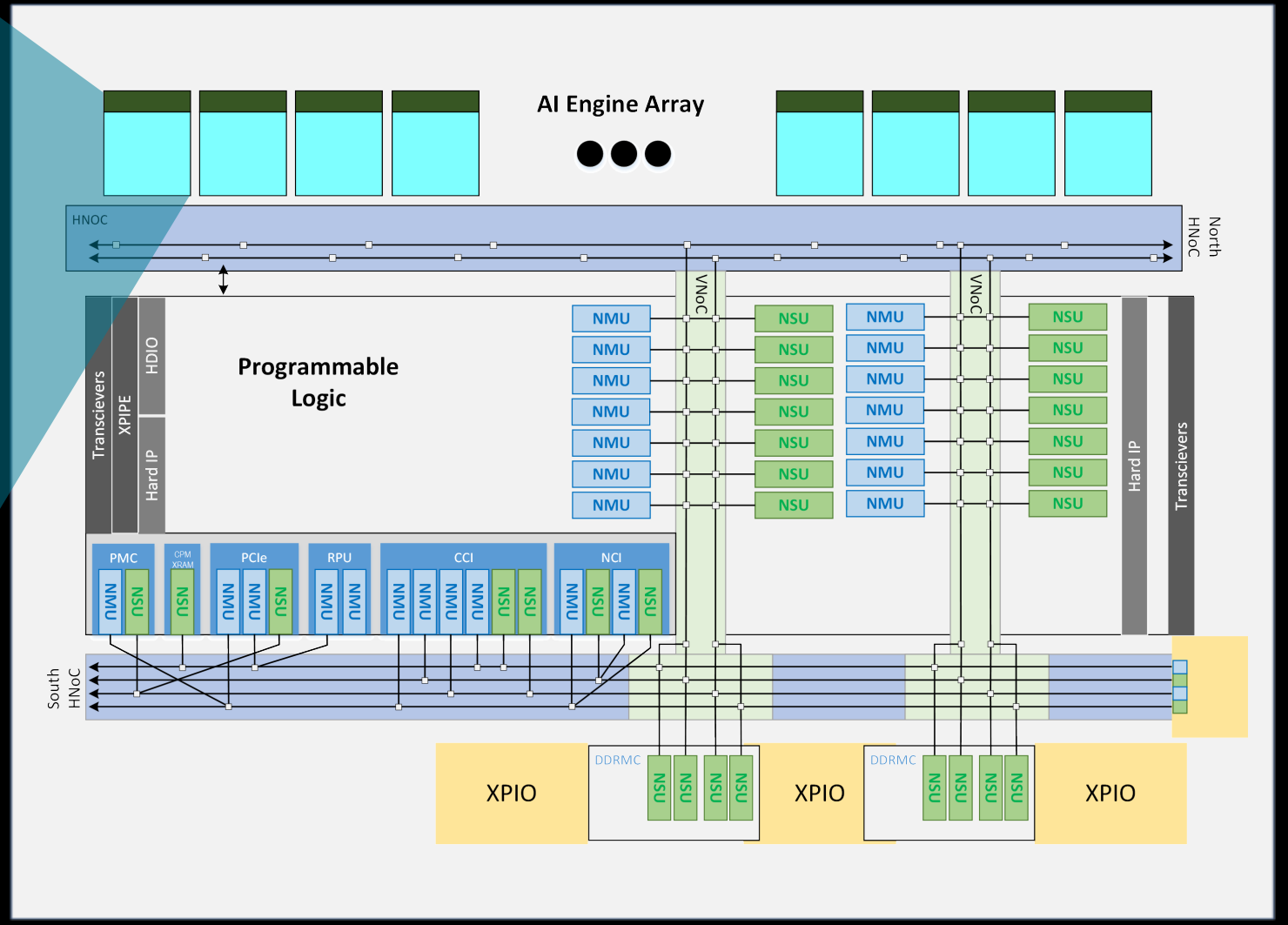
Exit: NSU



AMD
together we advance_

The NoC Dissected





- AIE NoC connections vary by part
- AIE elements share NoC Interface tile

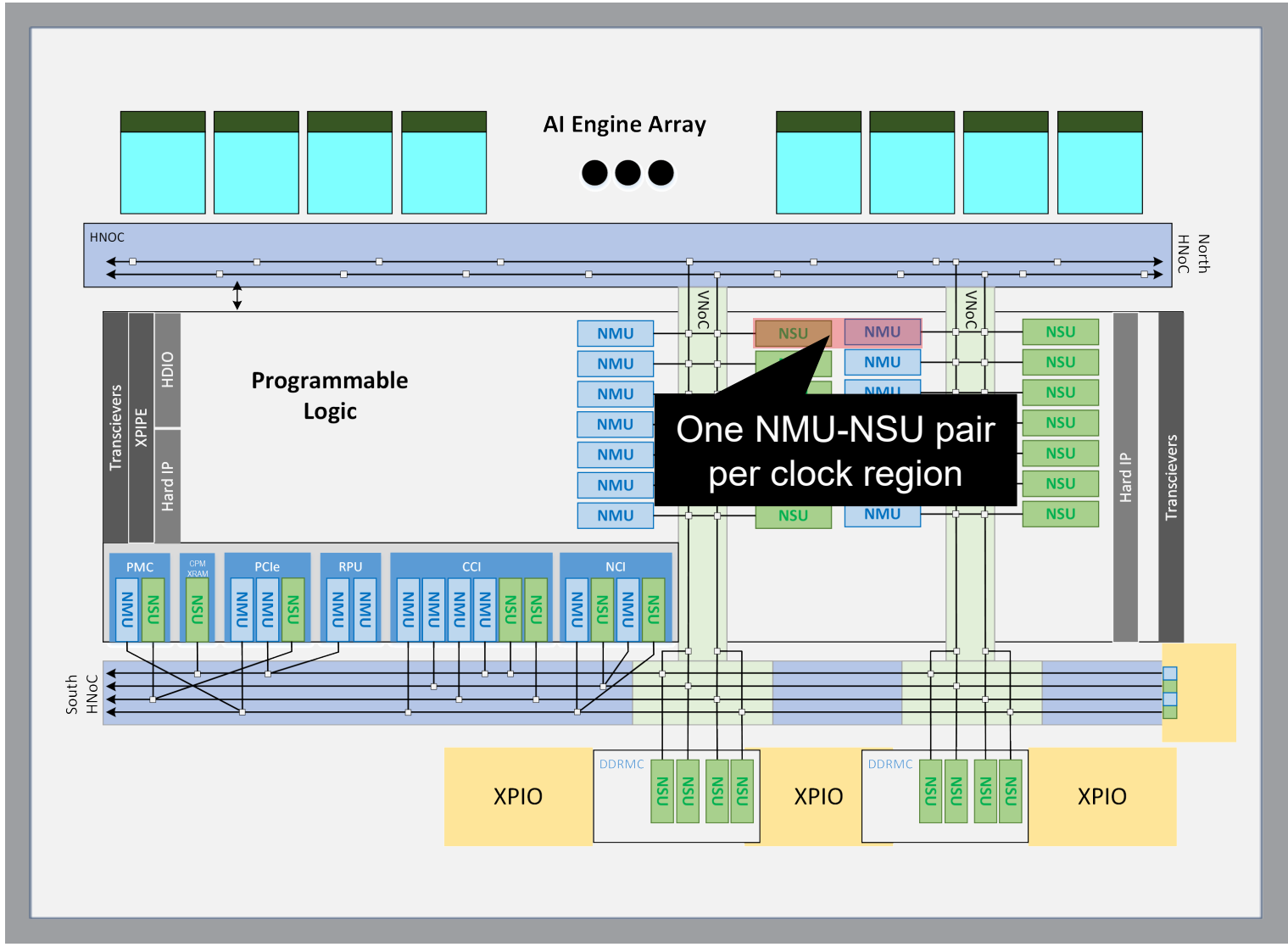
The NoC Dissected

NoC – network on chip subsystem

Entry/Exit Points

Entry: NMU

Exit: NSU



AMD
together we advance_

The NoC Dissected

NoC – network on chip subsystem

Entry/Exit Points

Entry: NMU

Exit: NSU

NPS at Connection Points

Buffer and retransmit

NoC Channels

128 bit wide

Dual-simplex pair of connections

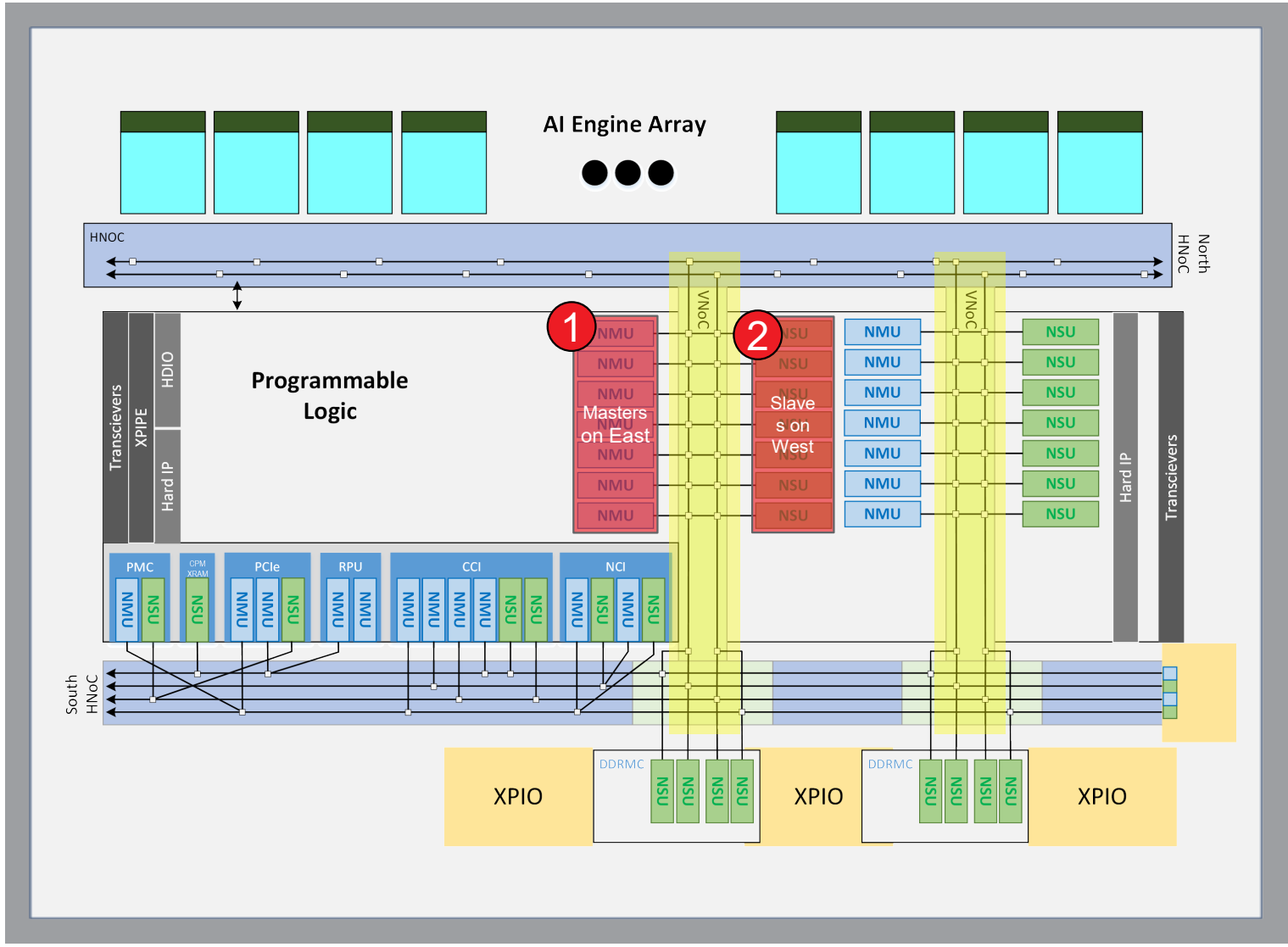
NoC Routes

Horizontal

Vertical

DDR Memory Controller

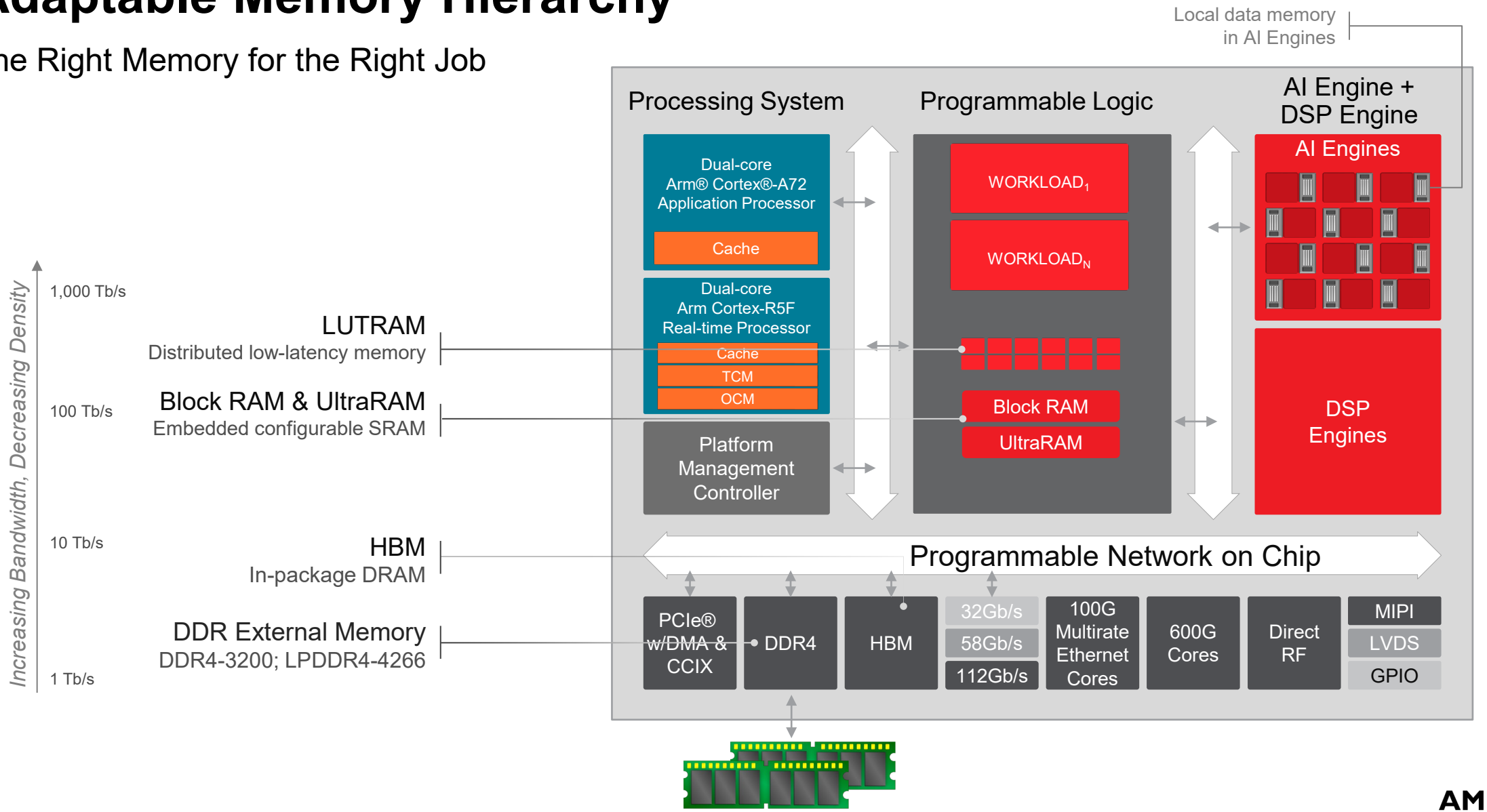
Manages access to memory



AMD
together we advance_

Adaptable Memory Hierarchy

The Right Memory for the Right Job



AMD
together we advance_

CPM Architecture – Caching and PCIe Module

Main components

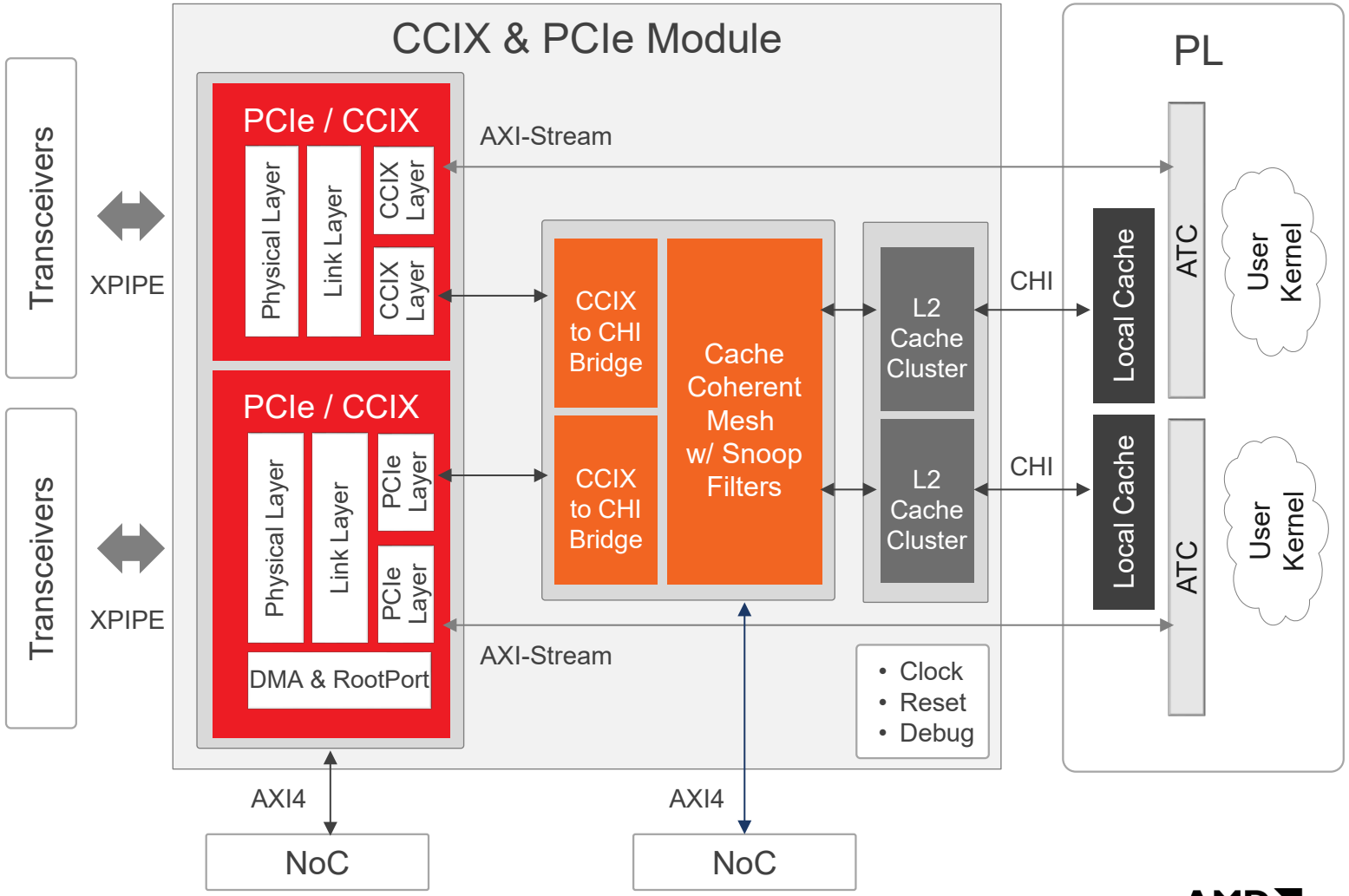
Two integrated PCI Express® controllers

- PCIe® 0 and PCIe 1
- CCIX capability
- Operate as endpoints in CCIX mode

Cache coherency block

- Specifies CPM cache and coherency architecture
- CCIX gateway to on-chip and off-chip processing nodes and memory

CPM connects to the transceivers, PL, and PS



GENERAL DISCLOSURE AND ATTRIBUTION STATEMENT

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this material, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this material, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD’s products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18.


AMD