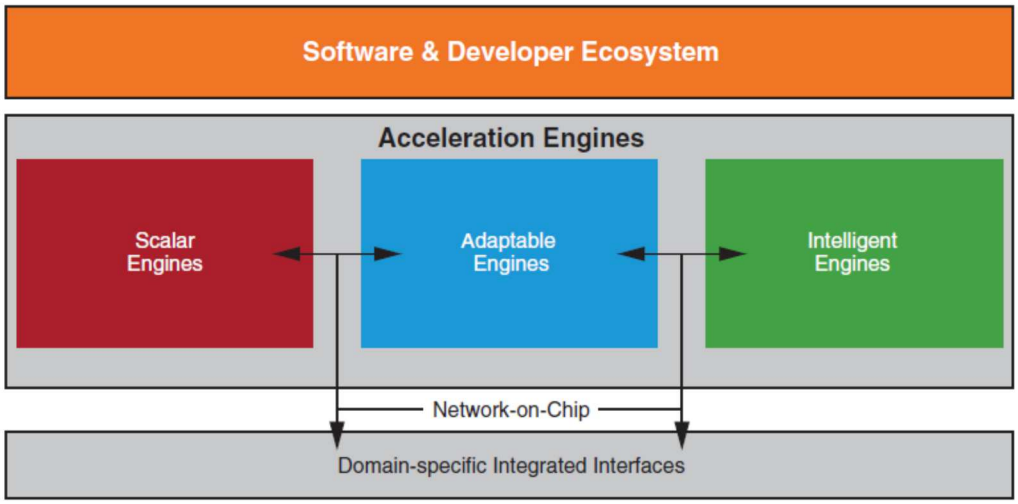
Electronics Industry Awards 2019 Entry

Xilinx® Versal™ Whole-Application Acceleration for Next-Generation Cloud and Embedded Computing

Xilinx® Versal™ devices are the first in a new class of processors called Adaptive Compute Acceleration Platform (ACAP) and address the three defining trends in computing today: the explosion of data; the dawn of artificial intelligence (AI); and the decline of Moore's Law.

Responding to the challenges presented by new, computationally intense workloads in cloud and embedded use-cases, the Versal concept enables whole-application acceleration leveraging a heterogeneous array of Adaptable Hardware Engines, Scalar Engines, and Intelligent Engines containing software-programmable vector cores, combined with leading-edge memory and interfacing technologies. These resources are tightly coupled using a Network On Chip (NOC) multi-terabit superhighway that gives designers the freedom to apply the right processing engine for the right workload.

Just as important as this high-performance architecture, Versal is designed to enable all types of developers – software engineers, hardware engineers, data scientists - to accelerate their whole application with optimized hardware and software. Moreover, they can instantly adapt both to keep pace with rapidly evolving technology.



Versal ACAP conceptual diagram.

Market Impact

These devices will enable the performance improvements needed by applications such as video transcoding, database, data compression, search, AI inference, genomics, machine vision, computational storage and network acceleration, which are deployed in diverse markets from cloud, to networking, automotive, wireless communications, edge computing, and endpoints.

Any-Developer Accessibility

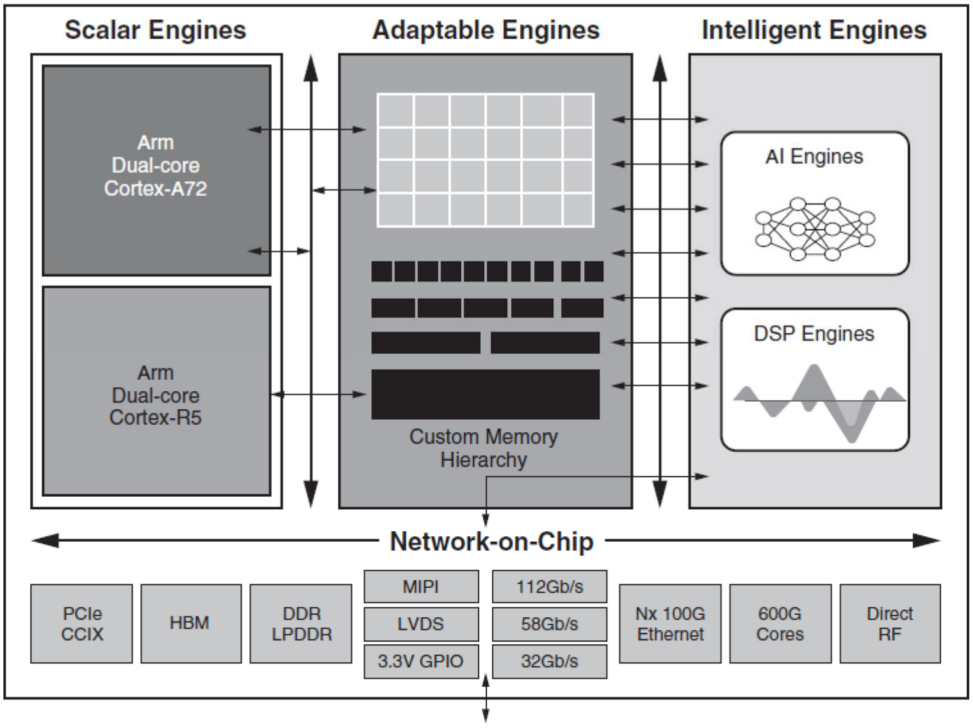
Critical to the success of Versal is the ability for device-level programming using familiar languages such as C/C++. To enable this, Xilinx has created a new and complete software development stack including a programming toolchain, performance-optimised software libraries and run-time software. AI developers using machine-learning frameworks such as TensorFlow, Caffe or MXNet can also compile for Versal devices.

Innovative Intelligent Engine Containing Advanced DSP Engines and AI Engines

Within the Versal family there are further opportunities to create devices that are optimized for particular types of applications and markets. Xilinx has initially introduced two series of Versal devices: Versal Prime and Versal AI Core.

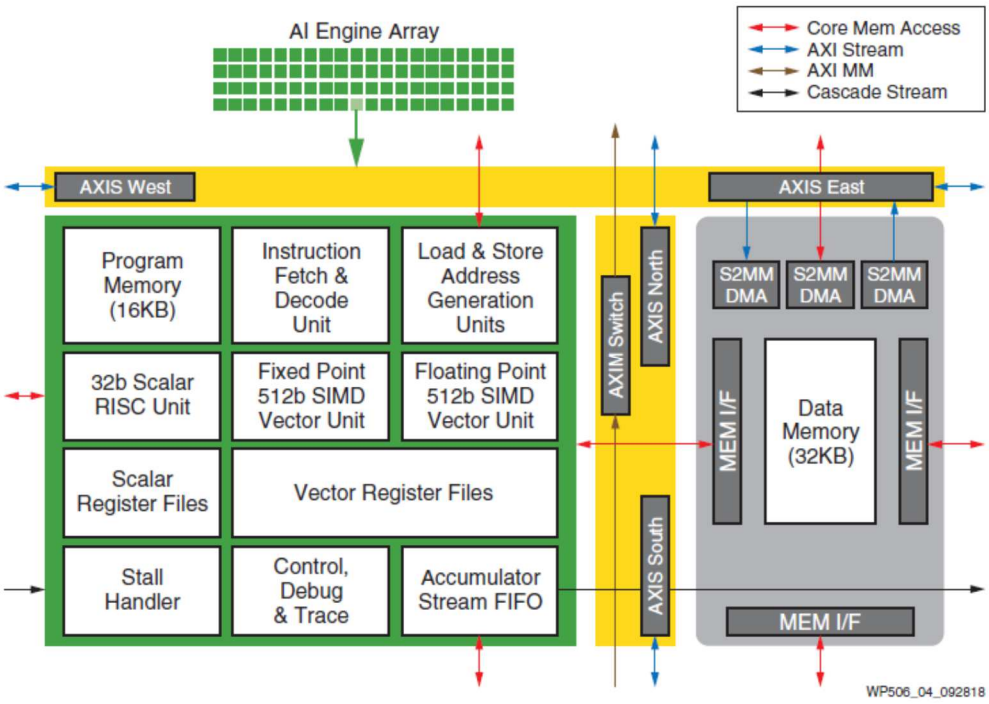
The Versal Prime series targets broad applicability across multiple markets and is optimized for connectivity and in-line acceleration of a diverse set of workloads. The Intelligent Engines of these devices contain large numbers of advanced DSP Engines, which are optimised for low-latency and high-precision floating-point workloads. There are 472 DSP Engines in the VM1102, rising to 3080 DSP Engines in the largest device in the series, the VM2902.

Versal AI Core devices are optimised for unprecedented AI inference throughput and performance. To achieve this, the Intelligent Engine block features a heterogeneous architecture that combines DSP Engines with Xilinx's innovative AI Engine. Devices range from the VC1352 containing 128 AI Engines with 32Mb memory and 928 DSP Engines, to the VC1902 with 400 AI Engines, 100Mb memory, and 1968 DSP Engines.



Versal AI Core series, featuring Intelligent Engines containing AI Engines and DSP Engines.

Xilinx's AI Engine can be visualised as an "FPGA of processors", comprising hundreds of highly flexible VLIW SIMD processing engines tightly integrated with programmable fabric, capable of achieving up to 140 TOPS @ INT8. Each processor in the array includes its own instruction and data memory, which neighbouring processors can also access directly. This memory architecture facilitates movement of data from processor to processor without the need for a large shared L2 cache, since each AI Engine has access to its neighbours' memory. Hence AI Engine processors can run at full rate without being starved by shared L2 cache structures or cache misses.



Details of the AI Engine: Part of the Versal Intelligent Engine.

Xilinx also plans to introduce the Versal Premium and Versal HBM series, as well as Versal AI Edge, and Versal AI RF devices, featuring further variations of the Intelligent Engine.

Versal Scalar Engines include a dual-core Arm® Cortex®-A72 application processor with 48KB/32KB L1 cache and 1MB L2 cache, as well as a dual-core Cortex-R5 real-time processor with 32KB/32KB L1 cache and tightly-coupled memory with ECC, and 256KB of on-chip memory with ECC.

The Adaptable Hardware Engines are comparable to traditional FPGA logic fabric, with up to two million system logic cells, up to 900,000 Look-Up Tables (LUTs), and generously distributed RAM, block RAM, and UltraRAM.

Performance

To give an indication of these devices' acceleration potential, the AI Engine delivers eight times the silicon compute density at half power consumption of traditional programmable logic solutions. Versal AI Core devices have demonstrated low-latency Convolutional Neural Network (CNN) execution

equivalent to about eight-times better AI-inference performance than industry-leading GPUs. For 5G communications, simulations show a four-fold bandwidth improvement over earlier architectures when implementing massive MIMO radio.

Conclusion

Xilinx Versal ACAPs are the first to combine heterogeneous domain-specific engines with programmable hardware and software, realising the whole-application acceleration needed to achieve low-latency performance and high energy efficiency that conventional CPUs and GPUs are unable to satisfy. At the same time, hardware and software adaptability delivers flexibility unmatched by ASICs or ASSPs, to keep pace with rapid technological advancements in neural networks and artificial intelligence.

By making such advanced performance and adaptability accessible to all types of developers, Versal ACAPs solve the major challenges facing embedded and cloud computing today.

Ends