WP505 (v1.1.1) September 29, 2020

# Versal:
## The First Adaptive Compute Acceleration Platform (ACAP)

*Introducing Versal ACAP, a fully software-programmable, heterogeneous compute platform that combines Scalar Engines, Adaptable Engines, and Intelligent Engines to achieve dramatic performance improvements of up to 20X over today's fastest FPGA implementations and over 100X over today's fastest CPU implementations—for Data Center, wired network, 5G wireless, and automotive driver assist applications.*

**ABSTRACT**

Recent technical challenges have forced the industry to explore options beyond the conventional "one size fits all" CPU scalar processing solution. Very large vector processing (DSP, GPU) solves some problems, but it runs into traditional scaling challenges due to inflexible, inefficient memory bandwidth usage. Traditional FPGA solutions provide programmable memory hierarchy, but the traditional hardware development flow has been a barrier to broad, high-volume adoption in application spaces like the Data Center market.

The solution combines all three elements with a new tool flow that offers a variety of different abstractions—from framework to C to RTL-level coding—into an adaptive compute acceleration platform (ACAP). This new category of devices, Xilinx's Versal™ ACAPs, allows users to customize their own domain-specific architecture (DSA) from these three programmable elements.

# Introduction

Recent technical challenges in the semiconductor process prevent scaling of the traditional "one size fits all" CPU scalar compute engine. As shown in Figure 1, changes in semiconductor process frequency scaling have forced the standard computing element to become increasingly parallel [Ref 1].

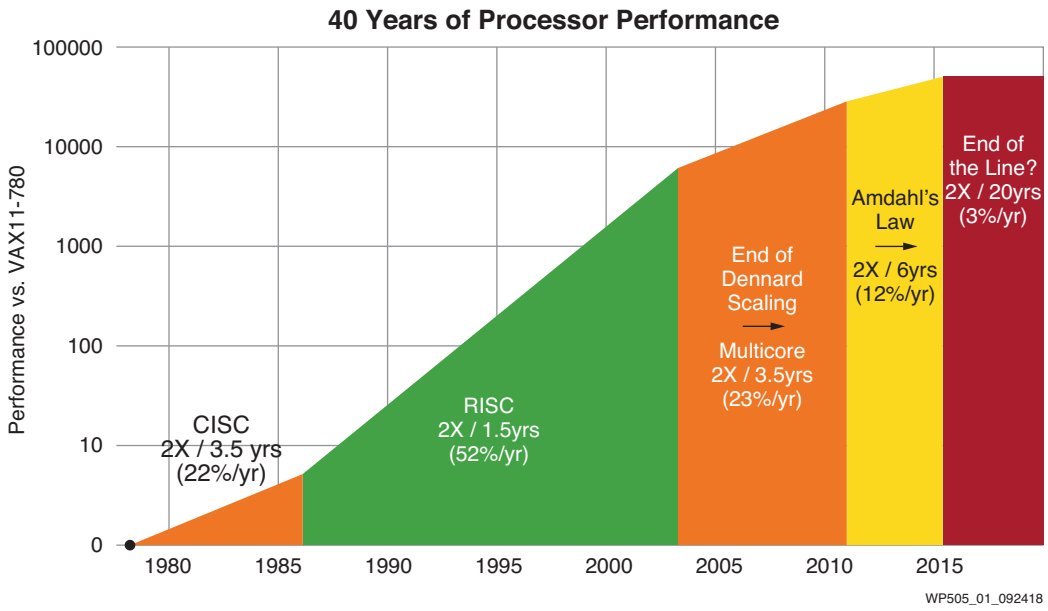


*Figure 1:* **Processor Performance vs. Time**

As a result, the semiconductor industry is exploring alternate domain-specific architectures, including ones previously relegated to specific extreme performance segments such as vector-based processing (DSPs, GPUs) and fully parallel programmable hardware (FPGAs). The question becomes, *Which architecture is best for which task?*

- **Scalar processing elements** (e.g., CPUs) are very efficient at complex algorithms with diverse decision trees and a broad set of libraries—but are limited in performance scaling.
- **Vector processing elements** (e.g., DSPs, GPUs) are more efficient at a narrower set of parallelizable compute functions—but they experience latency and efficiency penalties because of inflexible memory hierarchy.
- **Programmable logic** (e.g., FPGAs) can be precisely customized to a particular compute function, which makes them best at latency-critical real-time applications (e.g., automotive driver assist) and irregular data structures (e.g., genomic sequencing)—but algorithmic changes have traditionally taken hours to compile versus minutes.

See Figure 2.

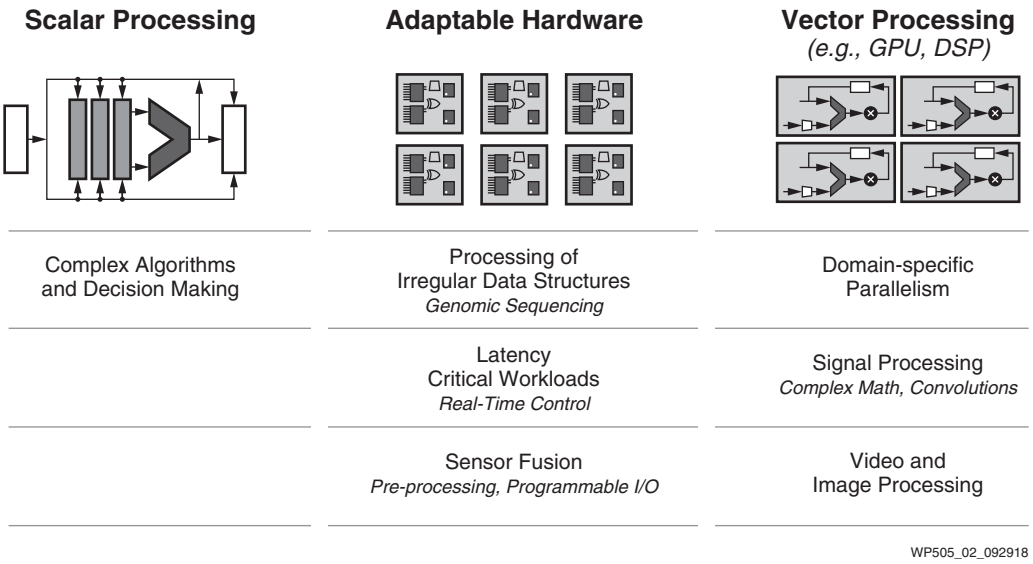


*Figure 2:* **Types of Compute Engines**

To answer the question, Xilinx is introducing a revolutionary new heterogeneous compute architecture, the adaptive compute acceleration platform (ACAP), which delivers the best of all three worlds—world-class vector and scalar processing elements tightly coupled to next-generation programmable logic (PL), all tied together with a high-bandwidth network-on-chip (NoC), which provides memory-mapped access to all three processing element types. This tightly coupled hybrid architecture allows more dramatic customization and performance increase than any one implementation alone. See Figure 3.

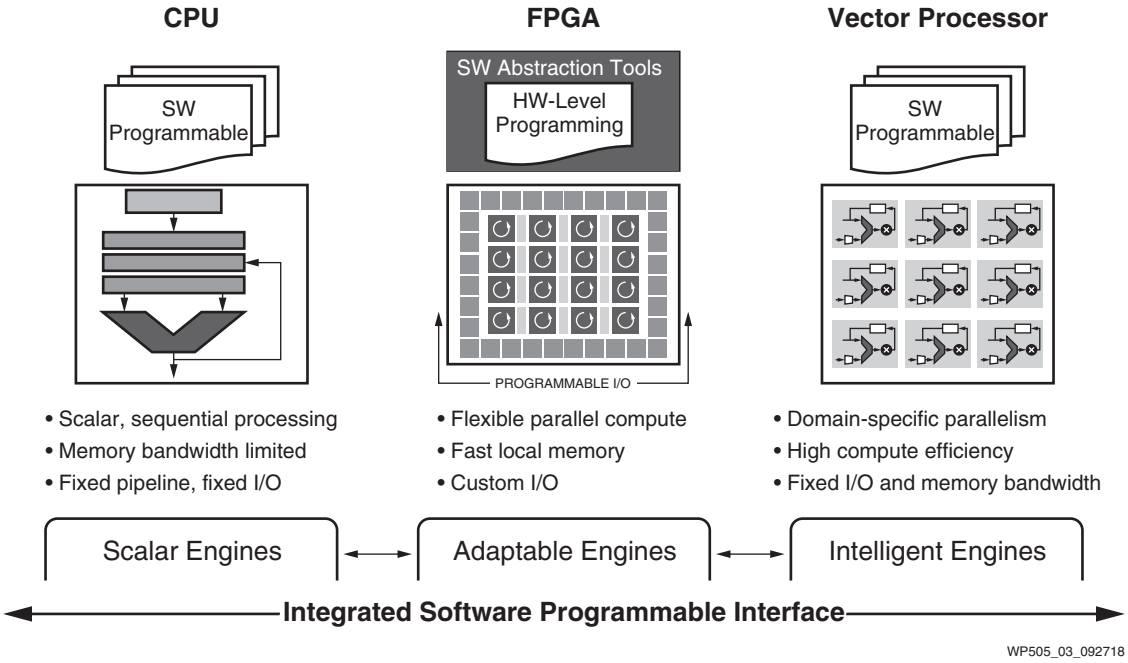


*Figure 3:* **Heterogeneous Integration of Three Types of Programmable Engines**

Such a dramatic increase in performance necessitates a similarly dramatic improvement in tools focusing on ease of use. ACAPs are specifically designed to work out of the box with no RTL flow required. ACAPs are natively software programmable, enabling C-based and framework-based design flows. The devices have an integrated shell that comprises a cache-coherent host interface (PCIe® or CCIX technology) with integrated DMA, a NoC, and integrated memory controllers, eliminating the requirement for RTL work.

The new ACAP architecture also yields a dramatic improvement in ease of use. It provides a fully integrated, memory-mapped platform for programming through a unified toolchain. The Xilinx toolchain supports multiple entry methods for every type of developer. For example, certain applications (such as AI machine learning inference) can be coded at the framework level (e.g., Caffe, TensorFlow); others can be coded in C using pre-optimized libraries (e.g., filters for 5G radio). Traditional hardware developers can still port their existing RTL to ACAP via the traditional RTL entry flow.

This white paper reviews the needs driving the change from the traditional CPU-based compute model, explores the other options in detail, and unveils the Xilinx Versal ACAP, the industry's first heterogeneous compute platform.

The three major advantages of an ACAP include:

1. **Software Programmability**—The ability to quickly develop optimized applications through software-abstracted toolchains.
2. **Acceleration**—Metrics for a wide range of applications from artificial intelligence, smart network interface cards, high density storage, 5G wireless, self-driving cars, advanced modular radar, and terabit optical networks.
3. **Dynamically Adaptable Reconfiguration**—The ability to reconfigure the hardware to accelerate new loads within milliseconds.

# Introducing ACAP: Hardware and Software Optimized for Parallel Heterogeneous Computation

ACAPs feature a mix of next-generation Scalar Engines, Adaptable Engines, and Intelligent Engines. The NoC connects them all together via a memory-mapped interface with an aggregate bandwidth of 1Tb/s+. In addition to the NoC, the massive memory bandwidth enabled by programmable logic (and integrated RAM blocks) enables programmable memory hierarchies optimized for individual compute tasks (avoiding the high latency and latency uncertainty inherent in other cache-based compute units). See Figure 4.

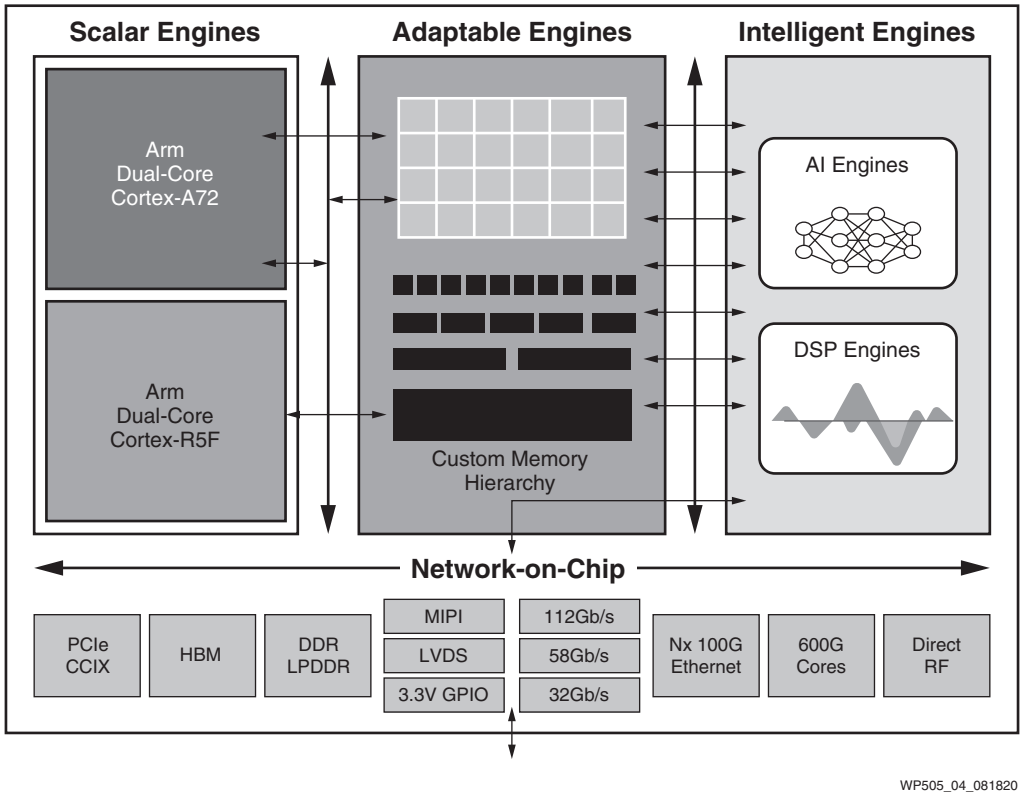


*Figure 4:* **Xilinx Versal ACAP Functional Diagram**

The Scalar Engines are built from the dual-core Arm® Cortex-A72, providing a 2X increase in per-core single-threaded performance compared to Xilinx's previous-generation Arm Cortex-A53 core. A combination of advanced architecture and power improvements from the 7nm FinFET process yield a 2X improvement in DMIPs/watt over the earlier 16nm implementation. The ASIL-C certified[1] UltraScale+™ Cortex-R5F Scalar Engines migrate forward to 7nm with additional system-level safety features based on learning from Xilinx's current automotive volume deployments.

The Adaptable Engines are made up of programmable logic and memory cells connected with the next generation of the industry's fastest programmable logic. In addition to supporting legacy designs, these structures can be reprogrammed to form memory hierarchies customized to a particular compute task. This allows Xilinx's Intelligent Engines to achieve a much higher cycle efficiency and a much higher memory bandwidth per unit compute than the latest GPUs and CPUs. This is key to optimizing for latency and power at the edge and for optimizing for absolute performance in the core.

The Intelligent Engines are an array of innovative very long instruction word (VLIW) and single instruction, multiple data (SIMD) processing engines and memories, all interconnected with 100s of terabits per second of interconnect and memory bandwidth. These permit 5X–10X performance improvement for machine learning and digital signal processing (DSP) applications.

These compute functions are mixed in different ratios and magnitudes to form the Versal portfolio of devices, as shown in Table 1.

1. https://www.xilinx.com/news/press/2018/xilinx-announces-availability-of-automotive-qualified-zynq-ultrascale-mpsoc-family.html

*Table 1:* **Versal Portfolio Devices, Markets, and Key Features**

| Versal Portfolio | Primary Markets | Key Features |
|---|---|---|
| Versal AI Core | Data Center, Wireless | Maximum intelligent engine compute |
| Versal AI Edge | Automotive, Wireless, Broadcast, A&D | Efficient intelligent engine count within tight thermal envelopes down to 5W |
| Versal AI RF | Wireless, A&D, Wired | Direct RF converters and SD-FEC |
| Versal Prime | Data Center, Wired | Baseline platform with integrated shell |
| Versal Premium | Wired, Data Center, A&D, Test and Measurement | Premium platform with maximum adaptable engines, 112G SerDes and 600G Integrated IP |
| Versal HBM | Data Center, Wired, Test and Measurement | Premium platform with HBM |

The Xilinx adaptive compute acceleration platform (ACAP) blends vector, scalar, and adaptable hardware elements to offer three compelling advantages:

- Software Programmability
- Heterogeneous Acceleration
- Adaptability

# Software Programmability

## Adaptable Acceleration Enabled by Adaptable Silicon

Versal ACAPs provide adaptable acceleration hardware that is easily programmable in software. The heterogeneous engines allow optimal acceleration of software applications, regardless of application type. The Intelligent Engine accelerates machine learning and common classical DSP algorithms. The next-generation programmable logic inside the Adaptable Engine accelerates parallelizable algorithms. The multi-core CPU provides comprehensive embedded compute resources for the remaining application needs. The entire Versal device is designed to be easily programmable using software without requiring hardware expertise. See Figure 5.

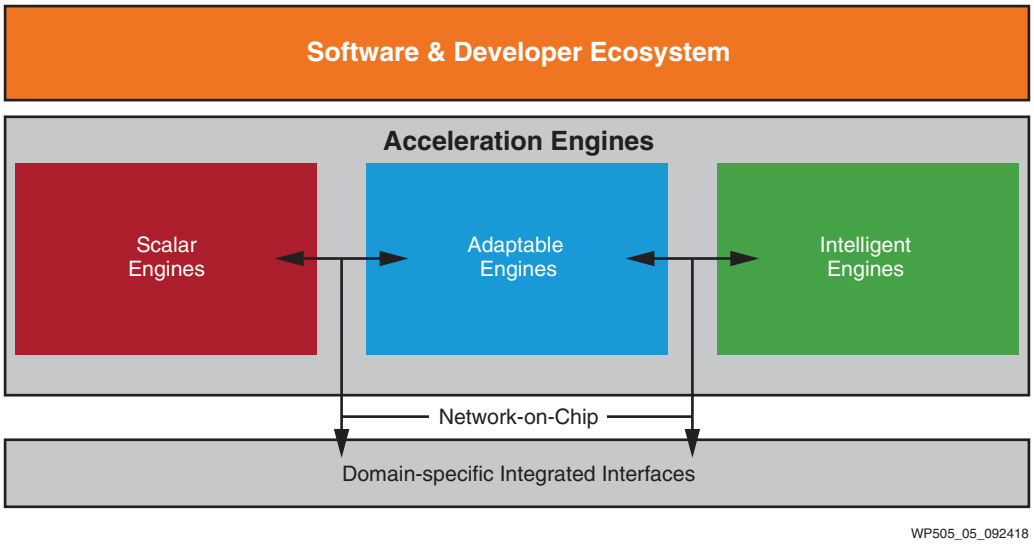


*Figure 5:* **Versal ACAP Top-Level Concept Diagram**

- Data and AI scientists can deploy applications built in a standard software framework, accelerating them by orders of magnitude using a Versal ACAP.
- Software application developers can accelerate any software application with a Versal ACAP, without needing hardware expertise, using Xilinx's unified software development environment.
- Hardware designers can continue to design with Vivado® Design Suite, while benefiting from reduced development time using the Versal platform's integrated I/O interfaces and NoC interconnect.

See Figure 6.

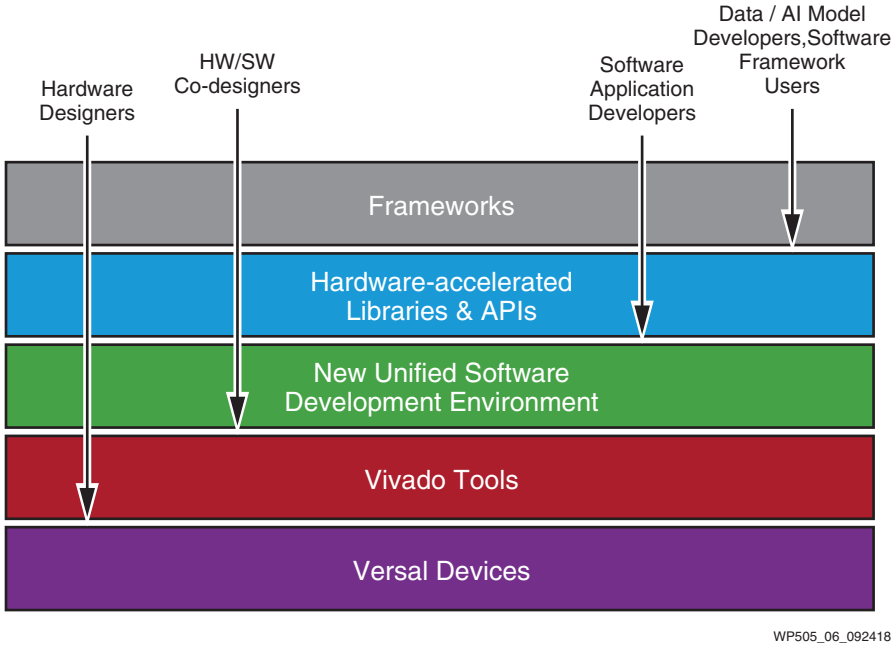


*Figure 6:* **Versal Platform Software Persona Concept**

## Dedicated Hardware to Improve Ease of Use and Application Efficiency

The adaptable interface logic provides easy access to off-chip interfaces. This includes a standard interface to an external host processor. In Data Center applications, the software application typically resides on a host CPU rather than the embedded microprocessors. The interface connecting the host CPU and the Versal platform's programmable resources is called the shell. The integrated shell includes a fully compliant cache coherent interconnect for accelerators (CCIX) or PCIe Gen4x16 interface to the host, DMA controllers, a cache coherency memory, integrated memory controllers, advanced functional safety, and security features.

The NoC makes each hardware component and the soft-IP modules easily accessible to each other, as well as to the software via a memory-mapped interface. It provides a standardized, scalable hardware framework, enabling efficient communication between the heterogeneous engines and the interface logic.

# Heterogeneous Acceleration

While programmable logic (FPGA) and vector-based (DSP, GPU) implementations have recently demonstrated great performance gains over CPUs, the true benefit of the ACAP architecture only comes into focus when developers leverage more than one type of the Versal ACAP's compute element to enable a tightly coupled compute model. In this case, 1+1+1 can be much more than 3.

Table 2 summarizes the benefits of Versal ACAP devices for various markets.

*Table 2:* **Versal ACAPs and Targeted Markets**

| Market | Benchmark | vs. CPU | vs. GPU | vs. FPGA | Notes |
|---|---|---|---|---|---|
| Data Center | Image Recognition (Inference) - Latency insensitive | 43X | 2X | 5X | GoogLeNet v1 (unlimited Batch size) |
| | Image Recognition (Inference) - 2ms latency | n/a | 8X | 5X | GoogLeNet v1 (< 2ms) CPU latency floor 5ms |
| | Risk Analysis | 89X | n/a | >1X | Value at Risk (VaR) for Interest Rate Swaps Maxeler result |
| | Genomics | 90X | n/a | >1X | Human Gene Analysis Edico Genome result |
| | Elastic Search | 91X | n/a | >1X | 91X lower latency BlackLynx result on 1TB data |
| Wireless 5G | 16x16 5G Remote Radio | n/a | n/a | >5X | >5X more radio bandwidth for 5G Remote Radio |
| | Beam Forming | n/a | n/a | >5X | >5X more compute |
| A&D Radar | DSP TMACs | n/a | n/a | >5X | Over 27 TMACs |
| | Algorithm iteration time | n/a | n/a | >100X | Software programmable intelligent engines compile in minutes |
| Automotive | Low latency inference (<2ms) | n/a | 3x | 15X | ResNet50 Batch=1 AI engine scales better to low latency safety critical ADAS and Autonomous Driving |
| | Enclosure types | 1 | 2 | 4 | ACAP portfolio breadth only one that covers <10W, 20W, 30W, and trunk-mounted enclosures efficiently |
| Wired | Encrypted Network traffic | n/a | n/a | 4X | ACAP integration of networking and encryption IP enables multi-terabit single-chip implementations |

# Data Center Artificial Intelligence: Machine Learning Inference Acceleration

As artificial intelligence starts to pervade modern life, the demand for enhanced compute efficiency has begun to drive innovation in the semiconductor space—but it is difficult for any single implementation to handle with maximum efficiency. This is one area where the tight coupling between vector processing and programmable hardware is invaluable.

There has been a lot of attention on the precision of the compute unit (FP32 versus FP16 versus INT16 versus INT8, etc.), but inattention to the divergence in memory hierarchy requirements between network types has caused many of the most recent AI inference engines to drop sharply in efficiency for different networks. For example, the current state-of-the-art machine learning inference engines require four HBM memories (7.2Tb/s of external memory bandwidth) to reach their peak performance—but their cache-based memory hierarchy only operates at around 25–30% efficiency and creates significant latency uncertainty for real-time applications. The solution is to augment the vector processing performed by intelligent engines with a programmable memory hierarchy, precisely optimized for each network type and enabled by the massive parallelism of FPGA logic.

For example, a Versal platform implementation of GoogLeNet enables extraordinarily high performance for latency insensitive applications, 43X more throughput than today's top-of-the-line Skylake Platinum CPU(2), and about 3X today's top-of-the-line GPU [Ref 2]—all at much lower power. See Figure 7.

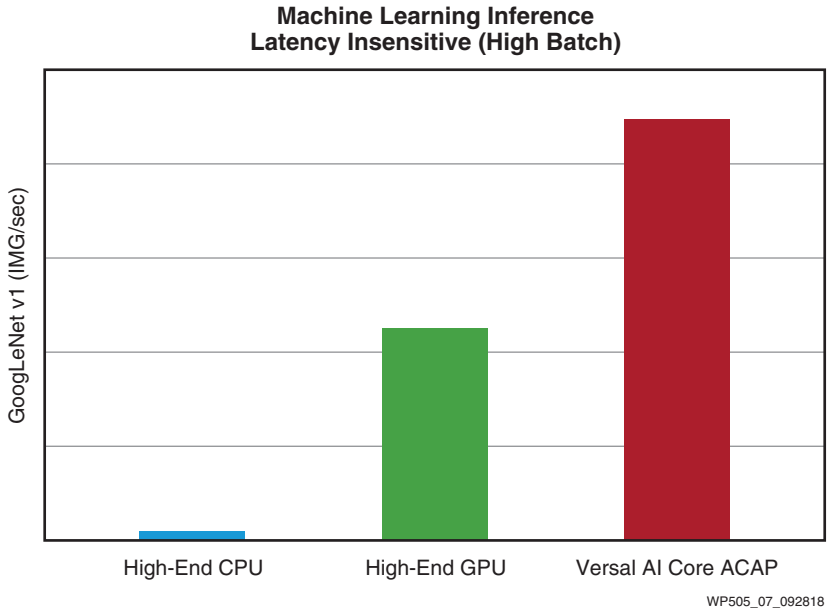


*Figure 7:* **GoogLeNet Performance (<7ms Latency) = 43X Higher Than a High-End CPU[1, 2]**

1. Measured on Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe.
2. V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services."

2. Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Canonical, Ubuntu, 16.04LTS, AMD64 Xenial image build on 2018-08-14, Intel Caffe. Git Version: a3d5b02, `run_benchmark.py` unmodified.

As Data Center applications for neural networks continue to get more advanced, multiple neural networks can be chained together, greatly increasing the need for low latency neural network performance. For example, real-time spoken word translation requires speech-to-text, natural language processing, a recommender system, text-to-speech, and then speech synthesis [Ref 2]. This means the neural network's portion of the total latency budget is multiplied by 5 for this application.

As the number of real-time applications continues to increase, it is important for Data Center customers to choose a technology that can scale to keep up with their future needs. Two trends are emerging:

- Deterministic latency is becoming increasingly important to improve software design efficiency [Ref 3].
- Neural network latency requirements continue to tighten as increasingly complex interactions are modeled (human interaction, financial trading), and safety-critical applications rise in importance (e.g., automotive, industrial).

These two requirements necessitate the removal of batching, which causes the performance of fixed, cache-based memory hierarchy of CPU and GPU-based solutions to degrade significantly. Even a high-end CPU caps out at 5ms latency, and below 7ms, even high-end GPUs degrade significantly in performance. Only the Versal ACAP achieves sub-2ms latency with acceptable performance. See Figure 8.

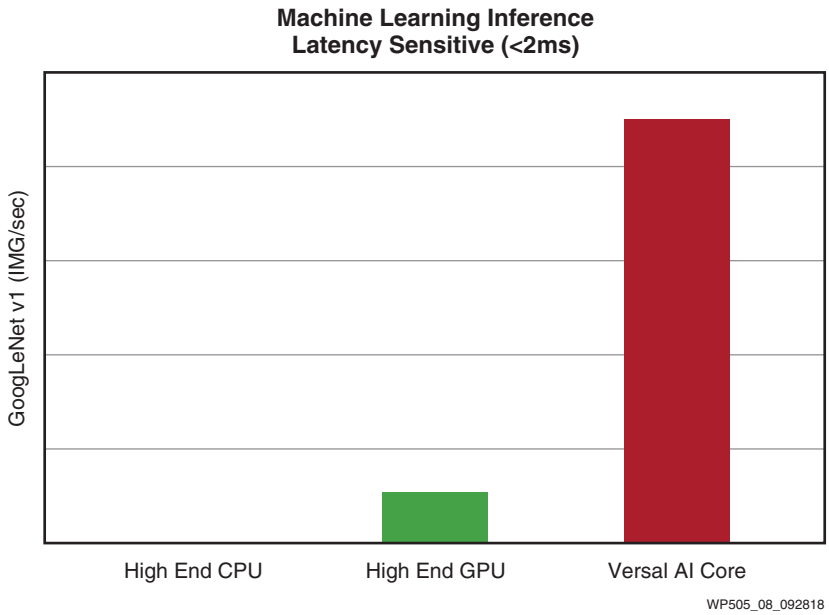


*Figure 8:* **Real-Time GoogLeNet Performance (<2ms Latency) = 8X Higher Than High-End GPU (Nvidia)[1, 2]**

1. Measured on Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe.
2. V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services."

As a result, the unique programmable memory hierarchy of ACAP-based solutions offers both the highest performance machine learning inference performance as well as unmatched scalability as future applications demand lower and more deterministic latency.

### *Data Center SmartNICs*

Network interface cards (NICs) began as simple connectivity. Over time, they became "SmartNICs" by adding additional network acceleration (encryption, hypervisor networking offload, virtual switching). Amazon had great success with their Annapurna project; it offloaded all hypervisor functions from their CPUs, enabling 100% of CPU cycles to be devoted to revenue-generating compute.

As SmartNICs evolve, Xilinx expects three new benefits to emerge: the ability to dynamically distribute and scale workloads over the Data Center Ethernet logic, the ability to have reconfigurable pools of acceleration that can run any compute acceleration function (maximize utilization of cloud resources), and finally, the ability to run compute functions inline with the network data plane.

Xilinx Versal ACAP devices enable the integration of NIC functionality with hybrid vector-based and programmable logic compute engines, all supported by Xilinx's deep portfolio of networking IP and world-class SerDes, including single channel 112G SerDes for next-generation NIC to TOR (top of rack) links.

Furthermore, these NIC resources can be dynamically reconfigured or redeployed on new workloads.

*Table 3:* **Types of Data Center Network Interface Cards**

| | Description | Features | Examples |
|---|---|---|---|
| **Type 1** | Basic Connectivity NIC | • Basic Offloads (Checksum, LSO, RSS)<br>• Single Root I/O Virtualization<br>• Some tunnel offloads (VXLAN, GRE0) | • Fortville<br>• ConnectX<br>• NetExtreme |
| **Type 2** | SmartNIC for Network Acceleration | • Encryption/Decryption (IPSec)<br>• Virtual Switch Offload (OVS, etc.)<br>• Programmable Tunnel Types | • **Xilinx Type 2**<br>• LiquidIO<br>• Annapurna<br>• Innova |
| **Type 3** | SmartNIC for Network Compute Acceleration | • Inline Machine Learning<br>• Inline Transcoding for Video<br>• Database Analytics<br>• Storage (Compression, Crypto, Dedupe) | • **Xilinx Type 3**<br>• MSFT (NIC + FPGA) |

## Data Center Storage Acceleration

FPGAs have long been used in storage drives to perform error correction and write-leveling tasks. Their flexible I/O allows superior design reuse, especially critical in the fast-moving world of flash technologies. In addition, many current database search and acceleration appliances already benefit from having FPGA-based acceleration next to the drives. (Maximal efficiency is achieved by placing the compute element directly next to the drives.)

With ACAP architecture, drive and DB acceleration vendors can add machine learning compute directly inside the drive (where FPGAs are already used), reducing data movement across the Data Center (and the associated latency, power consumption, and OpEx) by a factor of 10.

## 5G Wireless Communications

The insatiable hunger for wireless subscriber bandwidth has led to a blistering "10X every 10 years" pace of innovation in the wireless industry. At the 2020 Olympics, the industry will begin initial public demonstrations of the 5th generation of wireless technology, called "5G". The majority of these initial implementations will be built from existing Xilinx devices—particularly the wildly successful production 16nm RFSoC device, which delivers three key advantages:

- Integrated direct RF sampling rate ADCs and DACs
- Integrated LDPC and turbo soft-decision forward error (SD-FEC) correction code blocks
- Power-efficient DSP due to 16nm FinFET process technology

As the industry ramps to volume, two challenges emerge: the drive to wider spectrum at lower cost and the addition of machine learning inference technologies in the radio to enhance beam steering algorithms, enhance subscriber hand-off algorithms, and enable self-healing networks.

Traditionally, some wireless vendors have reduced cost by implementing vector-DSP-based ASICs. The inclusion of an intelligent engine into Versal ACAPs largely eliminates the traditional cost gap between ASICs and FPGAs, because it delivers 5X more single-chip TMACs. See Figure 9.

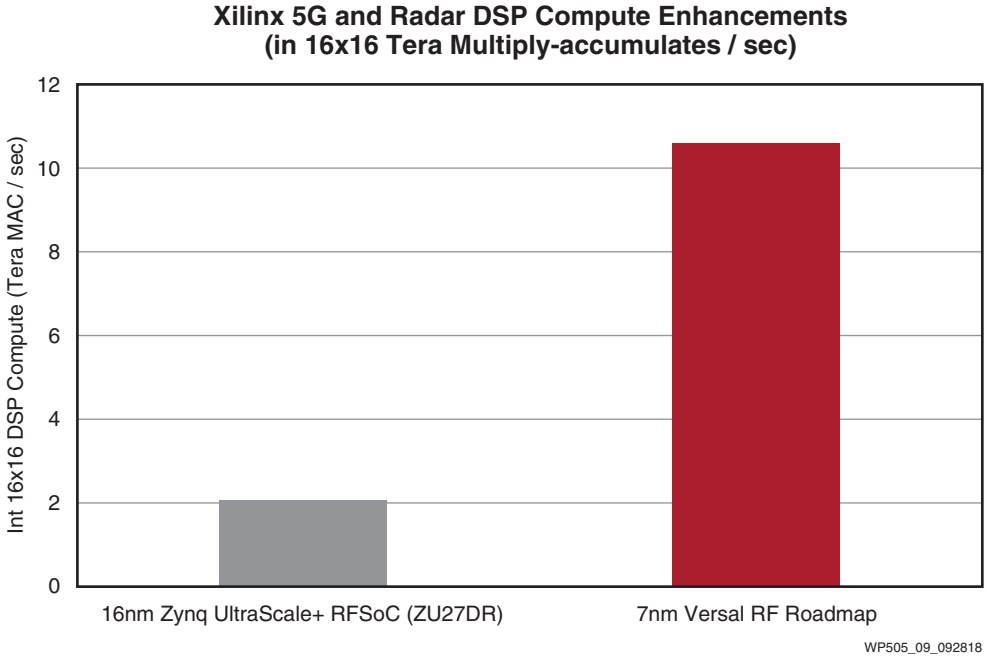


*Figure 9:* **Xilinx RF Compute Roadmap**

As a result, while a 16nm Zynq® UltraScale+ RFSoC can implement a 200MHz 16x16 remote radio unit (RRU), the 7nm Versal device roadmap can implement a full 800MHz 16x16 RRU. See Figure 10.

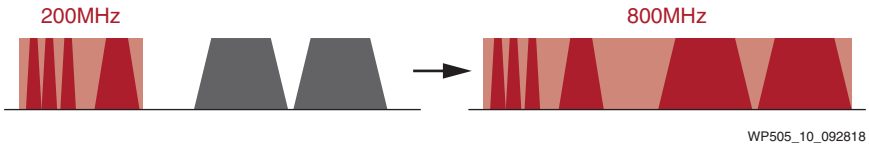


*Figure 10:* **Single-Chip Spectrum Possible with 16nm vs. 7nm Wireless Radio Devices**

The addition of power-efficient machine learning (with framework-level design flow) puts the ACAP-based Versal portfolio in a class by itself. This technology can enhance beam steering and subscriber hand-off algorithms by an additional factor of two over traditional programmatically-defined algorithms, approaching 85% of the theoretical limit. See Figure 11.

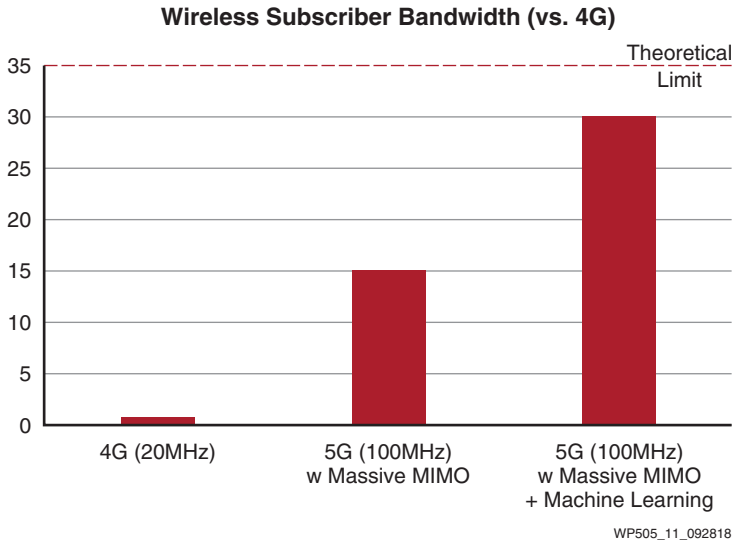


*Figure 11:* **Wireless Bandwidth Optimization vs. Theoretical Limit**

Xilinx is the only vendor in the industry to merge all four of these key technologies into a single chip: direct-RF sampling ADC and DAC, integrated SD-FEC codes, high-density vector-based DSP, and framework-programmable machine learning inference engines—the industry's first true 5G radio-on-chip.

## Aerospace and Defense

The massive parallel DSP capabilities of FPGAs have long been the backbone of many defense radar implementations. However, recent technical innovations in ADC technology have boosted ADC sample rates to 100s of gigasamples per second, requiring a commensurate increase in DSP capability.

The merger of powerful vector-based DSP engines with AI machine learning enables revolutionary new products in the aerospace and defense industries, such as advanced modular radar. The antenna spacing driven by high-frequency wavelengths dictates extremely small form factors. Xilinx offers devices with terabits per second of antenna bandwidth and up to 17 TMACs of INT24, or 24TFLOPS of 32-bit single-precision floating point DSP in a single package.

## Automotive Driver Assist (ADAS)

Xilinx has a long history in high reliability and thermally constrained systems in automotive, aerospace, satellite, medical, and commercial networking systems. Xilinx technology is designed up front to mitigate SEU effects and operate up to a temperature of 125°C. Combined with a focus on machine vision and machine learning, a heritage of reliability and quality makes Xilinx technology a natural fit for automotive driver assist (ADAS) and future self-driving car technology. To-date, Xilinx has shipped over 150 million FPGAs and SoCs into various automotive sockets, and over 50 million units specifically into ADAS applications. Automotive is Xilinx's fastest growing market segment over the last two years.

Xilinx's scalable set of automotive-targeted Versal ACAPs merge a power-efficient Scalar Engine with dual-core Cortex-R5Fs, programmable I/O, and low latency, intelligent AI Engines that enable power-efficient, functionally safe, AI-enhanced automated driving solutions with 15X more INT8 machine learning performance vs. today's FPGA-based ASIL-C certified(3) ADAS solutions. Furthermore, the ability to reprogram the entire device via over-the-air hardware updates improves system in-field versatility, which adds customer value. Lastly, Xilinx's programmable I/O gives vendors the flexibility and adaptability to change sensor types without the delays and cost of waiting for ASSP or GPU re-spins. See Figure 12.

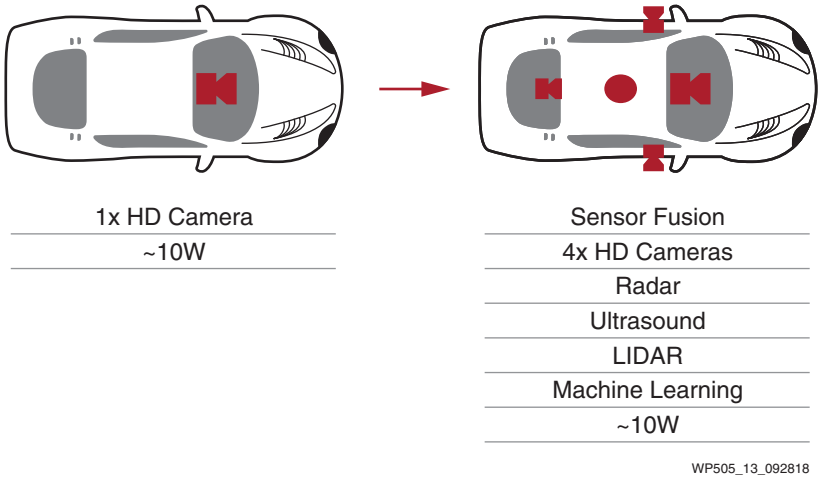


*Figure 12:* **Xilinx ACAP Devices Enable Sensor Fusion in Small Power Envelopes**

With all the innovation occurring in the automotive space, it is important to choose a processing device portfolio that offers code portability and scalability across multiple platforms, from 5–10W wind-screen mounted front camera designs to 20–30W in-cabin central modules to 100W+ liquid cooled trunk-mounted supercomputers, all with the same programming model. See Table 4.

*Table 4:* **Xilinx Automotive Product Breadth vs. Competition (Same Programming Model)**

| | (10W) Intelligent Endpoint (e.g., Front Camera) | (20W) Central Module (Basic, Passively Cooled) | (30W) Central Module (Advanced, Forced Air) | (100W+) In-trunk Supercomputer (Liquid Cooled) |
|---|---|---|---|---|
| Xilinx | ● | ● | ● | ● |

3. https://www.xilinx.com/news/press/2018/xilinx-announces-availability-of-automotive-qualified-zynq-ultrascale-mpsoc-family.html

*Table 4:* **Xilinx Automotive Product Breadth vs. Competition (Same Programming Model)**

| | (10W) Intelligent Endpoint (e.g., Front Camera) | (20W) Central Module (Basic, Passively Cooled) | (30W) Central Module (Advanced, Forced Air) | (100W+) In-trunk Supercomputer (Liquid Cooled) |
|---|---|---|---|---|
| Nvidia | | ❍ | ● | ● |
| Intel MobilEye | ● | | | |

Latency is an especially critical processing performance factor when considering vehicles traveling at automotive speeds. At 60MPH (100KPH), 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. 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 reliant on high batch sizes. Therefore, Xilinx has optimized the AI Edge series to operate at extremely high efficiency at low batch sizes. See Figure 13.

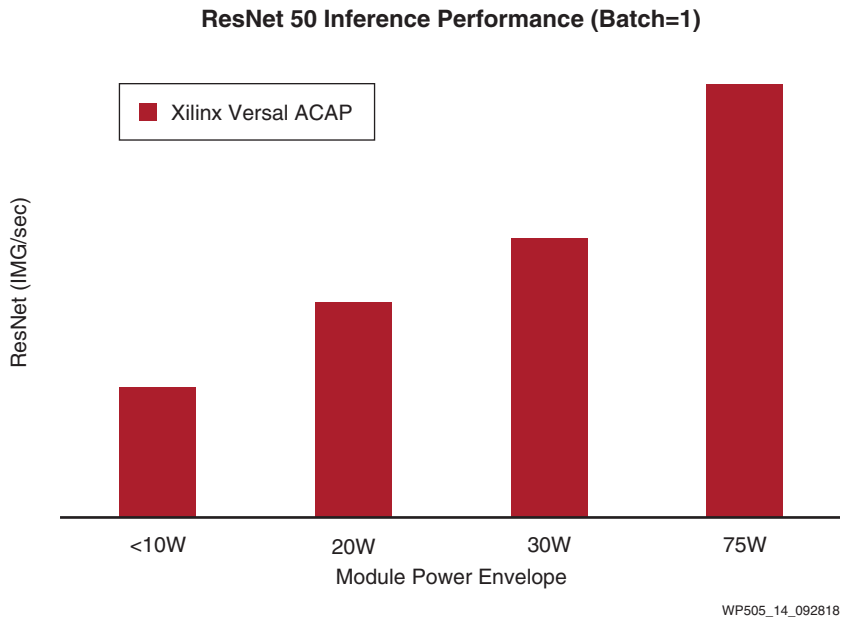


*Figure 13:* **Breadth of Low Latency Safety Critical Versal Portfolio**

Today's automotive ADAS/AD systems are demanding an increasing number of high-resolution cameras. Compute requirements scale with pixels, which means an image from an HD camera (1080x1920) requires significantly more compute vs. a Data Center standard 224x224 image. The high compute efficiency of Xilinx Versal devices are uniquely positioned to scale to meet the demands of higher resolutions.

## Wired Communications

Every piece of internet traffic today crosses multiple Xilinx FPGAs. FPGAs have long served as "glue logic" to enable network hardware to adapt to the changing needs of network operators. Xilinx's leadership in the most advanced 112G SerDes technology enables the industry's first implementation of new protocols and challenging optical, copper cable, and backplane standards, as well as current 58G PAM4 and 32G NRZ protocols like pre-standard PCI Express® Gen5. The

extensive portfolio of IP has enabled integration of standardized interfaces and has driven down cost and power. Xilinx's extensive portfolio of IP allows customers to mix and match, thereby differentiating at the hardware level.

As network operators continue to demand new features, the ability to quickly code and field-update adaptable hardware offers competitive advantage versus those stuck with legacy ASSPs.

Xilinx Versal ACAPs feature groundbreaking levels of integrated IP aligned with next-generation 600G wavelength plans, with full support for Ethernet and OTN standard 10G, 25G, 50G, and 100G SerDes rates, including:

- 10/25/40/50/100GE MAC/PCS/FEC with ±1ns IEEE Std 1588 timestamp, eCPRI, and TSN support
- 600G FlexE core channelizable down to 10G channels and high-density 400GE/200GE/100GE MAC/PCS/FEC
- 600G wire-rate encryption engine supporting MACSEC and IPSEC as well as bulk AES-GCM encryption
- 600G Interlaken with integrated FEC for PAM4 lanes
- SD-FEC for DOCSIS Cable LDPC applications

These dramatic improvements in SerDes enable:

- Single-chip 1.0Tb/s+ network line cards for OTN and edge router applications at similar power and superior flexibility versus commercial ASSPs
- Single-chip 2.4Tb/s+ encrypted Data Center interconnect (DCI) rack-mountable appliances, multiple instances per RU (see Figure 14)
- 400Gb/s+ cable modem termination systems (CMTS) with per-subscriber encrypted tunnels for advanced business and residential services

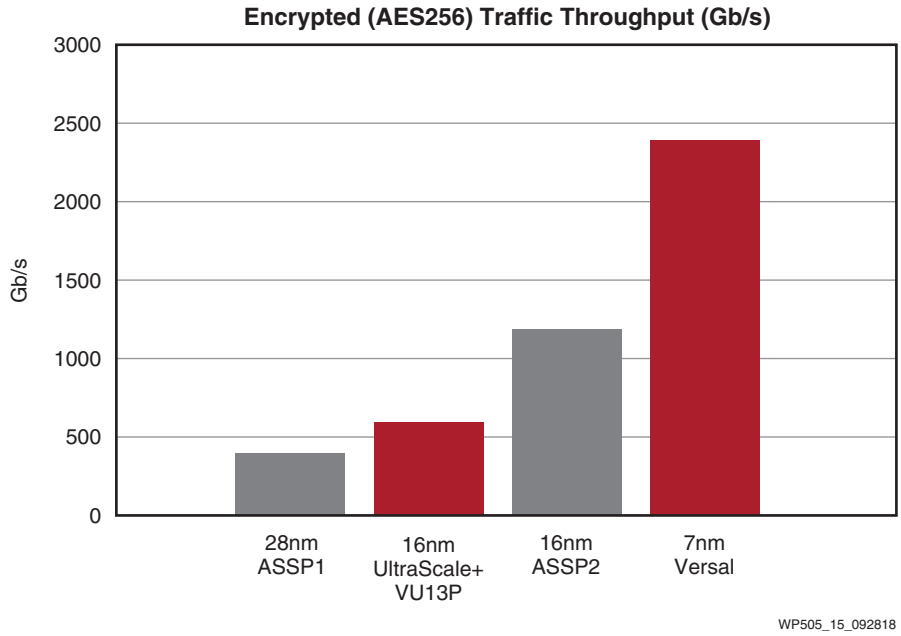


*Figure 14:* **Wired Communications: Single-Chip Encrypted Data Center Traffic[1, 2]**

1. Microsemi DIGI-G4 OTN ASSP: https://www.microsemi.com/product-directory/multi-service-otn-processors/4227-pm5990-digi-g4
2. Microsemi DIGI-G5 OTN ASSP: https://www.microsemi.com/product-directory/multi-service-otn-processors/5056-pm6010-digi-g5-otn-processor

# Adaptability

One of the primary benefits of programmable logic technology is that in-field hardware upgrades are possible. This is widely deployed today in 4G wireless and optical networks, as well as automotive autonomous driving products.

Xilinx Versal ACAPs extend this in-field upgrade functionality by enabling a higher level of abstraction (C or framework-level interfaces) and 8X faster partial reconfiguration, enabling much faster kernel swapping.

## Adaptable Hardware

The core value proposition of FPGAs has long been the ability to change designs in the field. Whether to correct bugs, optimize algorithms, or add completely new features, programmable logic offers unique flexibility versus all other semiconductor options.

Xilinx Versal ACAPs take this concept a step further, speeding up configuration time by nearly an order of magnitude, enabling dynamic swapping of partial bitstreams in milliseconds—hardware with the agility of software.

## Programmable Memory Hierarchy

Adaptable hardware augments Versal ACAPs by serving as a complement to optimize the efficiency of the new features of the ACAP architecture.

One of the biggest advantages of programmable logic is the ability to reconfigure memory hierarchy and thus to optimize for different compute loads. For example, even within the scope of neural networks focused on image recognition, the memory footprint and compute operations per image vary widely depending on the algorithm. Programmable memory hierarchy allows the programmable logic to be adjusted to optimize compute efficiency for each network it supports.

As a result, when implementing neural networks via a combination of vector processors and programmable logic, Versal ACAPs can achieve compute efficiencies of nearly 2X the leading GPUs, which implement vector processing with a fixed memory hierarchy. See Figure 15.

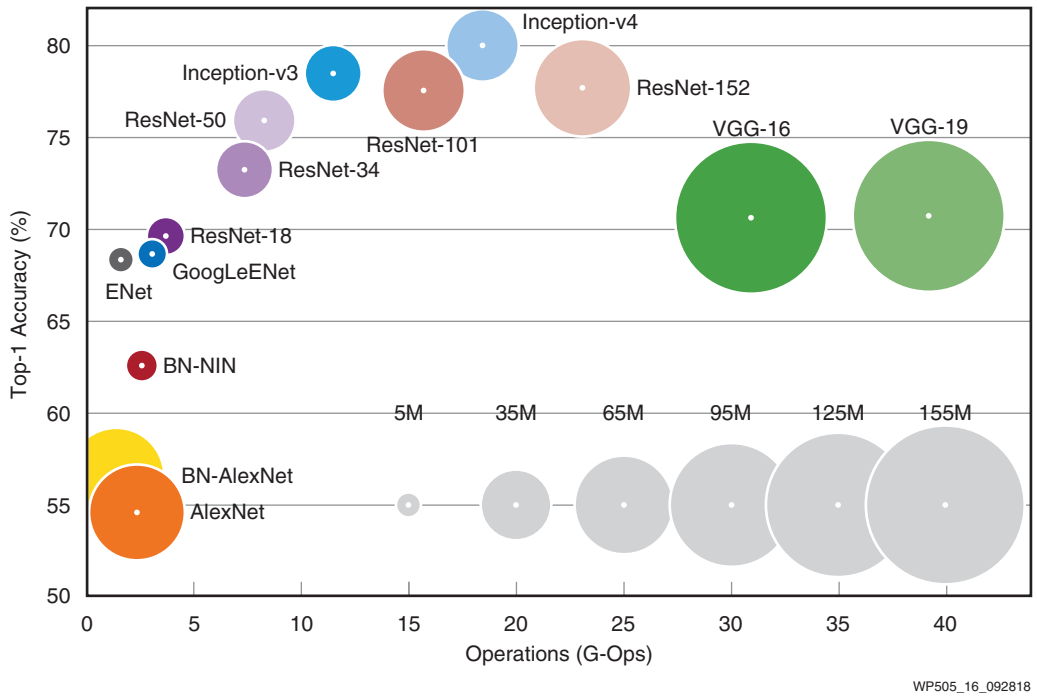


*Figure 15:* **Memory Utilization vs. Compute Operations by Neural Network Type**

## Dynamic Reconfiguration

Certain cost-sensitive, real-time applications can benefit from utilizing the device's inherent programmability to multiplex one set of programmable hardware between multiple logical functions with sub-millisecond Adaptable Engine partial reprogramming time. In the Data Center, this allows Versal ACAP devices to perform a much wider array of functions traditionally performed by a CPU when compared to a more limited vector processor like a GPU. (See Figure 16, [Ref 4])

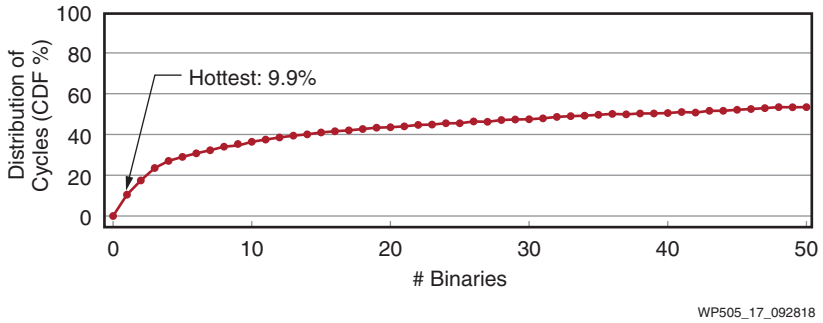


*Figure 16:* **No "Killer App"—Data Center Workloads Are Widely Distributed (Kanev)**

# Summary

Recent technical challenges have forced the industry to explore options beyond the homogeneous "one size fits all" CPU scalar processing solution. Vector processing (DSP, GPU) solves some problems but runs into traditional scaling challenges due to inefficient memory bandwidth usage. Traditional FPGA solutions provide programmable memory hierarchy, but the traditional hardware flow has been a barrier to adoption.

The solution is to combine all three elements with a new tool flow that offers a variety of different abstractions—from framework to C to RTL-level coding: an adaptive compute acceleration platform (ACAP).

The ACAP architecture significantly extends the capabilities of programmable logic alone. The hybrid of programmable logic and vector processing elements enables a disruptive increase in compute for applications in Data Center, wireless networking, automotive driver assist, and wired communications.

The merger of formidable AI machine learning compute, advanced networking, and encryption IP enables a new class of adaptable compute acceleration engines and SmartNICs in the Data Center.

The merger of pre-made artificial intelligence machine learning inference with dense DSP and direct-RF sampling ADC/DACs doubles the throughput of 5G wireless radios over in-house DSP-based ASICs, enabling single-chip sensor fusion between LIDAR, radar, and visual sensors in automotive driver assist (ADAS) applications.

Find out more about Xilinx's Versal portfolio of ACAP devices on the Xilinx website at: https://www.xilinx.com/products/silicon-devices/acap/versal.html.

# References

1. J. Hennessy, D. Patterson, *Computer Architecture: A Quantitative Approach* (6th Edition, 2019).
2. Nvidia, *Nvidia AI Inference Platform: Giant Leaps in Performance and Efficiency for AI Services, from the Data Center to the Network's Edge* (2018). Retrieved from nvidia.com, 2018.
3. N. Jouppi, C. Young, N. Patil, et al., In-Datacenter Performance Analysis of a Tensor Processing Unit™. In *International Symposium on Computer Architecture* (ISCA 2017). Retrieved from arxiv.org, 2018.
4. S. Kanev, J. Darago, K. Hazelwood, et al., *Profiling a warehouse-scale computer* (2015). Retrieved from google.com, 2018.

# Related Reading

1. H. Esmaeilzadeh, E. Blem, R. St. Amant, et al., Dark Silicon and the End of Multicore Scaling. In *International Symposium on Computer Architecture* (ISCA 2011). Retrieved from gatech.edu, 2018.
2. M. Horowitz, Scaling Power and the Future of CMOS. In *20th International Conference on VLSI Design* (VLSID 2005). Retrieved from semanticscholar.org, 2018.
3. A. Putnam, Large-Scale Reconfigurable Computing in a Microsoft Datacenter. In *IEEE Hot Chips 26 Symposium* (2014). Retrieved from microsoft.com, 2018.

# Revision History

The following table shows the revision history for this document:

| Date | Version | Description of Revisions |
|---|---|---|
| 09/29/2020 | 1.1.1 | Typographical edits. |
| 09/23/2019 | 1.1 | Updated 5G Wireless Communications. |
| 10/02/2018 | 1.0 | Initial Xilinx release. |

# Disclaimer



# Automotive Applications Disclaimer