AMD

Accelerate Your AI-Enabled Edge Solution with Adaptive Computing

Introducing Adaptive System-on-Modules (SOMs)

Executive Summary

AI-enabled applications are increasingly being deployed at the edge. Adaptive SOMs provide a comprehensive, production-ready platform to build accelerated, AI-enabled applications at the edge, including smart vision across cities, factories, and hospitals. Such applications require a large amount of processing to be performed with low latency, low power consumption and in a small footprint. To achieve this combination, the whole application (both the AI and non-AI processes) must be accelerated.

As AI models rapidly evolve, the acceleration platform must also be adaptable. This allows for optimal implementation of not just today's AI techniques, but tomorrow's as well. In this eBook, we give an introduction to AI at the edge and explain how SOMs can be a good production solution. We also introduce adaptive computing and detail the benefits that adaptive SOMs bring. Finally we detail considerations when selecting an AI-enabled edge solution.

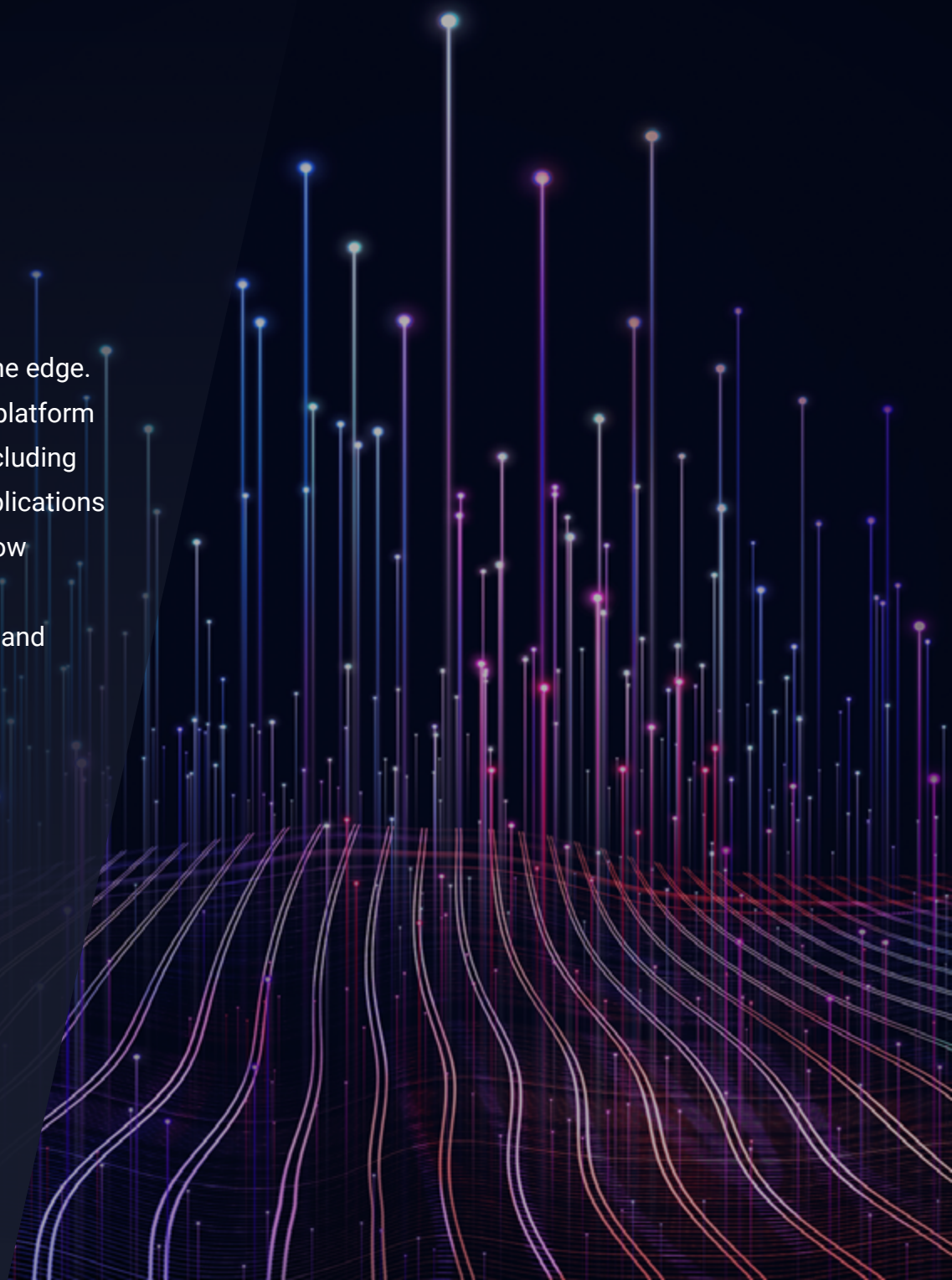


Table of Contents

CHAPTER 1

The Rise of AI-Enabled Edge Computing

SOM USE CASE:
CROP HEALTH ANALYSIS

High-performance AI inference is required with ultra-low power consumption to maximize functionality with minimal impact on battery life.

AI-enabled applications are increasingly being deployed at the edge and endpoint. A plethora of applications are being developed that use the latest AI techniques across a range of industries and geographic locations.

Cities are becoming smarter with automated vision applications that manage safety and alert emergency services when help is needed. Industrial IoT applications increasingly require high-performance AI inference processing as factories become smarter and more automated. Even the retail experience is beginning to change with smart retail bringing automated shopping experiences not thought possible a decade ago. All these applications must operate with utmost reliability, in harsh conditions, for 10 years or more. The applications require high performance, yet must be delivered in an efficient, reliable, and compact form factor.

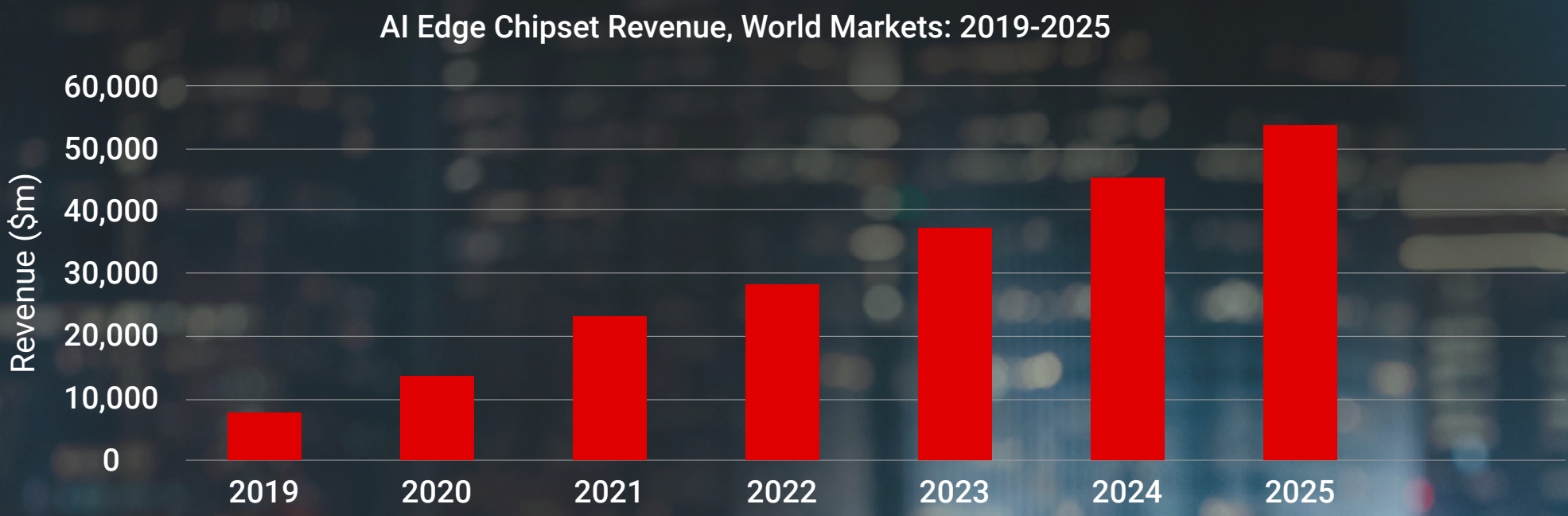


Figure 1. AI edge global chipset revenue is forecasted to exceed $50 billion by 2025. (Source: Omdia)

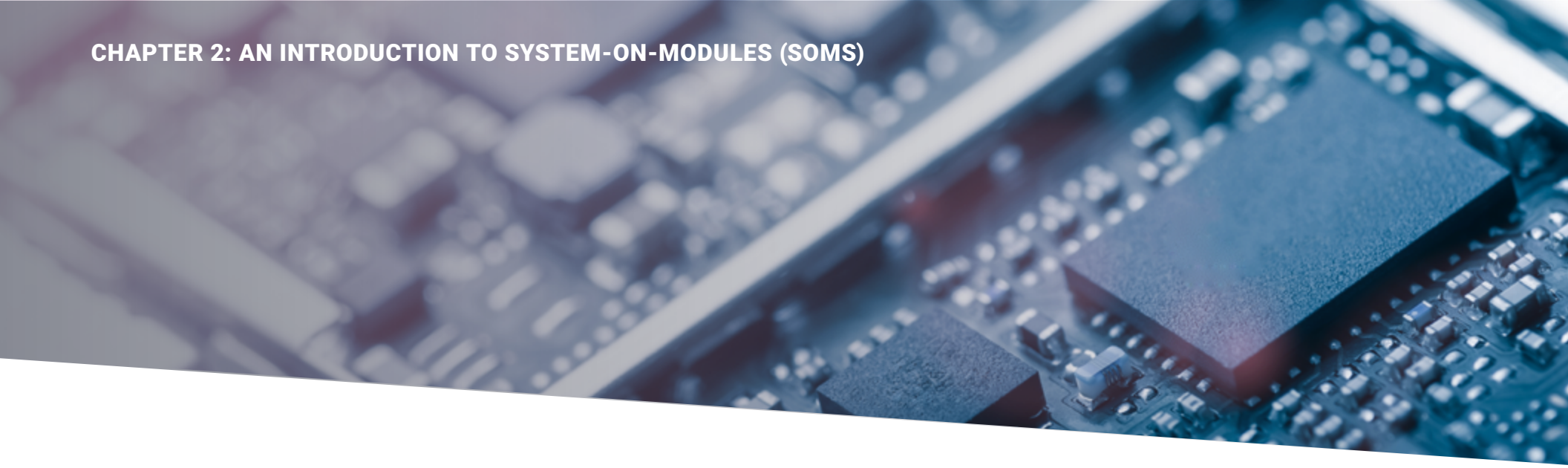
The trend towards increasing AI-enabled edge computing is highlighted by the market research firm Omdia, which estimates that global revenue for AI edge chipsets will surpass $50 billion by 2025.

CHAPTER 2

An Introduction to System-on-Modules (SOMs)

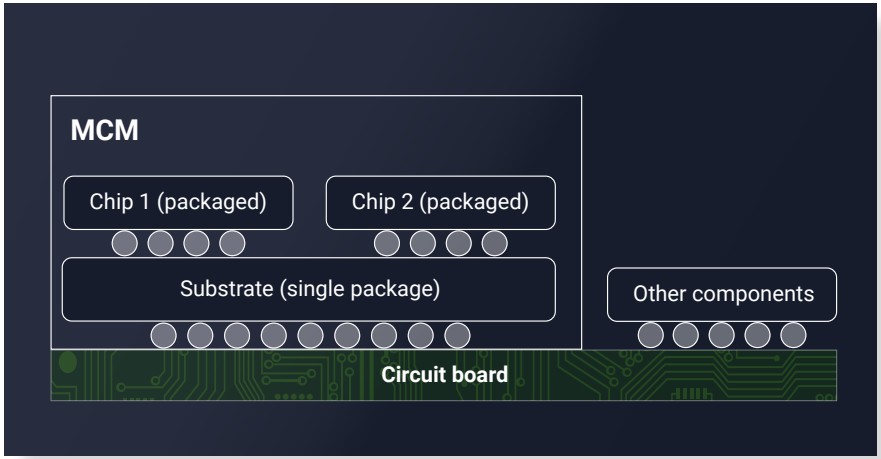
SOM USE CASE:
3D PRINTING ARM CONTROL

The 3D printing space is rapidly evolving and designers are looking to improve both speed and quality of the process using precise, deterministic control over a scalable number of axes of motion.



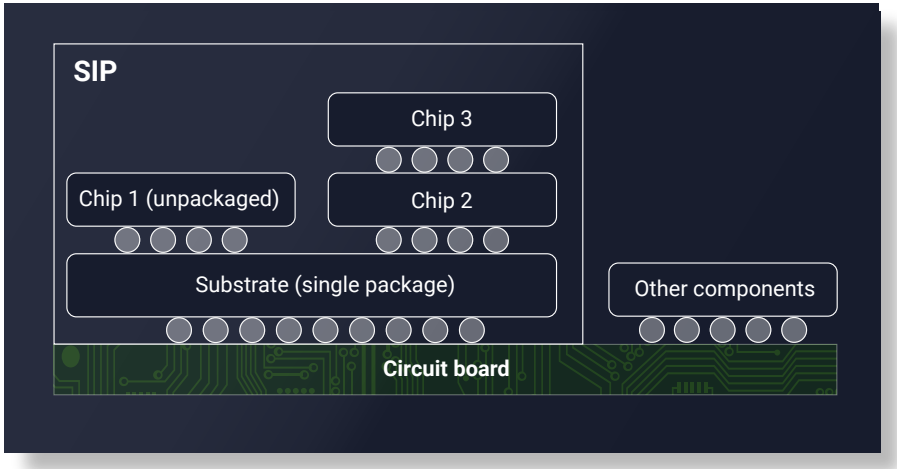
When building an AI-enabled edge application there are numerous options available. In many industries, it's common for a hardware design team to perform a "chip-down" development, where specific silicon devices are chosen, and a fully customized circuit board is developed for the application. While this does produce a highly optimized implementation, it can take significant development time and cost to reach production readiness.

To save the expense and time of a chip-down development, design teams may consider utilizing a more integrated solution such as Multi-Chip Module (MCM), System-in-Package (SIP), Single-Board Computer (SBC), or System-on-Module (SOM).



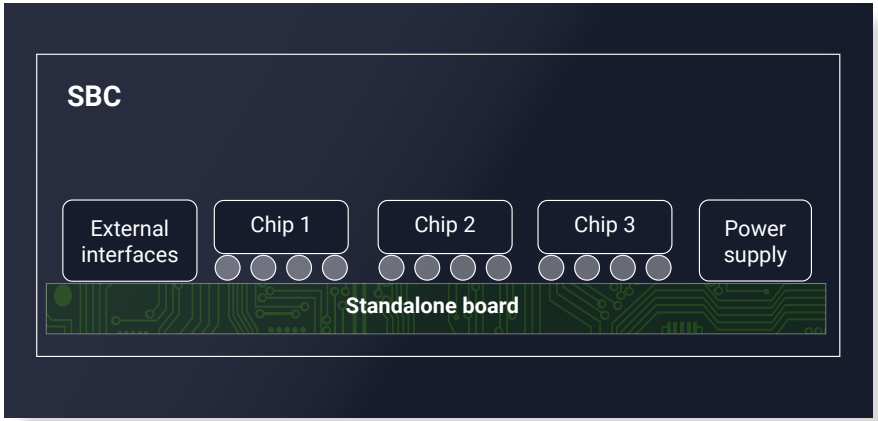
Multi-chip Module (MCM)

An MCM combines several (usually packaged) silicon devices within a single larger package. This has the advantage of simplifying circuit board design by reducing the number of individual devices on the board. Although there are often advantages in cost and performance, MCMs are only available for some specific applications and still require custom, complex circuit boards to be developed.



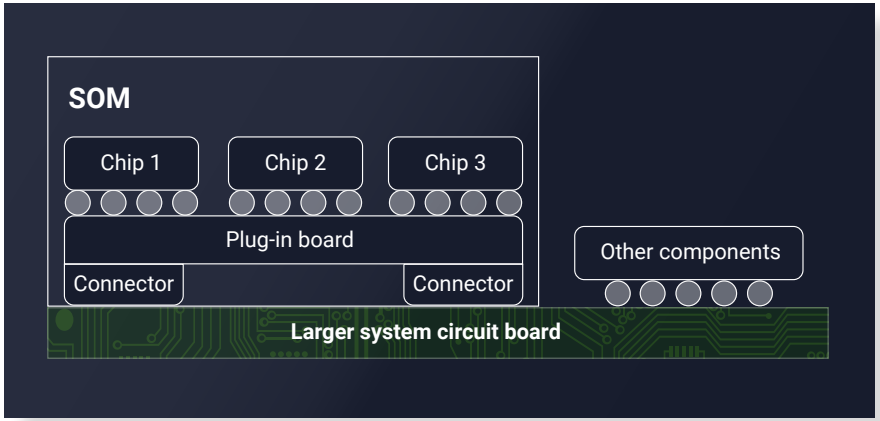
System-in-Package (SIP)

A SIP is like an MCM, except the individual silicon die are often unpackaged. Some SIPs also have dies that are stacked on top of each other in a 3D array. SIPs also can include discrete components such as resistors and capacitors, reducing the need to include these components on the circuit board. SIPs, like MCMs, still need a fully custom circuit board to be designed however – i.e. requiring a full, chip-down design flow.



Single-Board Computer (SBC)

An SBC is a complete computing system built on a circuit board and delivered as a standalone product. Typically, these include a small microprocessor, memory and input/output (I/O). SBCs are usually not built or validated for production deployment.



System-on-Module (SOM)

A SOM is a computing system like an SBC but is designed to connect into a larger solution, rather than be standalone. When built into high-performance edge applications, SOMs have many advantages over the other solution types mentioned. SOMs provide a complete, production-ready computing platform, and save significant development time and cost vs. chip-down development. SOMs can plug into a larger edge application, providing both the flexibility of a custom implementation with the ease-of-use and reduced time-to-market of an off-the-shelf solution.

"Using a SOM instead of a chip-down solution will save significant development costs. There is shorter development time and fewer mistakes are possible during development, which both lead to improved time-to-market. Furthermore, SOMs come pre-tested to reduce risk and make it easier for engineers to implement their custom designs. They don't have to focus on the infrastructure that's on the SOM. Instead, they can focus on the key differentiators they care about."

– Bryan Fletcher, Technical Marketing Director at Avnet

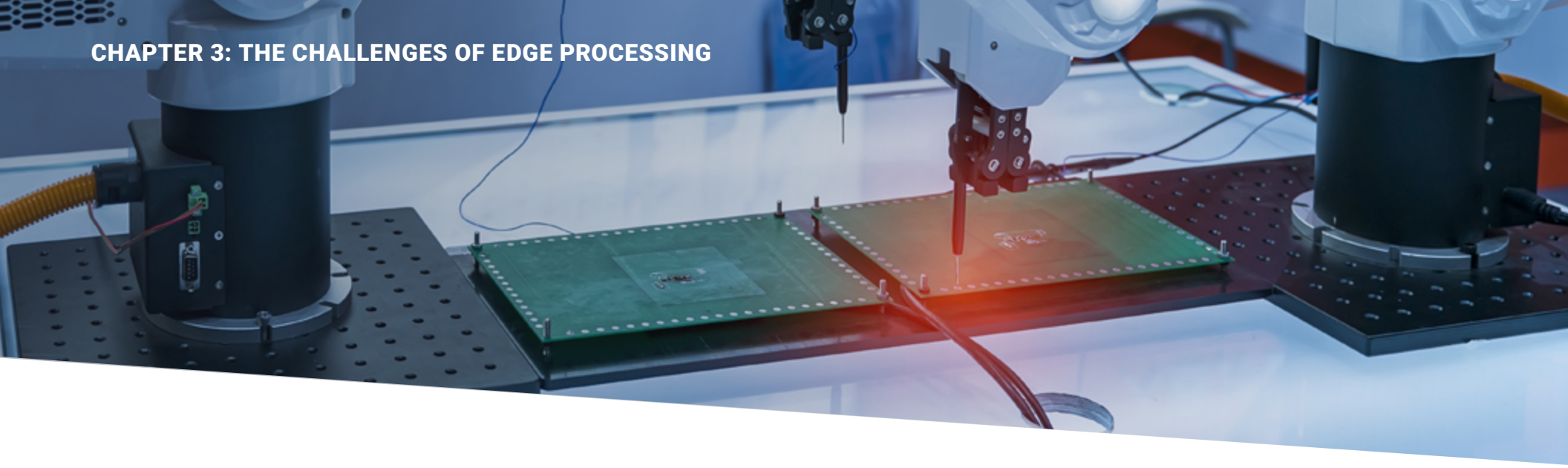
SOMs are therefore production-ready, proven designs that contain all the silicon components and interface connectors needed for their specific application area.

CHAPTER 3

The Challenges of Edge Processing

SOM USE CASE: DIGITAL PATHOLOGY

Typical procedures within radiology, pathology, dermatology, and ophthalmology are processing large image sizes requiring complex image processing, with AI workflows being particularly compute- and memory-intensive.



Edge computing is typically limited by power consumption, footprint, and cost. As processing demands increase, the challenge of providing the performance level required, within the limitations of edge processing, has increased exponentially.

CPUs have seen many improvements at the edge, but the gains have slowed in recent years. Unaccelerated CPUs struggle to provide the performance needed for the next generation of AI-enabled edge applications, especially when considering the tight latency requirements.

To implement advanced AI applications at the edge, a domain-specific architecture (DSA) is needed. DSAs provide a highly optimized implementation of an application vs. an unaccelerated CPU. They also provide determinism and low latency.

Whole application acceleration is required to implement efficient AI-enabled applications at the edge.

AI inference needs non-AI pre- and post-processing, all of which have higher performance requirements. A suitable DSA will be designed specifically to process the required data efficiently – both the AI inference, and the non-AI parts of the application (i.e. the whole application). Therefore, whole application acceleration is required to implement efficient AI-enabled applications at the edge (and elsewhere).

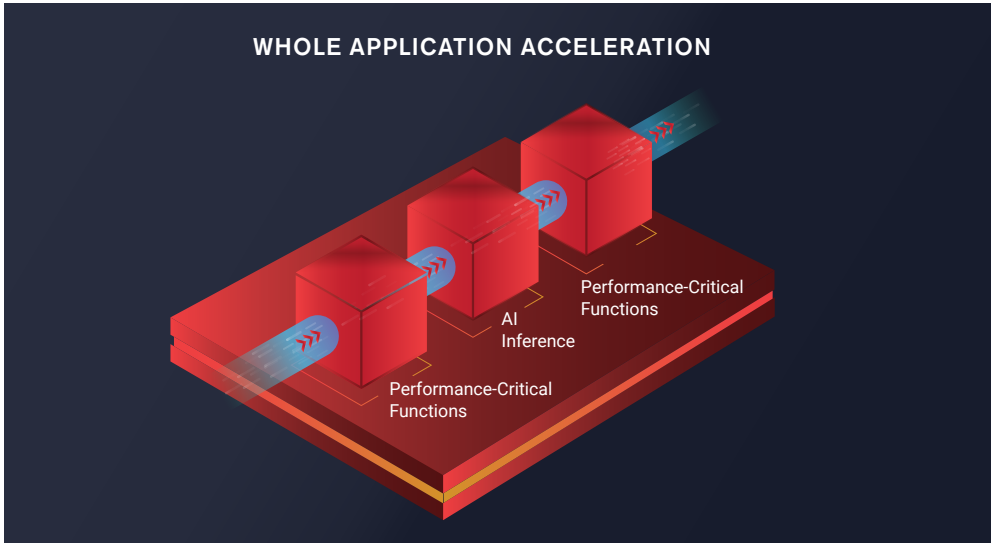


Figure 2. AI Inference is typically part of a larger processing system

Some ASSPs have been developed for AI edge applications, however, like any fixed silicon solution, they have limitations. Primarily, the pace of AI innovation is incredibly rapid, rendering AI models obsolete much quicker than non-AI technologies. Fixed silicon devices that implement AI can quickly become obsolete due to the emergence of newer, more-efficient AI models. It can take several years to tape-out a fixed silicon device by which time the state-of-the-art in AI models will have advanced significantly.

Security and functional safety requirements are also becoming more important for edge applications, often resulting in potentially expensive field updates. Systems with fixed hardware that cannot adapt to these changes will quickly become legacy systems. AI-enabled systems at the edge must be flexible enough to extend their operating life and maximize return on investment.

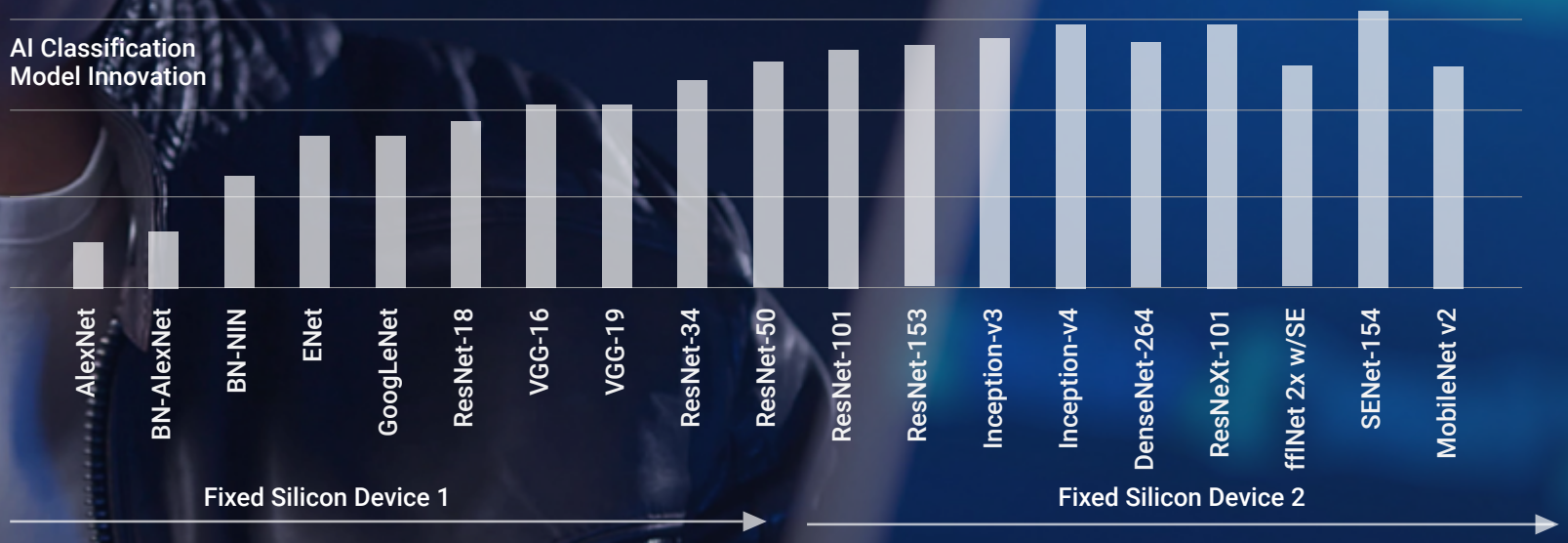


Figure 3. AI models evolve rapidly – much faster than silicon development cycles

CHAPTER 4

Introducing Adaptive Computing

SOM USE CASE: ROBOTIC CONTROL IN FACTORIES

A true industrial solution supports extended temperatures, harsh environments, and long lifecycles in order to be trusted in applications where downtime is impermissible.

One of the most-promising technologies for AI-enabled edge applications is adaptive computing. Adaptive computing encompasses hardware that can be highly optimized for specific applications such as Field Programmable Gate Arrays (FPGAs). In addition to FPGAs, new types of adaptive hardware have been recently introduced, including the adaptive System-on-Chip (SoC) which contains FPGA fabric, coupled with one or more embedded CPU subsystems.

Adaptive computing is more than just hardware, however. It also encompasses a comprehensive set of design and runtime software, that, when combined, delivers a unique adaptive platform from which highly flexible, yet efficient systems can be built.

Adaptive computing allows the hardware to be purpose-built for the application, yet still affords adaptation as needed, if workloads or standards evolve.

Adaptive computing allows for DSAs to be implemented without the design time and upfront cost needed by custom silicon devices such as ASICs. This allows rapid deployment of a flexible, yet optimized solution for any given domain, including AI-enabled edge applications.

Adaptive SoCs are ideal for such domain-specific processing because they combine the flexibility of a comprehensive, embedded CPU subsystem with the optimal data processing of adaptive hardware.

Adaptive computing allows the hardware to be tailored for the application, yet still affords adaptation as needed, if workloads or standards evolve.

CHAPTER 5

The Adaptive System-On-Module (SOM)

SOM USE CASE: ACCESS CONTROL

AI-powered edge applications like access control require high-performance, low latency implementation, but must remain within limited power and footprint requirements.

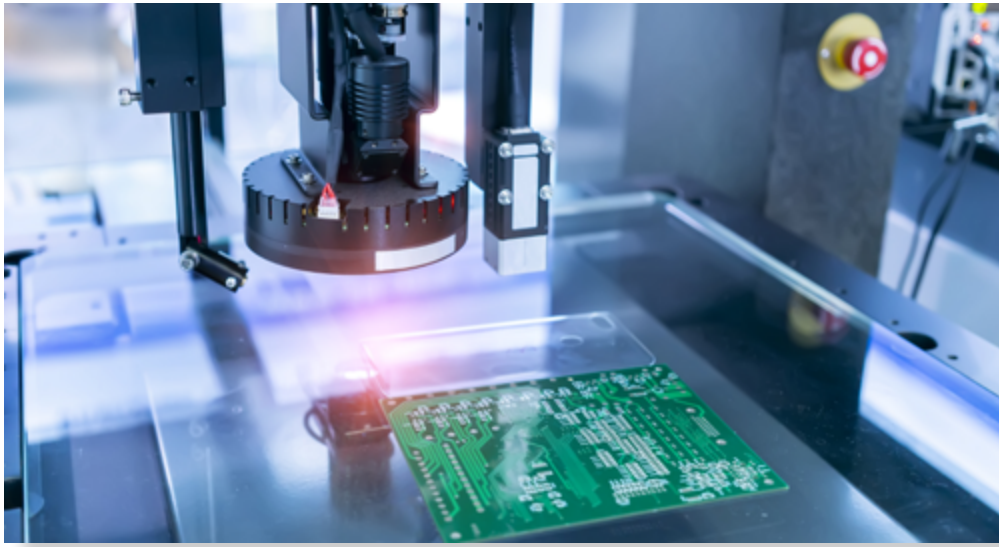
As we have seen, SOMs provide a good platform for edge applications. However, to achieve the performance required by modern AI-enabled applications, acceleration is needed. Adaptive computing provides the acceleration required for AI applications at the edge.

Typical application development teams are composed of several types of hardware engineers and software engineers. Circuit board designers are responsible for designing custom boards for the required application. In cases where adaptive computing (including FPGAs) is used, RTL designers have traditionally been responsible for configuring the adaptive devices using Hardware Description Languages (HDLs) such as Verilog and VHDL. And software developers write code running on embedded CPUs using languages such as C++ and frameworks such as OpenCV, which is commonly used for embedded vision applications.

Traditionally, all three disciplines are required to build applications using a custom circuit board, with custom hardware inside the adaptive SoC, and software running on the embedded CPU.

As previously mentioned, the design and manufacture of a custom circuit board containing the adaptive SoC and other physical components needed by the application is called 'chip-down' development. For decades, circuit board designers have used chip-down development to take advantage of adaptive computing. In fact, adaptive SoCs have been deployed across many AI-enabled edge applications using this method, including in tens of millions of automobiles.

Some applications still require custom hardware components to interface with an adaptive SoC and thus chip-down design is needed. However, an increasing number of AI-enabled edge applications need similar hardware components and interfaces,



even for vastly different end applications. As industries have moved towards standardized interface and communications protocols, the same set of components are suitable for a variety of applications, despite having vastly different processing needs.

An adaptive SOM for an AI-enabled edge application incorporates an adaptive SoC with industry-standard interfaces and components, allowing developers with no or minimal hardware experience to benefit from adaptive computing technology. An adaptive SoC can implement both the AI and non-AI processing, thus the whole application.

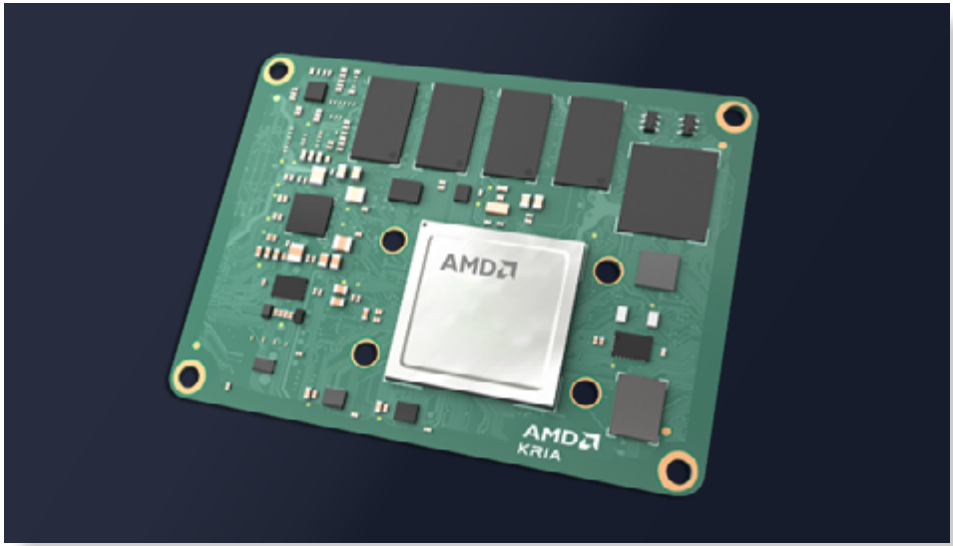


Figure 4. An example of an adaptive SOM

An adaptive SoC on an adaptive SOM enables a high degree of customization, without the need for a custom PCB. It is designed to be integrated into larger systems and uses a predefined form factor.

Adaptive SOMs make it possible to take full advantage of adaptive computing without having to do chip-down design. An adaptive SOM is just part of the solution, however. The software is also a critical consideration. We will cover software considerations in Chapter 7.

CHAPTER 6

Adaptive SOM Benefits for Hardware Developers

SOM USE CASE:
AUTOMATED GUIDED VEHICLES

Logistics companies require the latest AI models to enhance productivity and minimize downtime. This must be delivered in a compact, low power and rugged solution.

Typically, hardware developers build AI-enabled edge hardware by selecting their main silicon devices, then developing an edge form factor board to accommodate the chosen devices. As discussed, this chip-down development can be a costly and complex process with a long development cycle.

First, developers need to evaluate and select a device, build a prototype, and then prove the architecture works sufficiently well with the required software and AI models they plan to deploy. The evaluation process alone can take months. Once evaluation is completed, the production board needs to be developed, integrated, and manufactured before it can be deployed.

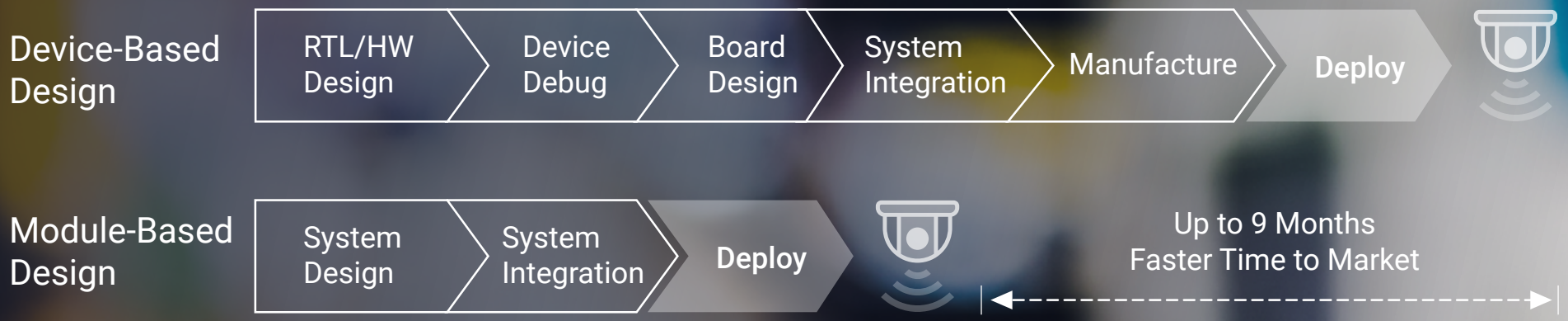


Figure 5. SOMs can help accelerate time to market and reduce development cost

Adaptive SOMs allow developers to use an off-the-shelf, production-ready solution, saving significant development cost and time. The illustration above shows how a SOM-based design for an embedded vision application can be completed up to nine months faster than a typical chip-down approach.

An adaptive SOM is advantageous for hardware developers because it allows the design to be changed late in the design process, vs. a SOM based on fixed-silicon technology.

CHAPTER 7

Adaptive SOM Benefits for Software Developers

SOM USE CASE: SURGICAL ROBOTS

Industrial control, communications, machine vision, machine learning, human-machine interfaces, cybersecurity, and safety are key technology considerations.

The advantages of adaptive SOMs are not just limited to hardware developers. Software developers can accelerate their design cycles as well by using pre-built configurations for the underlying adaptive SoCs.

Recent advancements in software tools, libraries, and frameworks can enable some design teams to use adaptive computing without burdening hardware engineers. The available comprehensive software platforms allow software engineers to utilize the capability of the entire adaptive SoC without needing specific hardware customization.

Solutions such as accelerated APIs and pre-built accelerated software platforms can be found freely available on adaptive SoC vendors' websites, as well as from third-party software vendors. Both also provide access to a range of AI models capable of performing common AI inference functions.

An adaptive SOM can provide a simple-to-use, out-of-the-box experience for software developers, from within familiar environments, including Python, C++, TensorFlow, and PyTorch.

In addition to prebuilt platforms and APIs, comprehensive software tools enable full customization of the adaptive hardware, enabling even more flexibility and optimization.

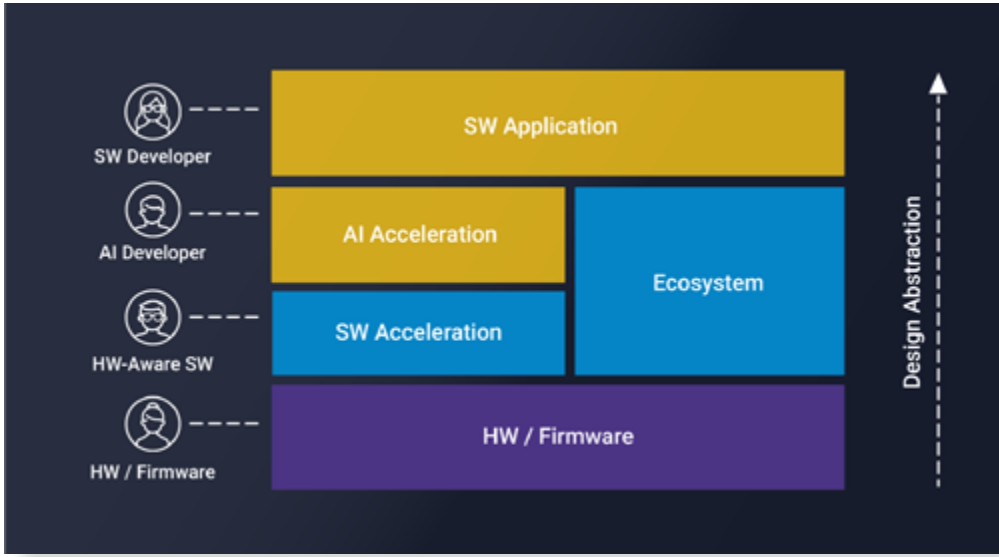


Figure 6. Adaptive SOMs can be programmed from several abstraction levels

CHAPTER 8

What to Look for When Selecting an Edge Solution

SOM USE CASE:
MOTOR CONTROL WIND TURBINES

Wind turbines need to be upgraded remotely and securely–throughout the world, onshore and offshore, and over the course of decades.

When evaluating a solution for edge applications, several factors should be considered:

Development Time and Cost

Successful design teams will look for a turnkey solution that accelerates development time and enables fast prototyping. More than just hardware alone, the solution should consist of an operating system, a board support package, reference designs, and open, system-oriented documentation with easy explanation and guidance. Design cycles can further be sped up using pre-built starter kits that allow developers to immediately start application development with a direct path to production using the same hardware and software. Reduced development costs especially help small and medium-sized companies benefit from adaptive computing.

Performance/Power Ratio

The solution should provide optimal performance per watt, which is especially important for applications at the edge typically operating within a pre-set power envelope of 10W and below. It should deliver domain-specific architectures with accompanying libraries and capabilities for trading off performance and power to suit the needs of a given application. Low latency performance is not always the same as high-latency performance. For example, GPUs typically batch data to achieve higher throughput, but at the cost of latency. The low-latency throughput is typically nowhere near as high as peak throughput.

Security

The solution should be able to support IEC 62443-grade cybersecurity, the security standard for the Industrial Internet of Things. This implies, at a minimum, that the system is built around a hardware-root-of-trust and includes support for secure and measured boot with remote attestation of software applications. Furthermore, the security solution should be able to adapt over the industrial lifecycle of edge equipment, typically 10-20 years, to keep pace with evolving threats over that same duration.

Future-Proofing

Beyond security, the platform should provide future-proofing to allow both hardware and software updates, even after the systems are deployed in the field via over-the-air (OTA) updates. The solution must be industrial-grade, with a lifecycle of at least 10 years in the field. For similar reasons mentioned with security and the industrial lifecycle, the solution should accommodate constantly changing AI models as approaches to AI have exploded in number in recent years and will continue to evolve for many years to come.

Ecosystem

The solution should feature a sizable ecosystem, offering a variety of services and solutions, including complementary hardware, software, and tools that are pre-integrated with the chosen AI platform or alternative AI platforms. And, the solution should include accelerated application libraries to give software developers the same jumpstart on design that reference designs do for hardware developers.

Adaptive SOMs bring the time-to-market advantage of a SOM-based solution, with the application optimization that can only be achieved with adaptive computing. This makes it an ideal approach for implementing AI-enabled edge applications.

"We've been recently seeing a lot of interest in SOMs for ML and AI applications. Many customers do not want to get involved in the specifics of complex designs. SOMs allow them to get to market fast and save development costs."

– Immanuel Rathinam, Associate Director and Head of R&D and Business Operations; iWave Systems Technologies Pvt Ltd.

CHAPTER 9

Summary

SOM USE CASE: AUTOMATED, MULTI-STORY CAR PARKING SYSTEM

Reliability is critical in many applications, with downtime having a high impact on the business operations.

Complex, AI-enabled workloads are increasingly moving to the edge. These applications require a large amount of processing to be performed with low latency, low power consumption, and in a small footprint. To achieve this, the whole application (both the AI and non-AI functions) must be accelerated.

As AI models rapidly evolve, the acceleration platform must also be adaptable. This allows optimal implementation of not just today's AI techniques, but tomorrow's as well.

SOMs provide an ideal edge processing platform. When coupled with adaptive SoCs, the resulting adaptive SOMs provide a comprehensive, production-ready platform for AI-enabled edge applications.

Companies that move to adaptive SOMs benefit from a unique combination of performance, flexibility, and rapid development time. They can enjoy the benefits of adaptive computing without the need to build their own circuit boards, something that has only recently been possible at the edge with the introduction of the AMD Kria™ portfolio of adaptive SOMs.

Kria adaptive SOMs are built around the AMD Zynq™ UltraScale+™ MPSoC architecture and give developers access to a turnkey adaptive computing platform. By standardizing the core parts of the system, developers have more time to focus on building in features that differentiate their technology from the competition.

To learn more about how an adaptive SOM can power your application, please visit the AMD adaptive SOM page at: **xilinx.com/kria**

Find out more:

[Achieving Embedded Design Simplicity with Kria SOMs]

[Kria K26 SOM: The Ideal Platform for Vision AI at the Edge]

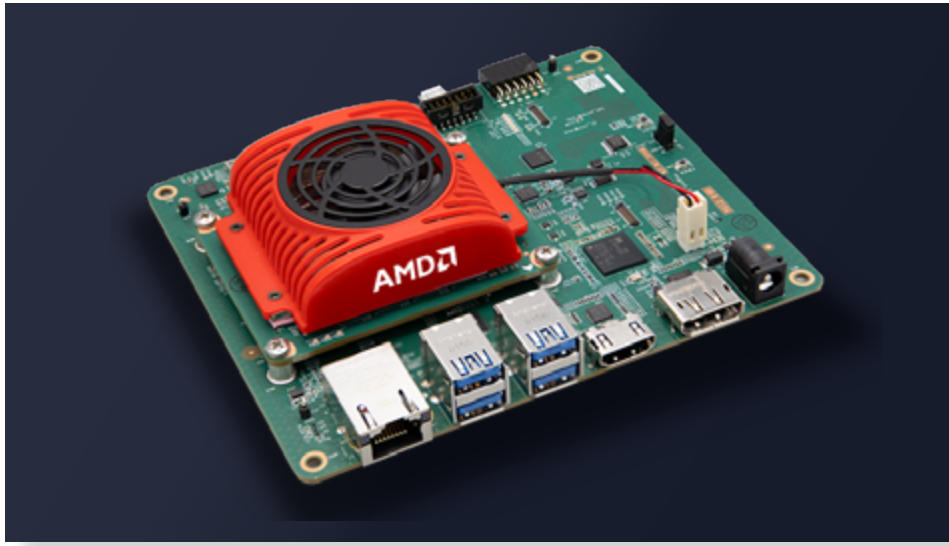


Figure 7. Kria KV260 Vision AI Starter Kit from AMD

Learn more about:

AMD Kria adaptive SOMs

Get in touch with us!

sales@amd.com

AMD