Overview of Alveo acceleration using Vitis HLS and Vivado

Bart Handels – FAE AECG Benelux

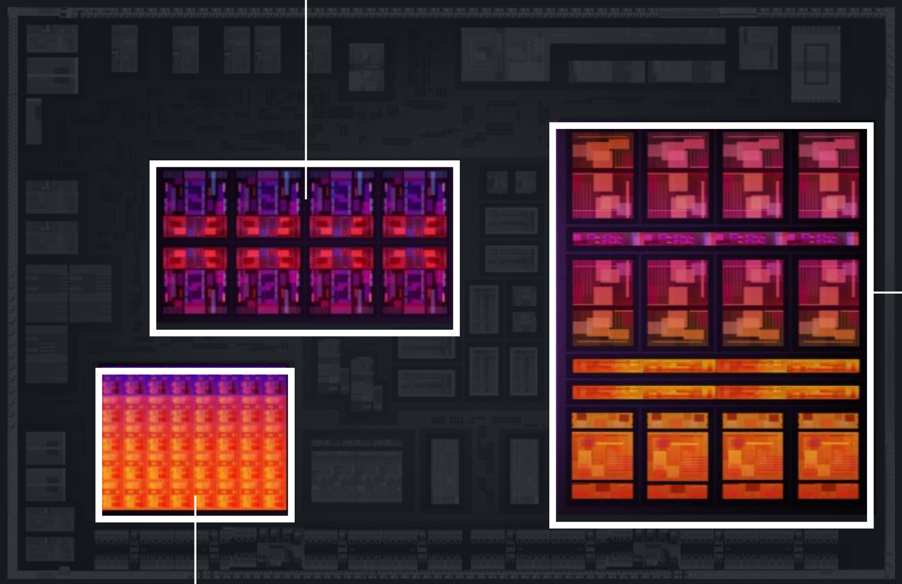
AMD
together we advance_

Ryzen AI 300 - Providing the Next Level of NPU, CPU, and GPU Architectures for Next-Gen AI PC Experiences

3rd Generation

AMD Ryzen™ AI

Best in class AI platform



AMD RDNA 3.5

Next-Gen GPU

Up to 16 Compute Units

ZEN 5

Next-Gen CPU

Up to 12 Cores, 24 Threads

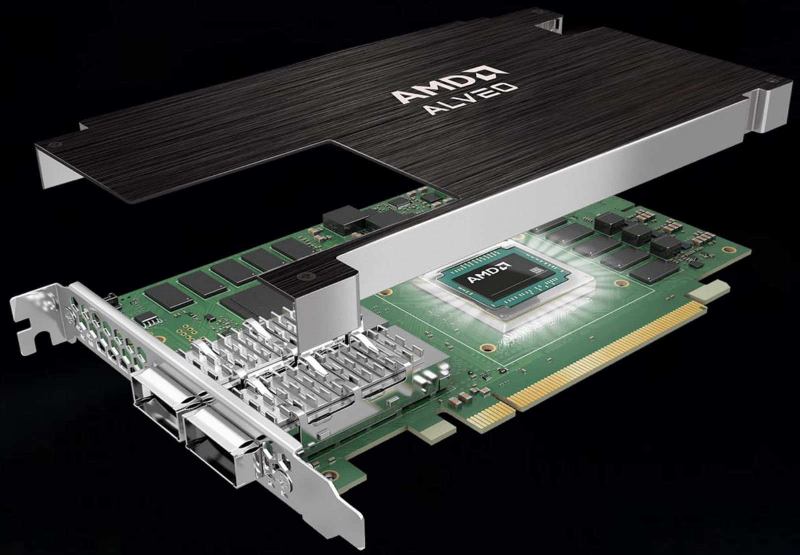
AMD XDNA 2

Next-Gen NPU

Industry-leading 50+ NPU TOPS

See endnote STX-04 and GD243

AMD together we advance_



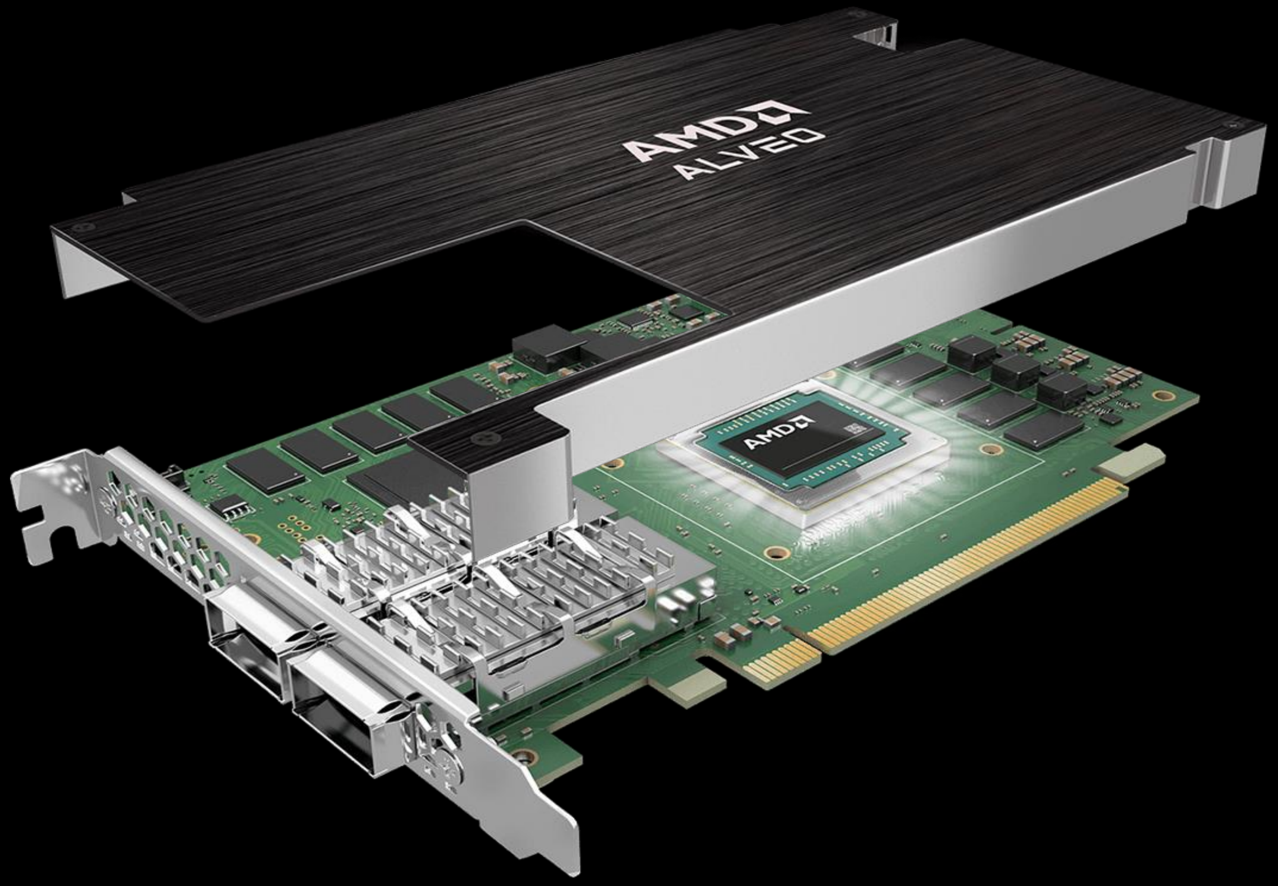
Alveo

AMD Alveo™ Accelerator Cards for Broad Application

Production FPGA PCIe® Cards
Fast Time to Deployment, Out-of-the-Box Ready

Traditional FPGA Design Flows
AMD Vivado™ Flow Support for Hardware Flexibility

Multi-Market Application
Data center workloads or specialized functionality

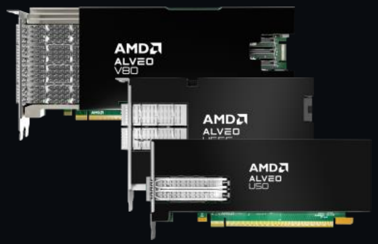


AMD Alveo™ Portfolio of Accelerator Cards

- General-purpose compute cards with high logic density, DSPs, on-board DDR and/or HBM
- Network accelerators to offload network and data center infrastructure tasks from CPUs
- Low latency and ultra-low latency cards for high speed networking
- Domain-specific solutions for AI inference and video streaming

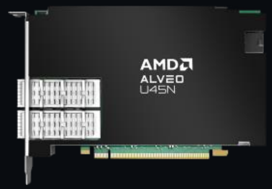
General Compute

Alveo™ U50, U55C, V80



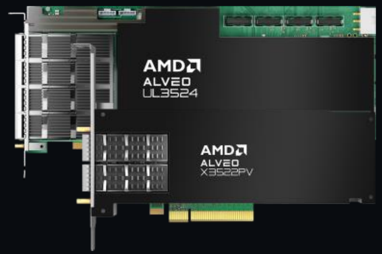
Networking

Alveo U45N



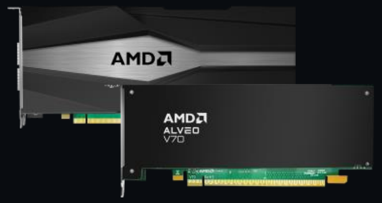
Ultra-Low Latency

Alveo™ UL3524, X3522PV



AI Engines

Alveo V70, VCK5000



Video Streaming

Alveo U30, MA35D

1: Alveo V70 supported by Vitis AI only
2: Vitis™ and Vivado™ not supported for Alveo U30 and MA35D, supported by AMD Media Acceleration SDK

Vivado™ Supported Alongside Vitis™ Unified Software Platform

Software Developers

- Frameworks
- Accelerated Libraries
- Application developers

AMD Vitis

Hardware Developers

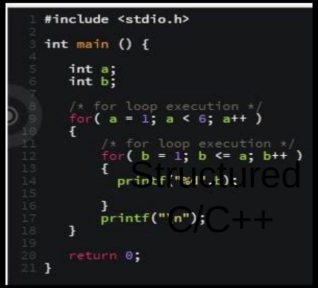
- Familiar FPGA Design Flows
- IP
- Reference Designs

AMD Vivado

Hardware Devices and
Accelerator Cards

Vitis HLS

Vitis HLS advantage



Automated C/C++ to RTL Conversion

Allows Significantly Faster Design Iterations

RTL Code

Significantly Accelerates Simulation

AMD Vitis

- Write initial design spec in HLS friendly C
  - Modify the C-code
  - Modify the C-code
  - Modify the C-code
  - Modify the C-code

Typical to have 10-20 iterations
Days long vs Weeks long for RTL based iterations

Input	RTL Simulation Time	C-Simulation Time	Acceleration
10 frames of video data	~ 2 days	10 seconds	~12,000X

Vitis HLS Design methodology in 3 steps

1 Define Performance Specification

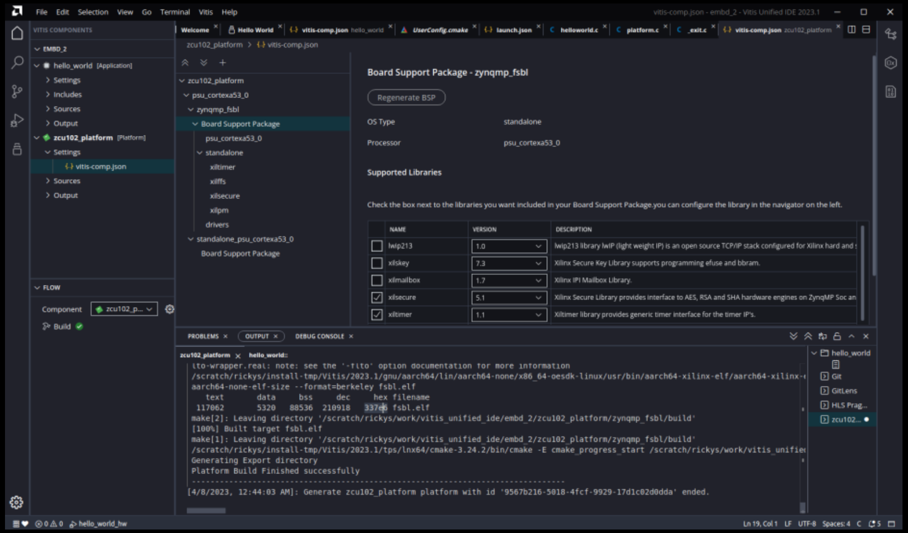
2 Build Macro Architecture

3 Refine Micro Architecture For Optimization

```
*/
#include "test.h"

void splitter(hls::stream<int>& in, hls::stream<int>& odds_buf,
                hls::stream<int>& evens_buf) {
    int data = in.read();
    if (data % 2 == 0)
        evens_buf.write(din: data);
    else
        odds_buf.write(din: data);
}

void odds(hls::stream<int>& in, hls::stream<int>& out) {
    out.write(din: in.read() + 1);
}

void evens(hls::stream<int>& in, hls::stream<int>& out) {
    out.write(din: in.read() + 2);
}

void odds_and_evens(hls::stream<int>& in, hls::stream<int>& out1,
                    hls::stream<int>& out2) {
    hls_thread_local hls::stream<int, N / 2> s1; // channel connecting t1 and t2
    hls_thread_local hls::stream<int, N / 2> s2; // channel connecting t1 and t3

    // t1 infinitely runs func1, with input in and outputs s1 and s2
    hls_thread_local hls::task t1(fn: splitter, in, s1, s2);

    // t2 infinitely runs func2, with input s1 and output out1
    hls_thread_local hls::task t2(fn: odds, s1, out1);

    // t3 infinitely runs func3, with input s2 and output out2
    hls_thread_local hls::task t3(fn: evens, s2, out2);
}
```

Vitis HLS Keeps evolving

- New Unified Vitis IDE based on THEIA
- Vitis Library as IP
- Vitis Library access through Github
- Arbitrary Precision Floating Point
- Adding RTL as black box in HLS
- Compile time and QoR improvements

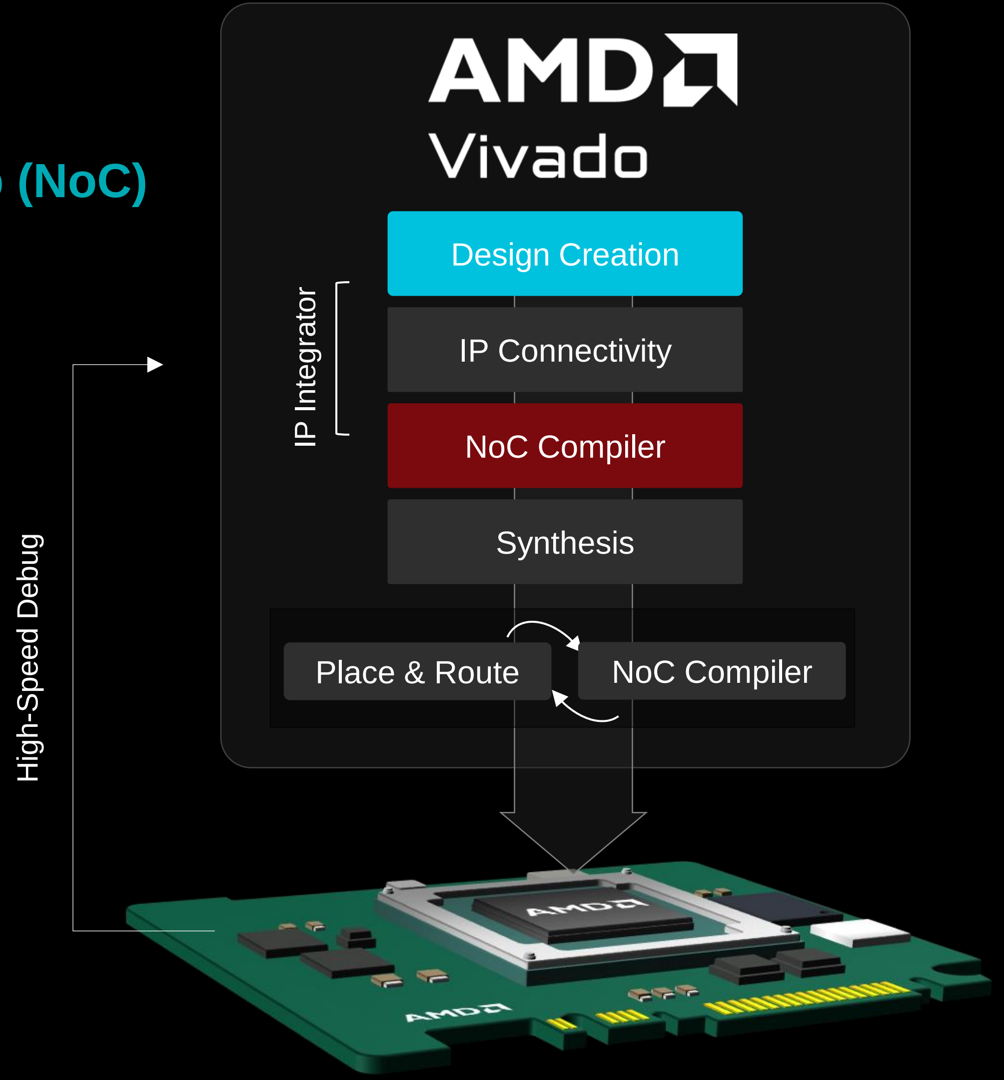


Alveo for Hardware developers

AMD Vivado™ Design Flow

for AMD Alveo™ V80 Compute Accelerator Card

- **Modular IP Integration with AMD Vivado & Network on Chip (NoC)**
  - Graphically connect hard/soft IP using Vivado IP Integrator
  - Streamlined, GUI-based flow with NoC Compiler
  - NoC ensures timing for critical interconnect paths
- **Traditional FPGA development Flow**
  - Design in RTL
  - Leverage familiar building blocks in IP integrator
  - Standard synthesis and P&R
- **High-Speed, Unified Debug Environment**
  - High-bandwidth, SerDes-based debug and trace
  - Fast readback
  - Cohesive debug across engines



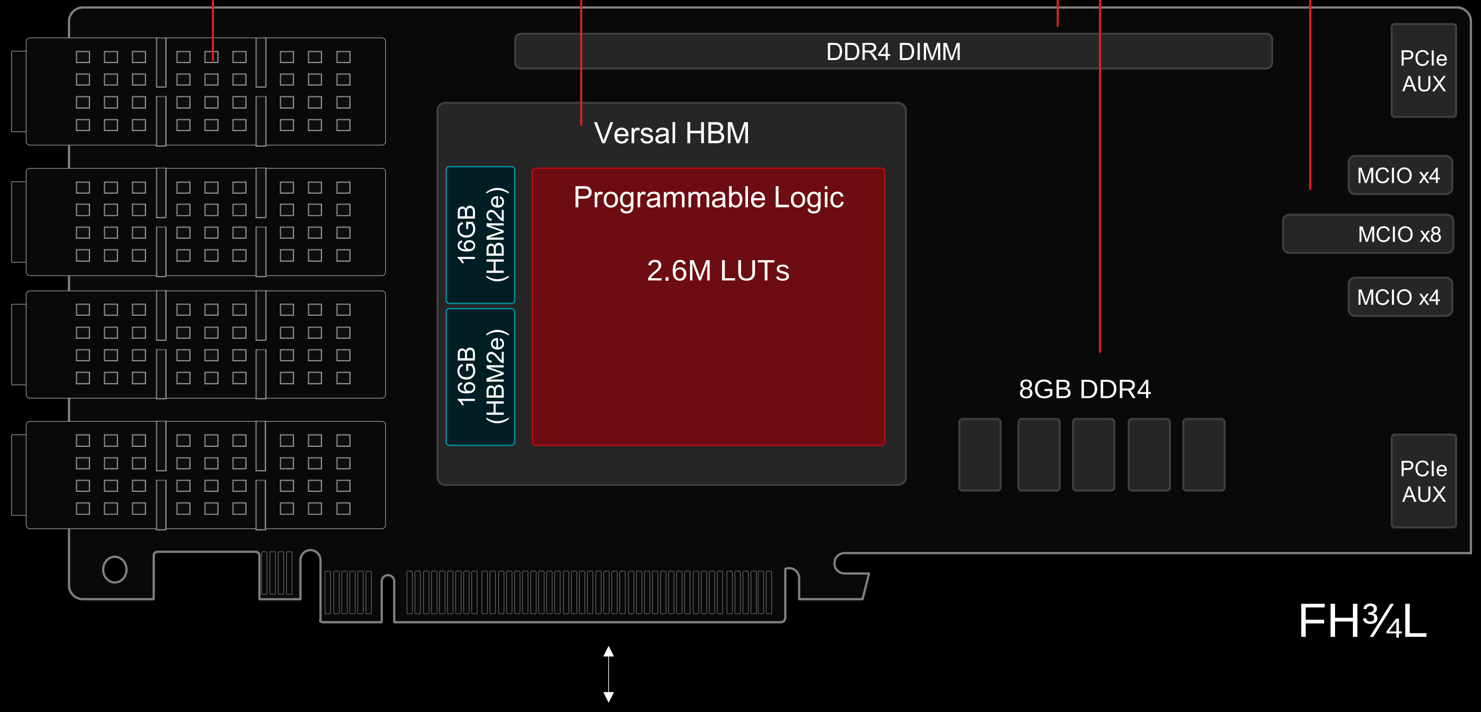
AMD Alveo™ V80 Specifications

800G Networking
- 4x200G
- QSFP56 ports

7nm AMD Versal Architecture
- 2.6M LUTs for flexible compute
- 10.9K DSP slices
- 32GB HBM at 820GB/s

On-Board DDR
- 8GB for Arm® processor management
- DIMM Expansion slot

MCIO Expansion
- PCIe® Gen5 connectivity
- Connect to NVMe

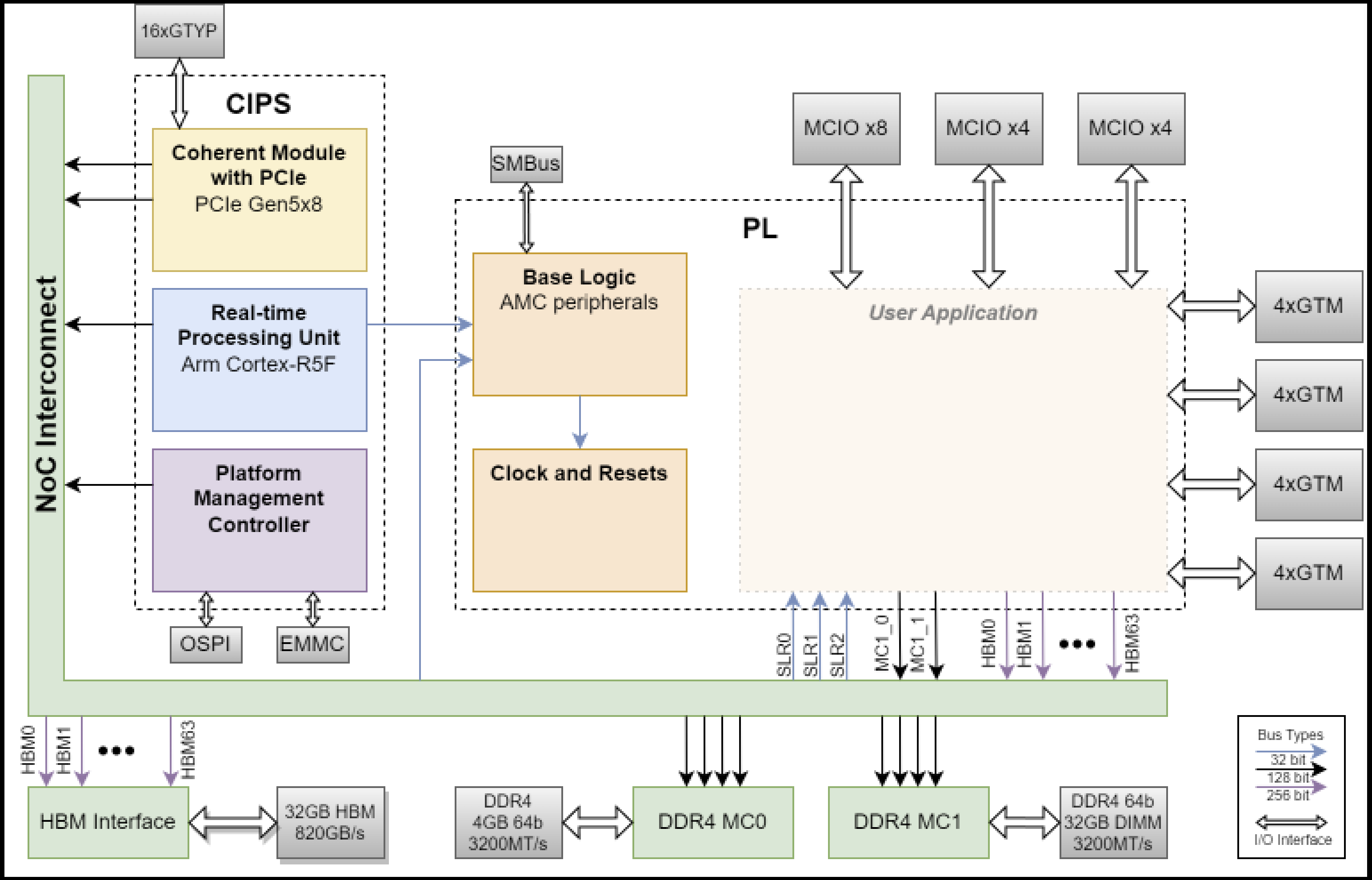


FEATURES	SPECIFICATION
Device Architecture	XCV80 (AMD Versal™ HBM adaptive SoC)
Logic Density	2.6 Million LUTs
HBM Capacity	32GB
HBM Bandwidth	820GB/s
DDR4 Capacity	32GB
Network Interface	• 4x200G (QSFP56) • Per port: 2x100G or 4x 10/25/40/50G
Expansion	PCIe® Gen5 over MCIO connectors
Form Factor	Full-height, ¾ Length (FH 3/4L), Dual-Slot
PCIe® Interface	PCIe® Gen4 x16, 2x Gen5 x8
Power (Electrical)	300W
Power (Thermal)	190W[1], passively cooled
Software	• AMD Vivado® Design Suite (RTL) • AMD Alveo Versal Example Design (AVED)

1: Total thermal power (TDP) is device and server dependent

AMD together we advance_

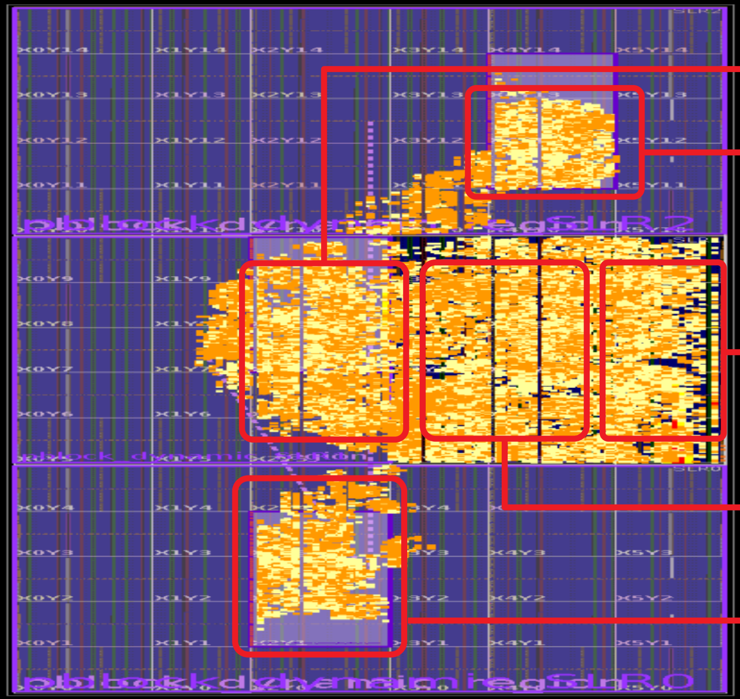
Alveo V80 AVED architecture



Integrated Shell Frees More FPGA Logic for Differentiation

AMD Virtex™ UltraScale+™ VU9P FPGA

200K LUTs Used for Infrastructure



Memory Controller

Memory Controller

PCIe®+DMA

Memory Controller

Memory Controller

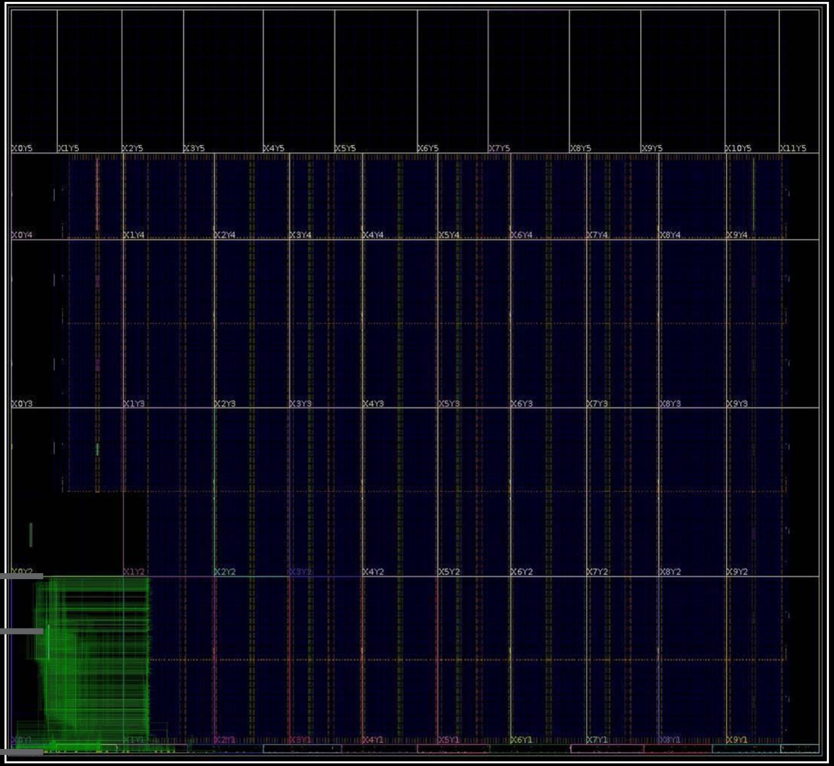
AMD Versal™ Device

Zero LUTs Used for Infrastructure

PCIe+DMA

Processor Subsystem

Memory Controllers



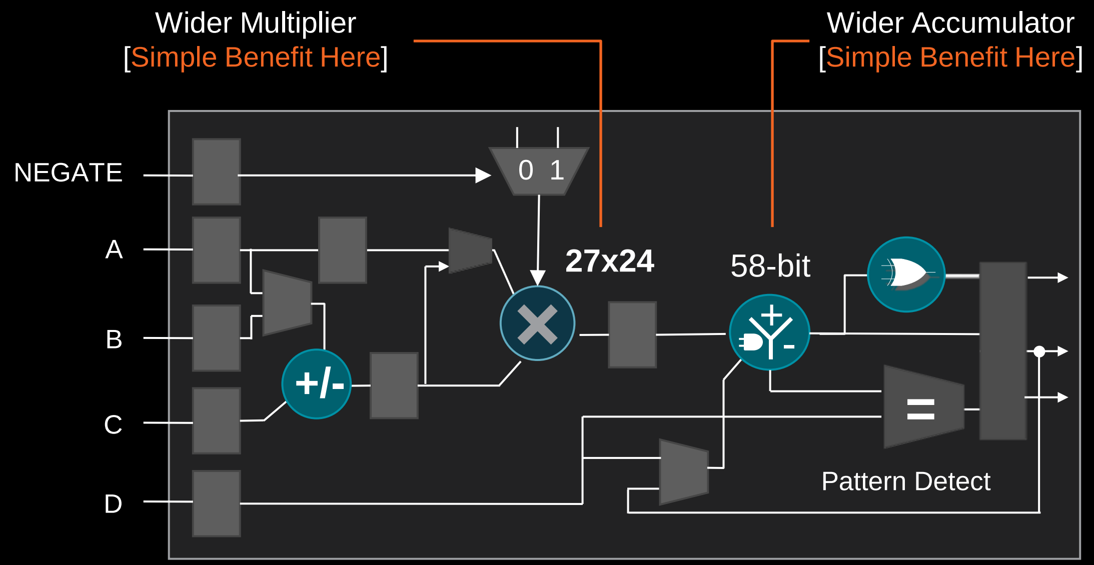
AMD

together we advance_

DSP Engines: Tunable for Precision, Accuracy, and Lower Power

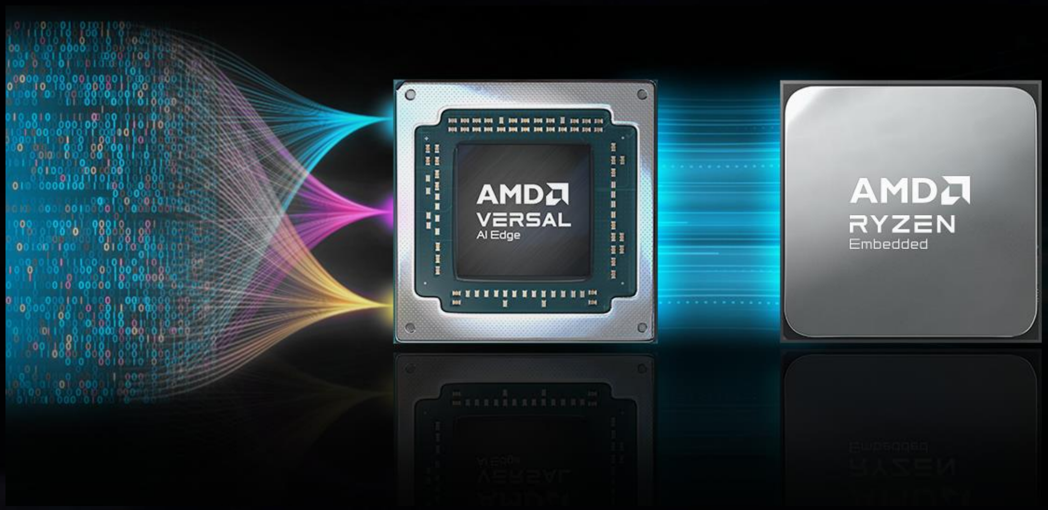
- **Enhanced compute architecture (DSP58)**
  - 1GHz performance (1.3X vs. UltraScale+™)
  - Variable precision fixed- and floating-point
  - Up to 70% power reduction vs. previous generation
- **Versatile data-type support for diverse applications**
  - Integrated 32- and 16-bit floating-point (e.g., HPC)
  - Integrated complex 18x18 (e.g., sensor processing)
  - 2.2X INT8 operation vs. previous gen (AI inference)
- **Code portability from previous generation**
  - Supports existing IP and LogiCore libraries
  - Compatible w/Model Composer, HLS, RTL import flows[1]

Enhancements per Block
(1GHz Performance)



1: Floating-point format requires instantiation of floating-point IP core (e.g., via IP integrator) vs. RTL inference

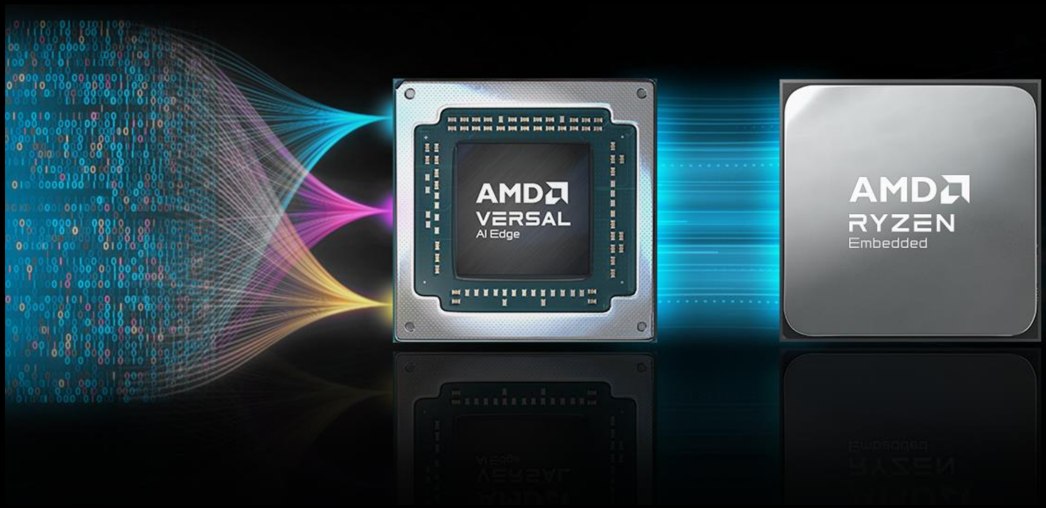
Embedded+



Wat is Embedded+ (Embedded Alveo?)

Embedded+ integrates AMD Ryzen™ Embedded processors with AMD Versal™ AI Edge adaptive SoCs on a single PCB

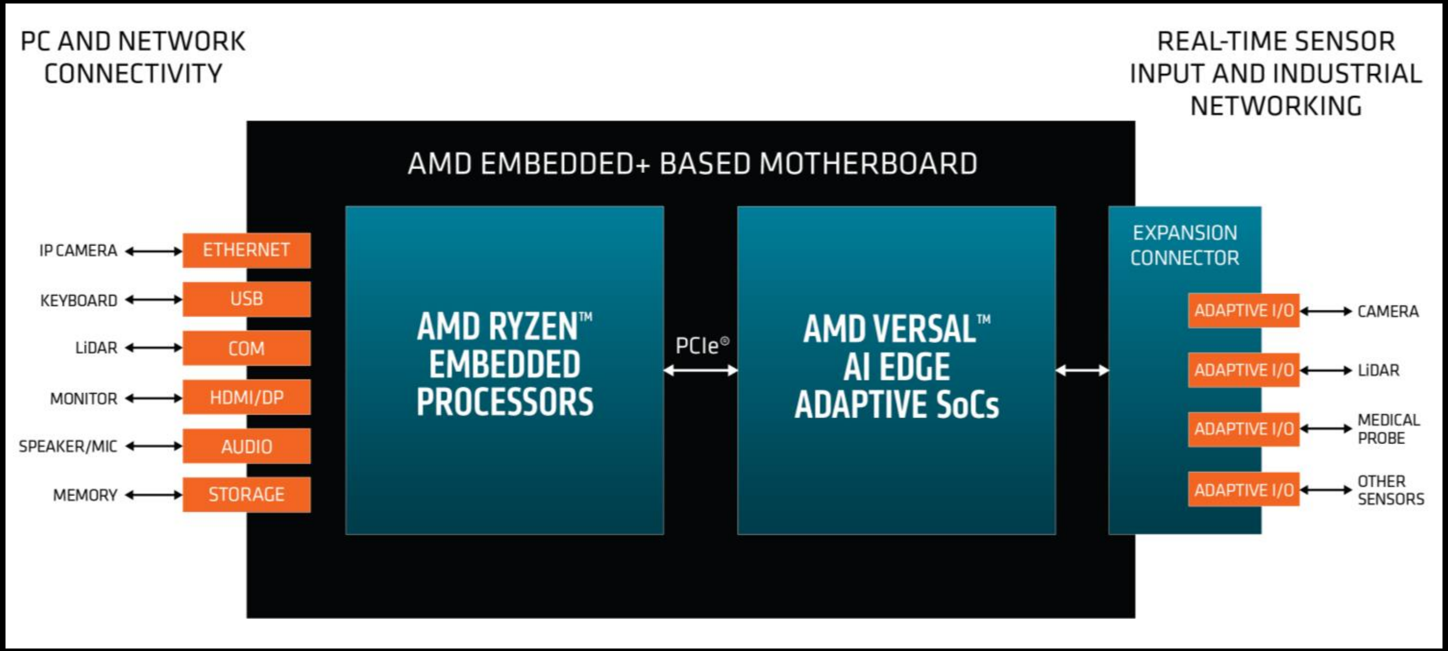
AMD makes the path to sensor fusion, AI inferencing, industrial networking, control, and visualization simpler with the Embedded+ architecture and ODM partner products

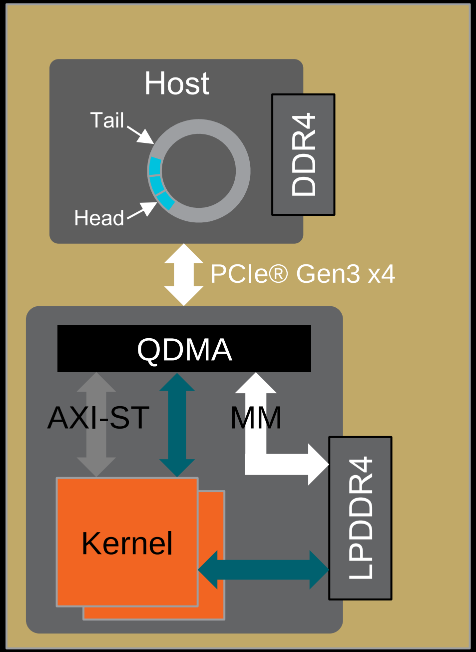


Sapphire Technology VPR-4616-MB

Embedded+ What makes something an Embedded+ board

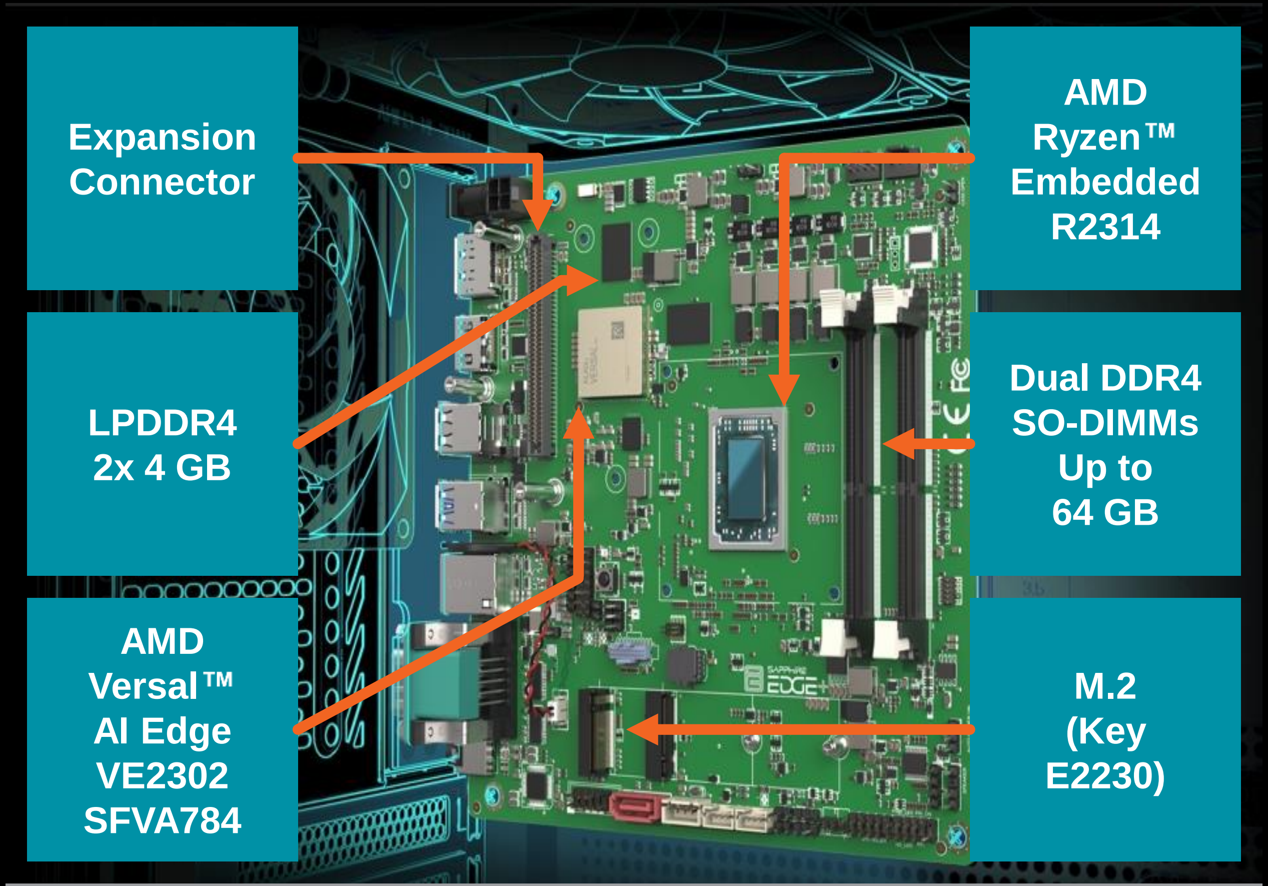
- Ryzen Embedded CPU
- Versal ACAP (connected by PCIe)
- Specified Expansion connector
- Base design based on XRT





I/O Type	Description	Device Available I/O	Routing
GTYP	High-speed transceivers	4 lanes	Differential pairs
XPIO	1.0-1.5V PL connected I/O	56	Differential pairs
HDIO	1.8-3.3V PL connected I/O	24	Single ended

EMBEDDED+ SOLUTION COMPONENTS



SAPPHIRE EDGE+ VPR-4616-MB

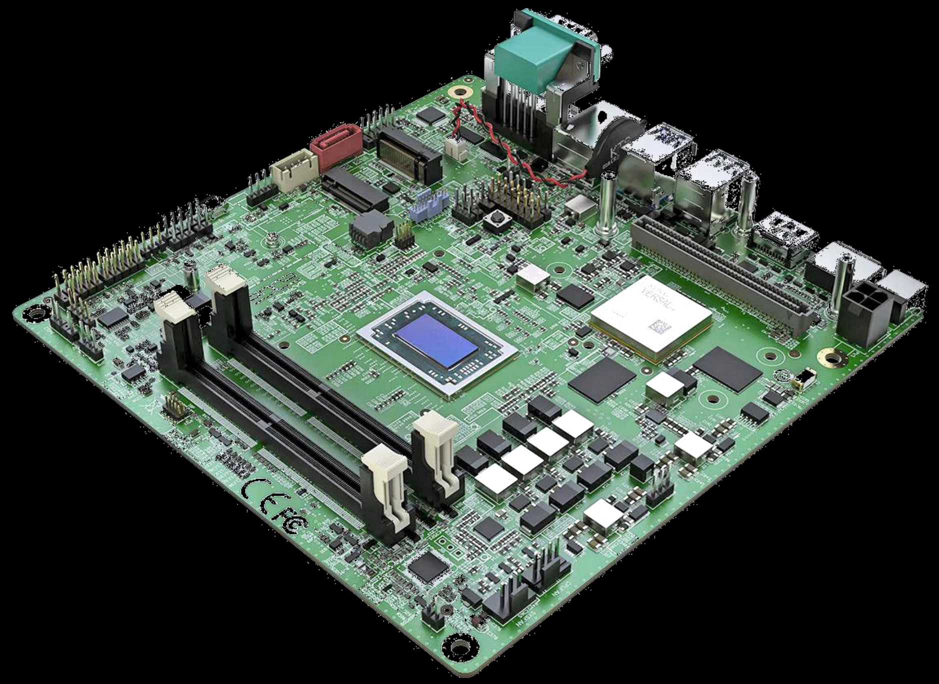
Software Infrastructure

- AMD Ryzen™ Embedded processors and Versal™ adaptive SoCs connected via PCIe® interface
- Inter-chip communication powered by XRT
- AMD Vitis™ AI and VVAS support (coming soon)

Example Design Roadmap

- 2D Filter (https://github.com/Xilinx/emb-plus-examples)
- Sensor fusion with AI inference
- AI-ML inferencing on video stream
  - Expansion connector source
  - Video decoder source
- AMR: 8x GMSL + LiDAR + GPS + IMU + WiFi
- TSN and other industrial Ethernet standards
- Machine vision frame grabber over 10 GE

ODM PRE-INTEGRATION AND PARTNERS (AVAILABLE NOW)



VPR-4616-MB platform features:

- Mini-ITX form factor (170 mm x 170 mm)
- AMD Versal™ AI Edge 2302 device
- AMD Ryzen™ Embedded R2314 processor
- Custom expansion connector for I/O boards
- Dual DDR4 SO-DIMM with 64 GB max capacity
- 1x M.2 (Key M 2580) with PCIe® Gen3 x4 and SATA for SSD
- 1x SATA3
- 1x 2.5 Gb Ethernet on motherboard
- 1x M.2 (Key E 2230) with PCIe x1 and USB2.0 for wireless / BT
- Dual displays – 1x HDMI plus 1x DP
- Discrete audio in/out
- 2x USB 3.2 Type A, 2x USB 2.0 Type A , 1x USB 3.2 Type C
- 1x RS232 / 422 / 485
- 12-19VDC
- OS support: Linux® Ubuntu® 22.04
- System version available: VPR-4616-SYS

SAPPHIRE Technology is a world-leading manufacturer and global supplier of innovative graphics, embedded, and GPU compute server solutions for Commercial and Consumer markets

AMD
together we advance_