WP518 (v1.0)

ACAP at the Edge with the Versal AI Edge Series

Versal™ AI Edge ACAPs are a series of devices targeted at edge computing. These devices allow engineers to achieve high performance per watt and low latency, while meeting critical environmental and safety certifications.

ABSTRACT

This white paper introduces the AI Edge series to the Versal™ ACAP portfolio, a domain-specific architecture (DSA) that meets the strenuous demands of systems implemented in the 7nm silicon process. This series is optimized to meet the performance-per-watt requirements of edge nodes at or near the analog-digital boundary. Here, immediate response to the physical world is highly valued, and in many market spaces is required.

The AI Edge series enables the Versal ACAP footprint to spread exponentially from core networks and data centers to the edge. Along with heterogeneous compute efficiency, the hallmarks of the AI Edge series are:

- Cacheless memory hierarchy (low latency and determinism)
- Hardened yet programmable infrastructure (high throughput, low power)
- Functional safety and security (trusted solutions enabled)
- Flexibility and programmability (solution can be differentiated or upgraded during development or even after deployment)

Versal AI Edge Series Architectural Constructs

The AI Edge series was developed to target commercial, industrial, and defense applications at the edge where balancing performance and power consumption, low latency, size and thermal constraints, and safety and reliability are paramount. As such, there are five main architectural constructs that form the basis for the devices in the AI Edge series to meet those requirements. Those constructs are:

1. Heterogeneous processing

2. Hardened infrastructure, I/O

3. Cacheless memory hierarchy

4. Functional safety and security

5. Flexibility and programmability for differentiation

Heterogeneous processing is evident in the composition of Scalar Engines, Adaptable Engines, and Intelligent Engines. Multiple engine types are required because no single processor type is capable of optimally performing all tasks required for an application. Scalar Engines are ideal for complex decision making and control. Adaptable Engines add flexibility to handle a diverse set of computationally demanding algorithms. Intelligent Engines are optimized for advanced signal processing, such as linear algebra and matrix math, which are well suited for 5G wireless systems and AI inference. The AI Edge series also introduces optimized Intelligent Engines for Machine Learning inference applications.

Having a variety of processing engines is only good as long as they are processing data and not stalled, so managing I/O bottlenecks is critical in any system. However, unlike an ASSP or ASIC, which are fixed function, ACAPs support a plethora of different algorithms and data flows. At the same time, power consumption is critical for the AI Edge series, so building infrastructure out of Adaptable Engines was not optimal. These requirements necessitated a completely different approach to the problem.

One of the keys was to develop a hardened network on chip (NoC) that was also programmable. The programmable NoC, unlike a typical NoC in an ASSP or ASIC, can be programmed to support different data flows between the high-speed I/Os and the various engines, as well as different throughputs, latencies, or even bit widths in the event of multiple algorithms on a single chip being optimized for different resolutions. This "secret weapon" is a major differentiator between ACAPs and FPGAs. It is also worth noting that there are alternative ways to route data through the Adaptable Engines, if necessary, for latency or determinism requirements. Xilinx has also hardened other key portions of the infrastructure such as memory controllers, PCIe®, and Ethernet MACs.

Similar to I/O bottlenecks, memory access is a crucial consideration for system development so the system does not stall. For this reason, as well as ensuring AI Edge-based systems can be low latency and deterministic for real-time performance, Xilinx instantiated a cacheless-memory hierarchy, illustrated in Figure 1. Unlike most ASSPs and ASICs, which have multiple levels of non-deterministic caches, all levels of the hierarchy are essentially shared RAM. The accelerator RAM is new to the AI Edge series with the intent for most datasets to be able to be stored on-chip to reduce power consumption and latency from off-chip memory accesses.
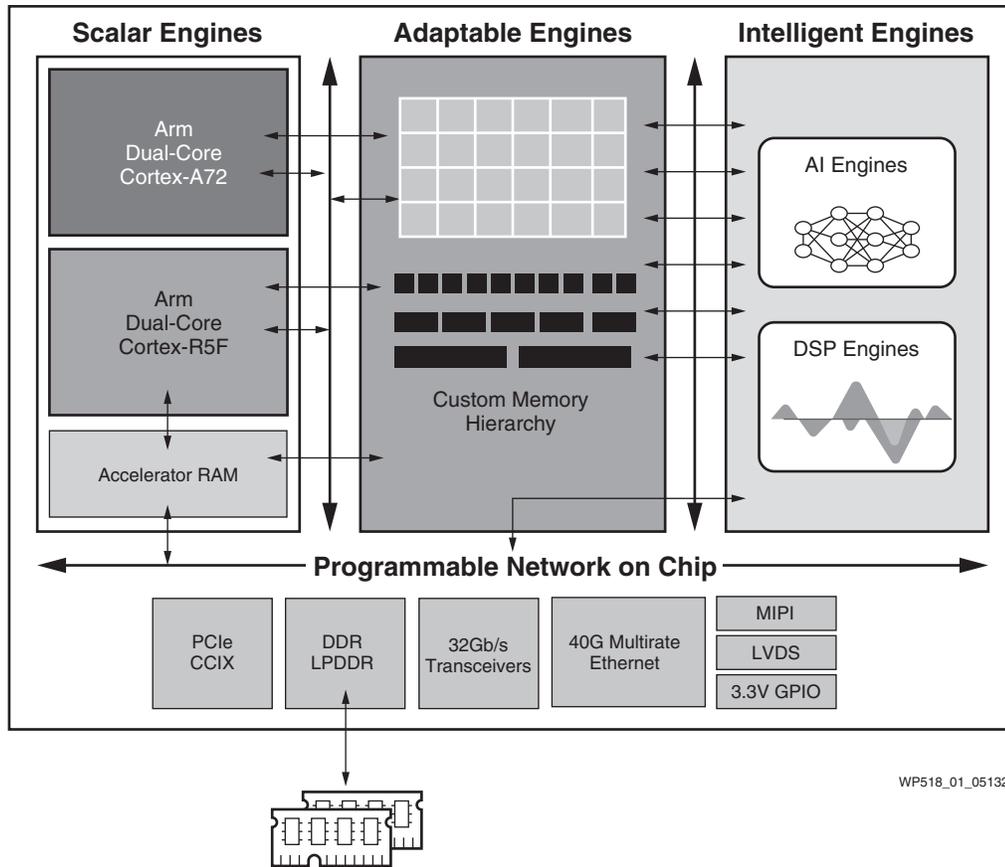
Figure 1: **Versal AI Edge Series Cacheless Memory Hierarchy**

Since many edge nodes or endpoints, such as automobiles, robots, and UAVs, have stringent safety, security, and quality requirements and certifications, the AI Edge series was designed to support the development of such systems needing to adhere to standards such as ISO 26262, IEC 61508, DO-178C, and DO-254. Implementation aspects such as processors running in lock-step, and the usage of redundancy throughout the devices, ensure the AI Edge ACAPs are capable of supporting systems with the highest levels of functional safety. In addition, ACAPs have the ability to create a hardware root of trust and attack protection for enhanced security.

The fifth architectural construct, flexibility and programmability for customer differentiation, has been the hallmark of Xilinx since its inception. However, with the Versal ACAP portfolio, Xilinx has taken the concept further than ever before with the goal that developers program with the tools and languages with which they are already familiar. The programming methodology for ACAPs provides multiple options, including some that are designed for AI and/or software developers who are not familiar with VHDL or Verilog. In those cases, AI developers can use machine learning (ML) frameworks such as TensorFlow or PyTorch and target their C++ algorithms and Python data flows to the Versal ACAP without using traditional tools needed to write RTL. In these cases, no place and route is required, just software compilation, resulting in a highly iterative and efficient development process. Embedded software developers can program in C using the Vitis™ software platform, and hardware engineers can continue to use Vivado® tools to program in VHDL or Verilog.

The combination of these architectural constructs differentiate Versal ACAP and the AI Edge series from ASSPs, ASICs, and GPUs.

Versal AI Edge Series Design Concepts

The AI Edge series architecture has been developed to meet stringent design goals around compute performance, data throughput, latency, determinism, power consumption/thermal constraints, scalability, functional safety, and security that are typical for applications related to the automotive, aerospace & defense, industrial, and healthcare IoT industries.

Designing for Compute Performance

The AI Edge series comprises Scalar, Adaptive, and Intelligent Engines so the correct processor can be targeted at the correct task.

Scalar Engines consist of a low-power domain and a full-power domain. The low-power domain provides real-time processing using two Arm® Cortex®-R5F processors. The processors can operate in lock-step in applications where the functional safety requirements are targeting a higher ASIL level. The processors come with tightly coupled memory for real-time task handling. They also have on-chip (SRAM) memory (OCM).

The full power domain is used where higher performance processing is needed along with memory management unit (MMU) capabilities. There are two Cortex-A72 cores along with a system memory management unit (SMMU).

Scalar Engines communicate through a low-latency NoC path to DDR. Scalar Engines also include dedicated communication components, such as PCIe, 1Gb Ethernet, and CAN-FD.

Adaptable Engines provide the programmable logic that ties all of the elements together and allows dedicated customized blocks to be created. Adaptable Engines include block RAM and UltraRAM that allow a customized memory hierarchy to be created. This is useful, for example, in implementing ML frameworks. The memory hierarchy and data movement can be precisely adapted to the network. The Adaptable Engines are also tightly coupled with programmable I/O, so Versal ACAPs can integrate to multiple sensors and interfaces, such as radar, LiDAR, GPS, and vision sensors.

Intelligent Engines comprise the AI Engines as well as DSP Engines. The AI Engines are a tiled array of VLIW and SIMD processing elements that provide high compute density for vector-based algorithms. The AI Engine tiles include vector processors for fixed and floating-point operations, and a scalar processor, as well as dedicated program and data memories. The array also has AXI data movement channels and support for DMA and locks. AI Engines execute 6-way VLIW instructions that are a mix of scalar and vector loads and stores on each cycle. The AI Engine array has a dedicated inter-tile communication network, as well as being in tightly coupled communication with the Adaptable Engines. This allows applications to seamlessly pipeline tasks and provide a complete solution with high compute performance, data movement and manipulation, and the right memory hierarchy for the task. For example, as self-driving car technology becomes more prominent, multiple neural networks might need to be chained together in sequence to perform complex tasks, exacerbating the issues with GPU implementations that are

reliant on high batch sizes. Therefore, Xilinx has optimized the AI Edge series to operate at extremely high efficiency at low batch sizes.

The AI Edge series introduces the AI Engine for Machine Learning (AIE-ML), which has been optimized for ML inference applications. Architectural enhancements have been made to the AIE-ML tiles to significantly increase TOPS/$, TOPS/watt, and throughput for AI and ML inference workloads like CNNs, RNNs, and MLPs. The memory hierarchy has been improved as well with the introduction of on-chip shared memory, called memory tiles, within the AIE-ML array. These memory tiles significantly reduce the need for programmable logic resources for ML applications while improving overall bandwidth and memory capacity.

DSP Engines provide high compute density and flexibility with support of 58-bit integer operations, and single precision floating point. They are suited for all signal processing tasks such as image conditioning, control systems, and medical and wireless applications, and they are tightly coupled with the Adaptable Engines to build customized accelerators that are exactly matched to the task.

Designing for Data Throughput, Latency, and Determinism

Real-time systems at the edge often need to be low latency and deterministic while processing significant amounts of data from multiple sensors. Latency is an especially critical processing performance factor when considering real-time systems, such as vehicles traveling at automotive speeds. At 60mph, the difference of a few handfuls of milliseconds in reaction time of different ADAS systems can have a significant impact on a system's effectiveness. The AI Edge series provides a number of features that enable this, including fast boot, the programmable NoC, hardened memory controllers, and I/O.

The platform management controller (PMC) manages the boot and setup of the device as well as monitoring and reporting while the device is in operation. Processing within the PMC is handled by two triple-mode-redundant (TMR) microcontrollers. The PMC performs power management, security, monitoring (voltage/temperature), and error reporting.

The programmable NoC provides unified memory-mapped access to all of the resources across the device. The NoC scales across the device family and is designed to support all of the DDR memory data movement needs as well as intra-device high-bandwidth data movement. The NoC is a packet switch network that performs high-speed point-to-point hops to move data around the device. Even without exploiting the inherent parallelism of the NoC, it is capable of moving greater than 2Tb/sec of data—and much more when the design mapping exploits parallelism. The routing within the Adaptable Engines can also provide high-bandwidth, low-latency data movement, as well as determinism when coupled to the internal cacheless-memory hierarchy.

All Versal devices have high performance I/O (XP I/O) paired with hardened DDR4/LPPDR4 memory controllers. The I/Os can be controlled directly by programmable logic to construct soft memory controllers or high-performance I/O interfaces (e.g., MIPI). The number of DDR controllers scales across the device series to give the right amount of memory bandwidth for the compute resources of the device.

Designing for Power and Thermal Management

Edge devices and endpoints might need to run on battery power or fit within a strict thermal envelope. Using all the hard IP in the AI Edge series, such as the programmable NoC, memory controllers, and interface MACs, is key to lowering power consumption.

In addition, Xilinx provides users with more silicon options than other processor vendors to enable the trade-offs between low power and performance, including:

1. The performance speed grade from -1 (low), -2 (mid), and -3 (high)

2. The operating voltage from low (0.7V), mid (0.8V), and high (0.88V)

3. A static power screen from low or standard

4. Temperature grade options from extended (0°C to 110°C), industrial (–40°C to 110°C), automotive and defense grade (–40°C to 125°C), and military grade (–55°C to 125°C)

AI Edge devices are also separated into multiple power domains for finer-grained power control. The PMC actively monitors and controls the power domains in the device.

As described before, effective use of the memory hierarchy and I/O is also important, because using external memory and I/O, particularly transceivers, can increase power consumption significantly. This is the main reason for the new accelerator RAM in the AI Edge series.

Another intriguing way to reduce power consumption is the novel usage of the Dynamic Function eXchange (DFX) capability, previously known as partial reconfiguration, in AI Edge ACAPs. DFX enables using the same silicon area for multiple functions that are mutually exclusive. For example, while a vehicle is parked and unoccupied, the device is performing security related functions. While the vehicle is being driven, it is performing safety- and control-related functions. When the vehicle is no longer being driven, a third set of tasks is loaded. DFX can also be used to future-proof a system by doing over-the-air upgrades while a system is still operational.

Designing for Scalability

Besides a number of optimizations for low power, the AI Edge series provides a range of size, weight, and power-friendly devices with a common programming model and platform to enhance code portability, both within the AI Edge series and across the Versal portfolio. The same development methodology can be applied to a small device such as a smart sensor running at 5–10W all the way up to a 70W+ device in an in-trunk autonomous driving solution. Development for AI Edge ACAPs can even begin today with AI Core devices and then moved to the targeted AI Edge device. See Table 1.

Table 1: **Versal AI Edge Series Resource Summary**

Application	Intelligent Sensor		Edge Aggregation			Accelerator	
Device	VE2002	VE2102	VE2202	VE2302	VE2602	VE1752	VE2802
AI Compute (INT4)	11TOPS	16TOPS	32TOPS	45TOPS	202TOPS	101TOPS	405TOPS
AI Compute (INT8)	5TOPS	8TOPS	16TOPS	23TOPS	101TOPS	101TOPS	202TOPS
AIE-ML Tiles	8	12	24	34	152	–	304
AIE Tiles	–	–	–	–	–	304	–
Adaptable Engines	20K LUTs	37K LUTs	105K LUTs	150K LUTs	375K LUTs	448K LUTs	520K LUTs
Processing Subsystem	Dual-Core Arm Cortex-A72 Application Processing Unit / Dual-Core Arm Cortex-R5F Real-Time Processing Unit						
32G Transceivers	–	–	8	8	32	44	32
PCIe	–	–	1	1	–	–	–
PCIe + CCIX	–	–	–	–	1	1	1
Estimated Power	6–9W	7–10W	15–20W	~20W	50–60W	50–60W	75W

Designing for Functional Safety

A requirement for many edge nodes, such as AD/ADAS devices, is that performance and power requirements must be met while meeting the required level of safety. Xilinx has a long history in high reliability, safety, and thermally constrained systems in automotive, aerospace, satellite, industrial, medical, and commercial networking systems. ISO 26262 defines automotive safety integrity levels (ASIL). ASIL levels are determined by performing a risk assessment. AD/ADAS tasks typically target ASIL B, C, or D, with D being the highest and most challenging level to provide. Similarly, SIL levels under IEC 61508 and ISO 13849 govern industrial safety applications.

The AI Edge ACAP can provide up to ASIL C/SIL 3 capability in the LPD domain when running the Cortex RPU processors in lock-step mode. The processors can also operate without lock-step at ASIL B. The other domains are capable of ASIL B.

The triple module redundant PMC processors continually monitor the safety and security of the device. The PMC provides a clean separation of device management from the other processing domains. This domain holds the root of trust for security and performs many chip-wide monitoring tasks for functional safety.

ASIL Decomposition

When building systems, higher ASIL levels can still be achieved with components at a lower ASIL level using ASIL decomposition; "sufficiently independent" architectural elements are implemented in parallel with redundant components. For example, an ASIL D task can be decomposed into two different ASIL B implementations that perform the function and cross-check the result.

AI Edge devices allow partitioning of AD/ADAS tasks while achieving ASIL B with very high levels of processing performance and power efficiency. This enables robust systems at ASIL C/D to be created.

Critical control and communication functions that need higher ASIL levels can run on the LPD with sufficient functional isolation from the rest of the device.

Functional Isolation

A single device might need to host different tasks with differing safety criticality. Devices need to be able to isolate tasks in one domain from those in another. It is necessary to guarantee freedom from interference. The AI Edge ACAP provides isolation between and within domains to meet these requirements.

Functional Safety Challenges

Implementing functional safety brings many challenges. If devices are not architected from the ground up with functional safety in mind, it can be difficult and costly (in lost performance) to meet the solution requirements.

The AI Edge series is ideal to meet the challenge of providing functional safety at the right performance level for the following reasons:

- Safety features are built in from the ground up with the end goal in mind. The Versal architecture is partitioned with functional hardware features in each domain, as well as global resources to monitor and eliminate common cause failures (CCF).

- Compute acceleration resources are provided with hardware resources to reduce or eliminate the need for software test libraries (STLs).

- Safety features are included across all domains, not just a safe island.

- The subsystems in each device are certified as a Safety Element out of Context (SEooC). This gives the system integrator the right starting information to perform safety analysis for the system.

- Safety standards involve the entire development process and toolchains. Xilinx development processes and tools are aligned to this.

- Devices can be flexibly partitioned while guaranteeing freedom from interference. Chip pervasive memory access control and routing in the programmable NoC provide robust isolation that is tailored to the target application.

- Custom safety features can be added in the Adaptable Engines as needed. There is tool support to determine the failure in time (FIT) rate distribution in the Adaptable Engines to guide the user in certifying their custom designs.

Designing for Security

Similar to functional safety, security must be designed in from the ground up to meet the needs of mission-critical systems. AI Edge ACAPs have a robust set of security features for physical and cybersecurity standards such as IEC 62443, including:

- Hardware root of trust
  - o Boot time firmware and image authentication
  - o Crypto engines to decrypt images
  - o Debug security
  - o Key management
  - o Secure and measured boot with remote attestation
- Attack protection
  - o Tamper detection and protection
  - o Security attacks protection
- Run-time security
  - o Trust zone and trusted execution environment
  - o Memory and peripheral protection

Versal AI Edge Series Use Cases

The AI Edge series was designed for key applications in the Automotive, Industrial, Vision, Healthcare & Sciences (ISM), and Aerospace & Defense markets, but is also broadly applicable to applications that require high performance per watt, including automotive forward-looking cameras, drones/UAVs, and robotics, where sensor fusion combines with AI in a small footprint.

Automotive Use Case

Historically, the Xilinx Automotive family of FPGA and SoC devices have provided unparalleled scalability and performance/power to allow customers to "right-size" the processing horsepower and cost for their ADAS and automated driving platforms. The AI Edge series follows in that tradition. To illustrate, a Smart Forward-Looking Camera (Edge Compute) use case is considered, as shown in Figure 2.

The automotive forward-looking camera represents a series of products focused on providing drivers a suite of convenience and safety features that include adaptive cruise control (ACC), lane departure warning (LDW), automated emergency braking (AEB), traffic sign recognition (TSR), and traffic jam assistant (TJA), among others. The widespread deployment of forward-looking camera products is being accelerated by regional New Car Assessment Program (NCAP) guidelines.
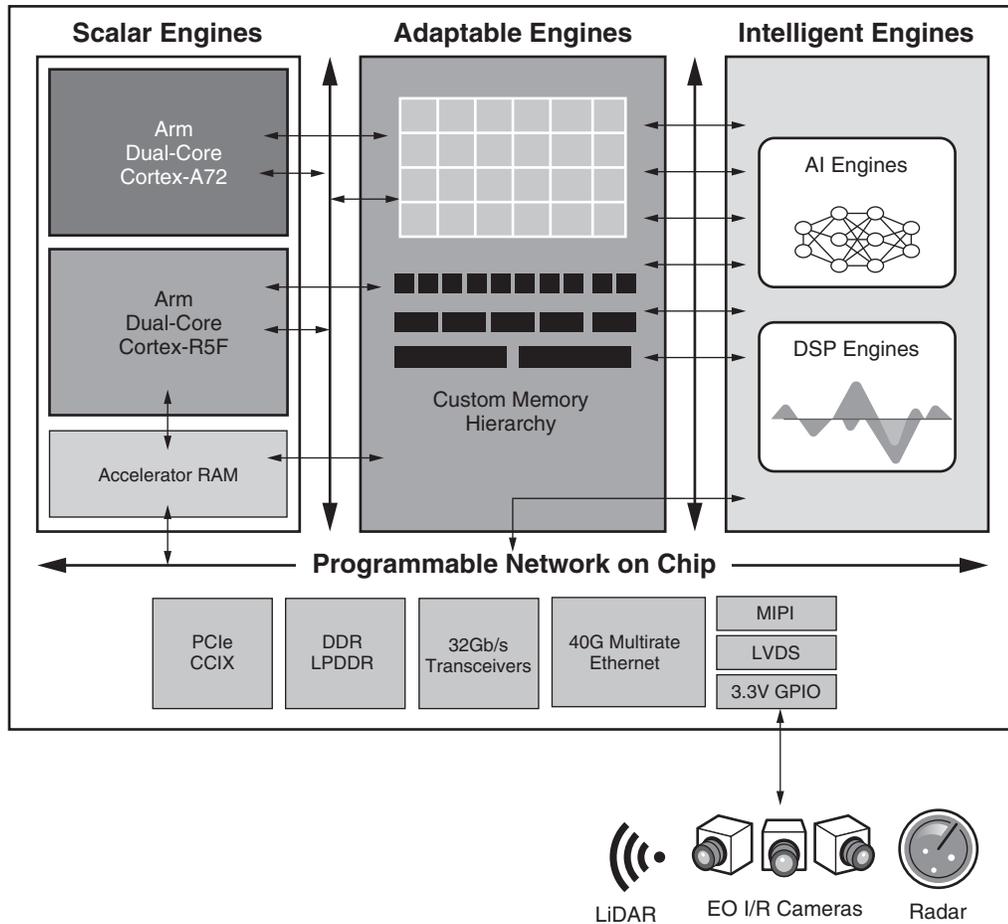
Figure 2: **Forward-Looking Camera Versal AI Edge Architecture**

Figure 2 illustrates a typical architecture and functional partitioning for the forward-looking camera use case. The I/O in AI Edge devices supports the preferred MIPI interface for cameras. It is important to know that the programmability of the I/O means that the number of MIPI channels and associated data lanes is limited only by the number of I/O pins, as these are not dedicated MIPI interfaces like those found on ASSP devices. This means an AI Edge ACAP-based forward-looking camera platform can support a single imager for mono processing, two imagers for stereo processing, or multiple imagers with various fields of view (FOVs) and resolutions.

A typical vision processing pipeline includes ISP, computer vision analytics (e.g., optical flow and object detection), and object classification. The vision data can also be combined/fused with data from other forward-looking sensors, e.g., radar/LiDAR. The interface to these sensors is commonly CAN/CAN-FD or Ethernet. The AI Edge ACAP provides hardened peripherals to support these interfaces, and the system designer has the option to expand the number of interfaces with soft IP cores that can be implemented in the Adaptable Engines' fabric. All of the vision processing, fusion, and object detection/classification functions are accelerated in the ACAP's adaptable and Intelligent Engine domains.

Of particular interest is CNN processing acceleration for object classification and scene segmentation. The Adaptable Engines can be used to fuse data from multiple sensors and manipulate/condition that data to create input feature maps for the CNN computational

processing, which is carried out by the AI Engines. For high utilization of the AI processing cores, the NoC, fabric UltraRAM/block RAM, and accelerator RAM memories, enable efficient datapaths based on performance and specific network needs.

Finally, the ACAP's Scalar Engines run the overall system application software to orchestrate the rest of the device and make the necessary assessment and vehicle control/alert decisions, which comprise the forward-looking camera features. The Cortex-R5F complex can be run in lock-step for safety critical decisions and vehicle communications.

Industrial, Vision, Healthcare, and Sciences Use Case

Robotics are the ultimate expression of industrial automation systems. Precision control, deterministic communications, machine vision, responsive AI, cybersecurity, and functional safety are key technology considerations when building cobots, medical surgical robots, medical assistant robots, and automated guided vehicles (AGV) for factories and warehouses, as well as other applications of industrial and healthcare robotics. Furthermore, these wide ranging technologies need to seamlessly converge and interoperate as part of one compact system. Versal AI Edge ACAPs enable a modular approach to robotics by providing a common embedded hardware and software platform that can be reused and scaled to dramatically reduce time to market, form factor, and total cost of ownership, while maximizing intelligence and adaptability of these assets. Specifically, the AI Edge series enables:

- Precise, deterministic control over a scalable number of axes of motion via parallel processing of control loops without impacting cycle times

- Connectivity over multiple, fragmented industrial Ethernet standards, including support for OPC UA and DDS over time-sensitive networking (TSN)

- Support for numerous and heterogeneous sensor inputs (e.g., RGB, Infrared, time-of-flight, accelerometer, LiDAR, and environmental), enabling sensor fusion and high-level sensors techniques

- Real-time analytics and ML supporting predictive maintenance, model predictive control, remote diagnostics, digital twin, and additional edge intelligence use cases

- An embedded software infrastructure, enabling industry-leading support for mixed criticality applications and information technology / operational technology (IT-OT) convergence

- Compliance for IEC 61508 SIL 3 and ISO 13849 PLe CAT4 functional safety

- IEC 62443 cybersecurity, supporting secure and measured boot with remote attestation

- Smaller physical footprint and power through highest levels of integration

- On-chip acceleration for real-time motion planning to maximize efficiency and safety, reducing the need to be tethered to an external industrial PC

Many new use cases are enabled and simplified by the AI Engines. For example, the AI Engines in combination with two Cortex-A72 processors running Linux can be used to improve robot operating system (ROS) responsiveness by a factor of 10X in handling the complex mathematics required by such frameworks. The AI Engines make ROS perform as a very responsive system, allowing for much faster, more precise movement. Proprietary robotic frameworks can also benefit

from the seamless integration between the Cortex-A72 processors and the AI Engines for improving response time.

Linux is also aided by type 1 hypervisor technology supported by the Cortex-A72 processors and XEN Linux kernal with cache coloring, allowing higher predictability of the entire system.

Python libraries, such as Numpy, Scipy, and Matplotlib can be used interactively with the AI Engines. Accelerating Numpy mathematics by an order of magnitude compared to the Cortex-A72 processors alone makes it possible to handle and transform multiple streams of machine data coming from sensors and from the network into executable local information, thereby avoiding unneeded data transfer both to the cloud and to on-premises computing systems.

Security also benefits from the localized high-performance processing offered by the AI Engines and Cortex-A72 cores, e.g., sensitive data does not need to leave the factory floor or surgery room in its raw form, exposing the system to possible data eavesdropping or exploitation.

Furthermore, the AI Engines can execute advanced control algorithms like model predictive control (MPC), as well as machine learning-based or traditional algorithms for locally-based predictive maintenance.

The benefits are very tangible: improved performance, optimized productivity, maximized equipment uptime, and over the long term, extended life for that asset, greatly reducing the total cost of ownership. See Figure 3.
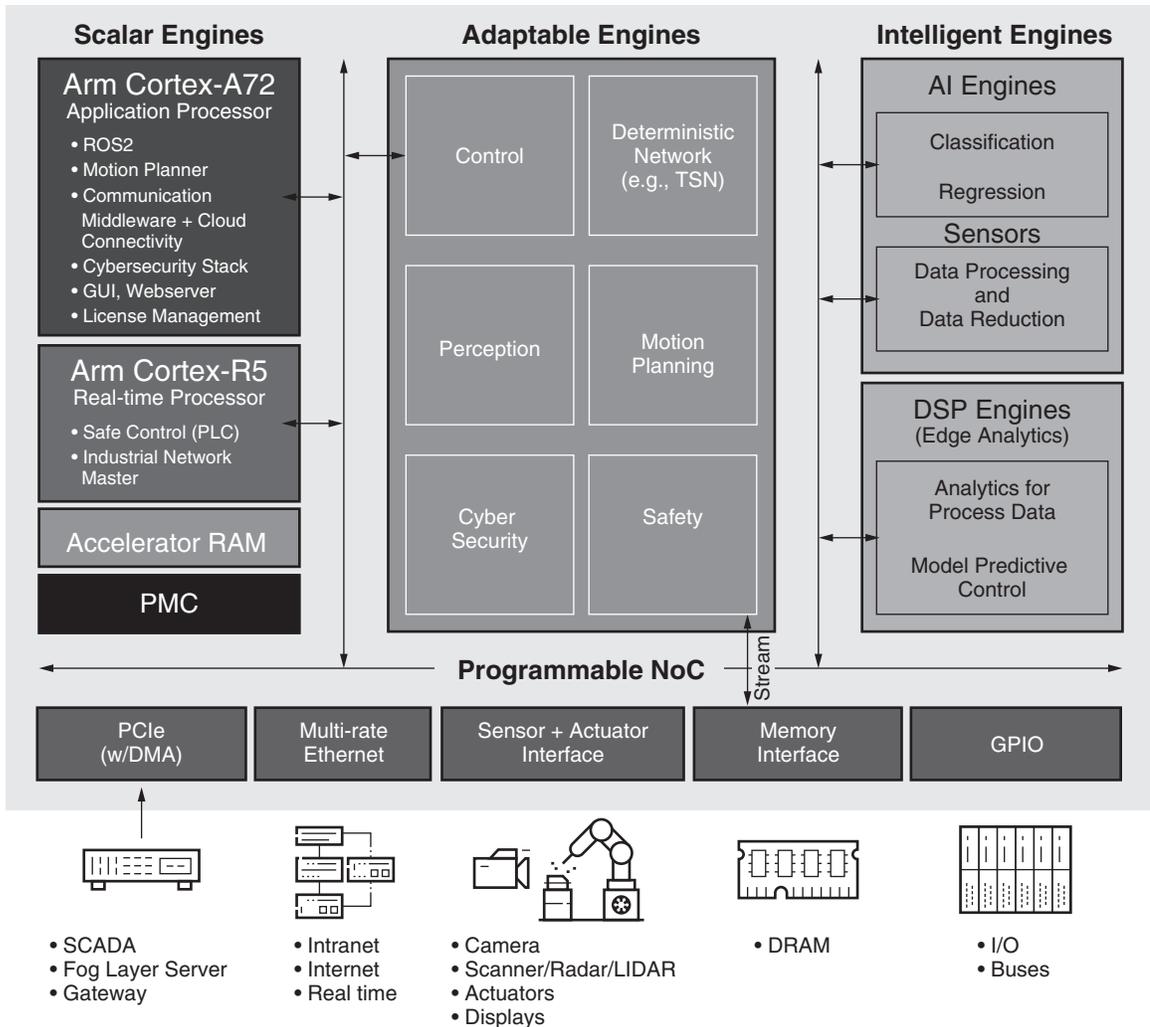
Figure 3: **Vision-Guided Collaborative Robots with Versal AI Edge Architecture**

In machine vision scenarios (Figure 3), the AI Engines implement tasks such as object detection, recognition, classification, re-identification, region of interest, and many other emerging algorithms in a more efficient (performance/watt) manner than previously possible.

AI systems not only need to be safe systems, but AI Engines can be used to augment traditionally rigid functional safety controls, producing more dynamic context-based execution. Collaborative robots that interact very closely with humans benefit most, because the AI Edge ACAP's safety qualification, advanced sensors, vision and radar-based environment awareness, and predictable and fast reaction time make the collaboration extremely safe and effective.

The AI Edge devices are not traditionally constrained industrial or healthcare IoT edge devices; rather, their ability to solve computationally involved problems closest to the analog-digital boundary make them ideal for the new Digital Twins concept. In most industrial and medical scenarios, the cycle time of the system, typically operating at 1ms or faster, requires intelligence at the edge. In either case, where real-time response is required, or alternatively, offloaded to cloud computing, Xilinx offers users the needed flexibility to solve challenges where the physics of the problem dictate.

Aerospace and Defense Use Case

Xilinx has a 30+ year track record of commitment to the industry, including support for defense-grade, mil-temp, radiation tolerant, and radiation hardened devices, as well as enabling DO-178C and DO-254 system-level safety certifications. The AI Edge series continues this heritage with low-size, weight and power (SWaP) devices optimized for real-time, high-performance applications in the most demanding environments.

A good example of an Aerospace & Defense use case for the AI Edge series is a multi-mission situationally aware payload for a drone or UAV. The multi-mission nature can include radio for communications data links and navigation, radar for target-tracking, and identification friend or foe (IFF), electro-optical sensors for visual reconnaissance, and other sensor fusion. The situational awareness can be derived from post-processing and AI algorithms for signals intelligence (SIGINT), and electronic warfare derived from the sensor data, with the goal of ascertaining actionable intelligence.

While it is extremely challenging to aggregate and process all the sensor data within a strict power budget, the AI Edge series is uniquely adapted for just such a task. A plethora of programmable I/O is available to support sensor inputs from radio, radar, cameras, and other sensors. Moving all the data in a highly efficient manner requires hardened infrastructure like the programmable NoC and low-power DDR memory controllers. Sensor fusion of all the data is a good task for the Adaptable Engines, which can also do any signal conditioning that might be required.

The cacheless-memory hierarchy ensures that all the data can be accessed quickly and in a deterministic fashion. Taking all the data and deriving intelligence from it is a good fit for the Intelligent Engines, particularly the AI Engines. They can use various neural networks to do waveform and/or object classification, as well as target tracking and IFF; as the algorithms evolve, they can be upgraded over the air. The Scalar Engines are optimal for doing assessment and decision making, as well as command and control. If necessary, mission-critical security features such as the Cortex-R5F processors running in lock-step and triple modular redundancy (TMR) can be used to ensure safety and security. And since there is a mix of small to large UAVs in use on the battlefield, the AI Edge series devices ensure scalability to meet the different constraints of each UAV with regard to SWaP-optimized performance. See Figure 4.
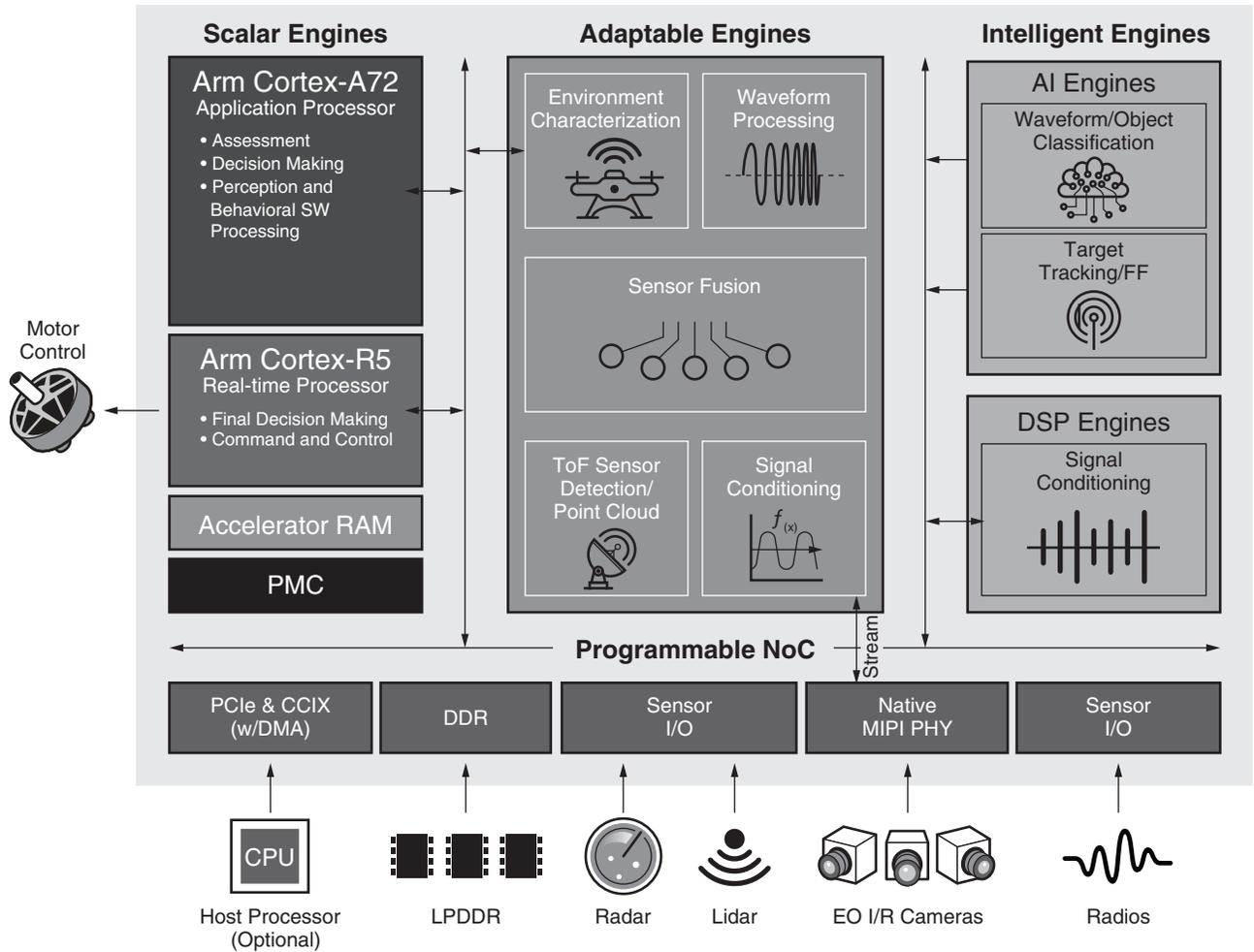
Figure 4: **Multi-Mission Situationally Aware UAV Payload Versal AI Edge Architecture**

As shown in Figure 4, the ability of the Versal AI Edge devices to be the single "brain" of a multi-mission situationally aware UAV payload makes it the ideal SWaP-friendly processing solution to support all the mission-critical requirements, now and in the future.

Conclusion

This white paper has outlined the benefits of the Versal ACAP architecture as applied to the AI Edge series. This series of devices scales from sub-10W intelligent edge devices all the way to >100W in-trunk or on-premises supercomputer systems or payloads.

The architecture has the flexible resources to accelerate the whole application. Whether the sub-tasks include low-latency inferencing, sensor fusion, and/or managing an array of interfacing needs, the AI Edge series can be adapted to the task. This adaptability makes the best use of available power budgets.

In addition, functional safety and security were designed into the series from the ground up, ensuring that AI Edge ACAPs are capable of meeting system-level safety certifications.

To learn more about Xilinx's Versal AI Edge series, go to: https://www.xilinx.com/products/silicon-devices/acap/versal-ai-edge.html

Acknowledgments

Related Reading

1. Xilinx Website, Automotive Qualified Zynq UltraScale+ MPSoC Family
2. Xilinx White Paper WP505, *Versal: The First Adaptive Compute Acceleration Platform (ACAP)*
3. Xilinx White Paper WP506, *Xilinx AI Engines and Their Applications*

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
06/09/2021	1.0	Initial Xilinx release.

Disclaimer

Automotive Applications Disclaimer