Xilinx Versal™ AI Core

The Xilinx® Versal™ AI Core series delivers breakthrough artificial intelligence (AI) inference acceleration with integrated AI engines that deliver over 100X greater compute performance than today's server-class CPUs**.** Versal AI Core series devices are optimized for compute-intensive applications, offering the Versal portfolio's highest compute for maximum AI and machine learning (ML) workload acceleration.

Versal introduces a new class of hardware adaptive SoCs to the industry: the adaptive compute acceleration platform (ACAP). It is a multi-core, heterogeneous compute platform that dynamically adapts at the hardware and software levels to target a wide range of applications and AI workloads.

Versal ACAPs are conceived to be natively programmable by software developers and data scientists, as well as hardware developers. Any developer can now harness the capabilities of adaptable hardware with the latest innovations in AI acceleration to create custom architectures optimized for cloud, 5G wireless, and edge applications.

Versal ACAP devices counter the slowing of Moore's Law and enable the proliferation of AI across industries. They combine next-generation scalar engines for embedded compute, adaptable engines for hardware programmability, and intelligent AI engines and digital signal processor (DSP) engines for AI inference and advanced signal processing. With leading-edge memory and interfacing technologies also integrated, Versal delivers powerful heterogeneous acceleration for any application.

The Versal AI Core series has five devices for scalability across end-applications, offering 128 to 400 AI Engines. The series includes dual-core Arm Cortex™-A72 application processors, dual-core Arm Cortex-R5 real-time processors, more than 1,900 DSP engines optimized for high-precision floating point with low latency, and 1.9 million system logic cells of adaptable engines for hardware acceleration and differentiation. For high-speed compute and network interfacing, the series also contains PCIe Gen4, and CCIX host interfaces, power-optimized 32G transceivers, up to 4 integrated DDR4 memory controllers, and up to 4 multi-rate Ethernet MACs. All of this is interconnected by a state-of-the-art network-on-chip (NoC) for multi-terabit per-second bandwidth with native software programmability.

Versal AI Core devices boast the Versal portfolio's highest compute and lowest latency, enabling breakthrough AI inference throughput and performance. They are optimized for compute-intensive applications including compute acceleration for the data centre, 5G radio and wireless beamforming, radar applications, and automated driving technology—delivering unprecedented signal processing performance, workload acceleration, and diverse forms of AI for the cloud, network and edge on a single platform.

Innovation of AI Engines

The Versal ACAP AI Engine array introduces a unique architecture to overcome the greatest challenge of AI inference: compute efficiency. Hundreds of flexible ASIC-class VLIW and SIMD processing engines are tightly interconnected through a programmable fabric and feature an efficient memory structure that enables all processors to run continuously at full rate up to 130 TOPS @ INT8 at low latency. This allows specialized hardware customization (domain-specific architectures) to achieve maximum

efficiency for any given neural network type without re-spinning the chip. Neural network architectures are changing quickly and need this rapidly adaptable hardware.

The AI Engine delivers five times the silicon compute density at half the power consumption of traditional FPGAs and enables low-latency Convolutional Neural Network (CNN) execution resulting in 1.5X greater AI-inference performance compared to industry-leading GPUs[1].

Hardware Adaptability and "Whole Application Acceleration"

Beyond performance, Versal AI Core's adaptability and heterogeneous architecture deliver advantages over other solutions that focus on a subset of AI workloads. AI algorithms are diverse across industries and evolve rapidly, even within a single domain. This fast pace of innovation outpaces silicon design cycles. Fixed-architecture ASSP and ASIC chips inevitably arrive in the market behind the evolutionary curve of AI networks, and hence contain a compute engine and memory hierarchy that is sub-optimal to the latest algorithms, resulting in much lower compute efficiency. Versal adapts the architecture to the algorithm to address both the compute density and memory bandwidth needed as algorithms evolve.

Furthermore, ML is rarely a stand-alone workload; it's typically integrated into a larger application. As a complete heterogeneous compute platform, Versal AI Core devices infuse deep learning as "an element" of a larger application that has other pre/post-processing requirements. This is what makes Versal ACAPs so uniquely powerful: its integration of diverse processing engines—hardware adaptable and software programmable—to accelerate the whole application.

Testimonials

In April 2020, Xilinx and Samsung announced a partnership where Samsung will utilize the Xilinx Versal ACAP for worldwide 5G commercial deployments. 5G is a new technology that is still developing its infrastructure requirements, and industry specifications will continue to advance. Xilinx is the leader in adaptive computing technology and can provide the means to adapt to the evolving 5G technology. Versal ACAPs are providing a universal, flexible, and scalable platform that can address multiple operator requirements across various geographies.

"Samsung has been working closely with Xilinx, paving the way for enhancing our 5G technical leadership and opening up a new era in 5G," said Jaeho Jeon, executive vice president and head of R&D, Networks business, Samsung Electronics. "Taking a step further by applying Xilinx's new advanced platform to our solutions, we expect to increase 5G performance and accelerate our leadership position in the global market."

"Samsung is a trailblazer when it comes to 5G innovation and we are excited to play an essential role in its 5G commercial deployments," said Liam Madden, executive vice president and general manager, Wired and Wireless Group, Xilinx. "Versal ACAPs will provide Samsung with the superior signal processing performance and adaptability needed to deliver an exceptional 5G experience to its customers now and into the future."

[1] CNN Performance Img/Sec (ResNet50, batch=1) vs. Nvidia Tesla T4 GPUs. Based on T4 performance numbers published in "NVIDIA Tesla Deep Learning Product Performance" at www.developer.nvidia.com