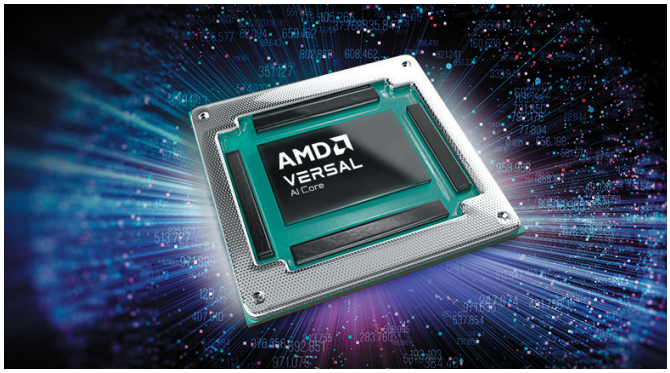
SOLUTION BRIEF

AI Inference with Versal™ AI Core Series

- 2.7X performance/watt vs. competing FPGAs[1] for cloud acceleration
- Accelerates the whole application from pre- to post-processing
- Adaptable to evolving AI Algorithms

CHALLENGE

Applied machine learning techniques have now become pervasive across a wide range of applications, with tremendous growth in vision and video in particular. FPGA-based AI/ML acceleration has already shown performance and latency advantages over GPU accelerators, but next-generation CNN-based workloads demand compute density beyond what traditional FPGA programmable logic and multipliers can offer. Fabric-based DSP blocks offer flexible precision and are still capable accelerators, but the bit-level interconnect and fine-grained programmability come with overhead that limits scalability for the most compute-intensive CNN-based workloads.



SOLUTION: VERSAL AI CORE SERIES FOR AI COMPUTE ACCELERATION

The AMD Versal™ AI Core series is a highly integrated, multicore, heterogeneous device that can dynamically adapt at the hardware and software level for a wide range of AI workloads, making it ideal for cloud accelerator cards. The platform integrates next-generation Scalar Engines for embedded compute, Adaptable Engines for hardware flexibility, and Intelligent Engines consisting of DSP Engines and revolutionary AI Engines for inference and signal processing. The result is an adaptable accelerator that exceeds the performance, latency, and power efficiency of traditional FPGAs and GPUs for AI/ML workloads.

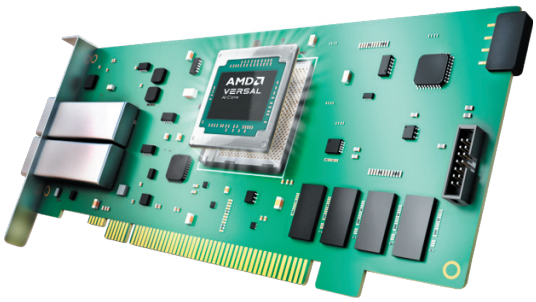
AI Engines for Breakthrough AI/ML Inference

Within the Versal platform is a unique architecture for AI inference—the AI Engines—which are an array of software programmable vector processors with flexible interconnect and tightly coupled local memory—ideal for CNN-based inference and delivering 2.7X performance/watt over competing 10nm FPGAs.[1] AI Engines deliver compute density, power efficiency, and low latency not possible with GPUs and traditional FPGA architectures, all while retaining hardware adaptability to evolve with AI algorithms.

2.7X

Performance/Watt vs. FPGAs[1]

Versal AI Core device for
AI Accelerator Cards



Whole Application Acceleration

Machine learning is typically integrated into a larger application rather than a stand-alone workload. As a complete heterogeneous compute platform, the Versal AI Core series leverages its diverse engines to infuse deep learning as “an element” of a larger application that has other pre/post-processing requirements, delivering end-to-end application acceleration.

Complete Development Environment for SW Developers

Fully supported by the Vitis™ AI development environment, which consists of optimized IP, tools, libraries, models, and example designs, Versal AI Core devices enable data scientists to compile TensorFlow, PyTorch, and Caffe models using Python or C++ APIs in minutes—without prior FPGA knowledge.

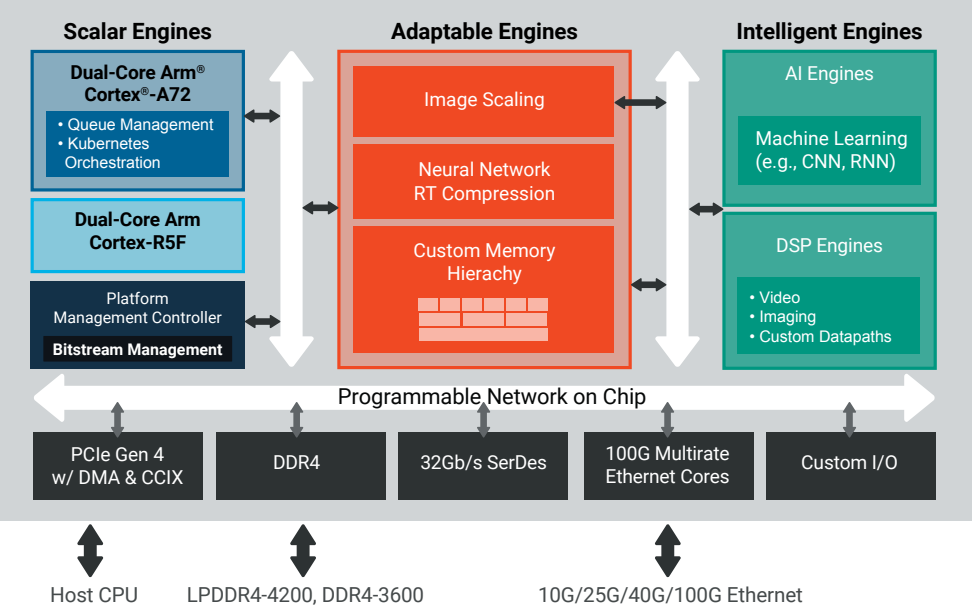
1: Versal AI Core VC1902 device vs. Intel Agilex AGF027 FPGA, ResNet50 v1.5 (frames-per-second/watt)

VERSAL ADAPTIVE SOC IMPLEMENTATION

AI Compute Accelerator with Versal AI Core Series

The Versal AI Core series solves the unique and most difficult problem of AI inference—compute efficiency—by coupling ASIC-class compute engines (AI Engines) together with flexible fabric (Adaptable Engines) to build accelerators with maximum efficiency for any given network, while delivering low power and low latency. Through its integrated shell—enabled by a programmable network on chip and hardened interfaces—Versal SoCs are built from the ground up to ensure streamlined connectivity to data center compute infrastructure, simplifying accelerator card development.

VERSAL AI CORE VC1902 IMPLEMENTATION



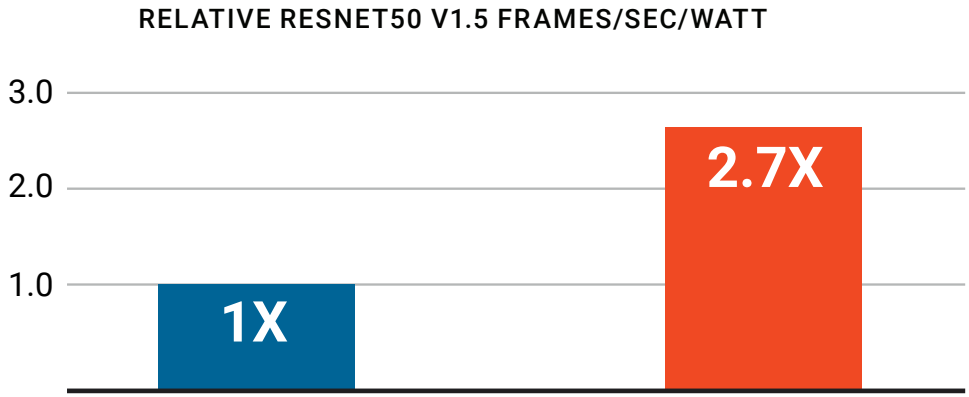
PLATFORM HIGHLIGHTS	
Adaptable Engines	> Custom memory hierarchy optimizes data movement and management for accelerator kernels > Pre- and post-processing functions including neural network RT compression and image scaling
AI Engines	> Tiled array of vector processors, flexible interconnect, and local memory enabling massive parallelism > Up to 133 INT8 TOPS with the Versal AI Core VC1902 device, scales up to 405 INT4 TOPS in the portfolio > Compiles models in minutes based on TensorFlow, PyTorch, and Caffe using Python or C++ APIs > Ideal for neural networks ranging from CNN, RNN, and MLP; hardware adaptable to optimize for evolving algorithms
Scalar Engines	> Arm processing subsystem for queue management and Kubernetes orchestration > Platform management controller for security, power management, and bitstream management
Programmable Network on Chip (NoC)	> Seamlessly integrates all engines and key interfaces > Simplifies kernel and IP placement, reducing soft logic needed for connectivity > Streamlines programming experience for software and hardware developers
Integrated Shell	> Comprises hardened host interface, programmable NoC, and Scalar Engines > Ensures streamlined device bring-up and connectivity to off-chip interfaces, making the platform available at boot > Delivers pre-engineered timing closure and logic resource savings, simplifying development of accelerator cards

AMD
VERSAL

BENCHMARK

ResNet50 v1.5 Performance Comparison

Shown below is a comparison of measured results on Versal devices as submitted to the ML Perf Data Center v1.0, and projected performance of competing 10nm Intel Agilex FPGAs.



	Intel Agilex AGF027-2		Versal AI Core VC1902[1]
Peak INT8 TOPS	61 TOPs[2]	2.2X	133 TOPS
Logic Shell	117k ALMs	62% Less	45k LUTs[3]
Frames/Sec	2,246	2.9X	6,471
Total Power	92 Watts[4]	6% Less	87 Watts[5]
Frames/Sec/Watt	**24 FPS/Watt**	2.7X	**66 FPS/Watt**

1: Measured results of VCK5000 development card based on Versal AI Core VC1902

2: Assumes 30% compute efficiency for Intel Agilex FPGA 18x19 multipliers and 40% compute efficiency of AI Engines

3: Integrated shell reduces logic required for connectivity, 45K LUTs required for run-time SW & deep-learning processor support

4: Based on Quartus Power & Thermal Calculator 2021.2, assumes SmartVID and claimed static power savings

5: Device power estimates, based on Xilinx Power Estimator (XPE) available at https://www.xilinx.com/products/technology/power/xpe.html

TAKE THE NEXT STEP

- For more information on Versal AI Core series, visit www.xilinx.com/versal-ai-core
- To try the above benchmark yourself, visit www.xilinx.com/versal-performance-elevated
- To start designing for cloud acceleration and edge computing, visit www.xilinx.com/vck5000
- To start designing on a Versal AI Core Evaluation Kit, visit www.xilinx.com/vck190
- To contact your local AMD sales representative, visit Contact Sales

AMD VCK5000 Versal Development Card
www.xilinx.com/vck5000