WP539 (v1.2) February 15, 2022

System-Level Benefits of the Versal Platform

Learn about the system-level benefits of Versal® ACAPs and comparative performance to competing programmable-logic based devices.

ABSTRACT

Moore's Law has fueled the technological prosperity of the last 50 years, but it is generally believed now that Gordon Moore's 1965 forecast about the pace of innovation no longer holds true today. Continuing the silicon architectures of yesterday cannot meet the expanding demands of tomorrow's workloads. Frequently highlighted by today's leaders in the field of computer architecture [Ref 1], to meet these workload demands, the industry has entered a new golden age of computer architecture, giving rise to domain-specific architectures.

The Xilinx® Versal portfolio offers a disruptive architecture, combining best-in-class 7nm programmable logic with scalar processing engines, spatial processing hardware engines, and vector processing intelligent engines, along with leading-edge memory and interfacing technologies to provide a foundational platform for adaptable domain-specific architectures across a range of markets and applications.

This white paper evaluates the Versal architecture's system-level performance across a set of domain applications and compared to competing programmable-logic based devices.

Introduction

Over the past several years, the computing industry has witnessed a massive explosion of data and a surge of machine learning (ML) and AI applications. The result is an ever-increasing need for higher throughput and real-time computing capabilities, while also retaining adaptability to keep up with evolving workload requirements and changing protocols. See Figure 1.

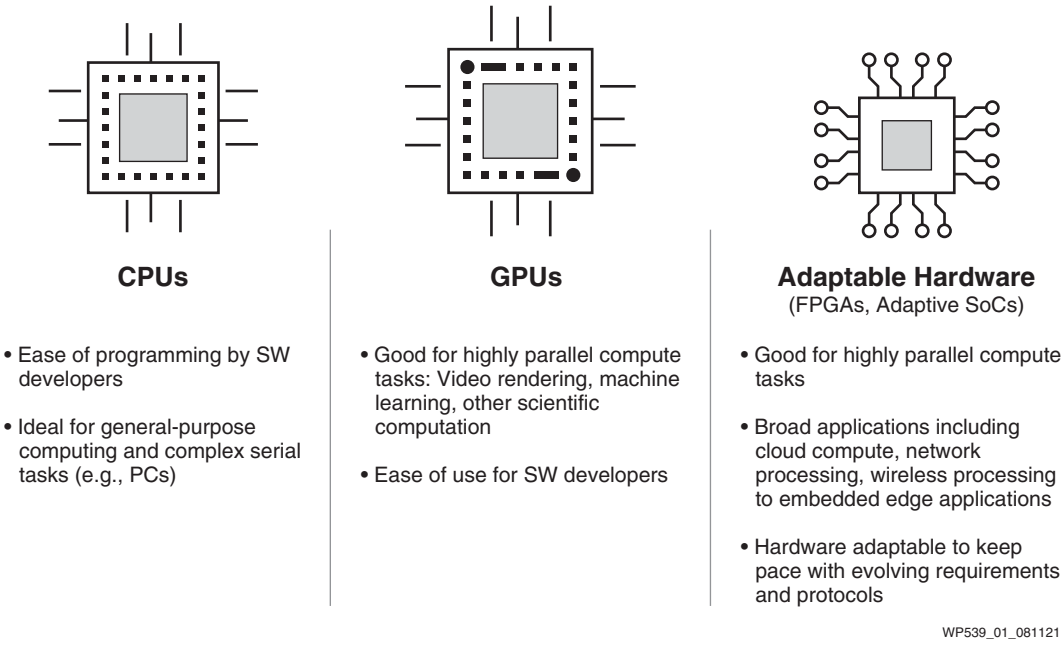


Figure 1: **Device Type Comparison**

Versal ACAPs are ready to shape the products of tomorrow in a broad range of markets and applications: data center networking, storage and compute acceleration, AI acceleration from edge to cloud, 5G wireless, wired applications, autonomous driving, and Aerospace & Defense markets, as well as many others.

System-Level Performance

The Versal architecture is not a traditional FPGA architecture. Since its inception, the drive has been to provide much higher system-level gains than incremental fabric Quality of Results (QoR) performance. Specifically, Xilinx aimed for up to 5X system-level performance over previous-generation and alternative programmable-logic-based architectures. The Versal architecture delivers this by hardening foundational IP such as AI Engines, programmable network on chip (NoC), 100G Ethernet MRMAC, 600G Ethernet DCMAC, 400G High-Speed Crypto Engines, 600G Interlaken, and hardened memory controllers.

Major Challenges

The Versal architecture addresses three major challenges:

- System-level performance per watt
- Energy-efficient compute and data movement functions
- Metal scaling limitations in programmable logic

Improved system-level value is more than just delivering simple, raw performance. Comparing performance without consideration of power is examining only half of the issue. Power impacts overall system cost, both in increased operating costs and increased costs for advanced cooling. For example, Google says that system total dissipated power (TDP) is correlated with total cost of ownership (TCO) with an R^2 of 0.78 [Ref 2].

While traditional programmable logic can provide a tremendous amount of flexibility, that flexibility comes at a cost. A function implemented in hardened gates can be 10X more power efficient than a soft implementation on programmable logic. For example, the ASIC hardening of foundational compute and data movement functions, such as PCIe® DMA to CPU host, vector-vector and matrix-matrix operations, memory access and data movement, high-speed protocol engines, and encryption, frees up more programmable logic and platform resources to developers to innovate industry-changing, adaptable, domain-specific architectures.

Metal scaling limitations have also been addressed in the Versal architecture. While transistor delays continue to improve, metal scaling has become a major challenge, which is made worse in FPGAs because they generally have more metal interconnect and are more heavily loaded than ASICs.

The chart in Figure 2 shows this trend. Plotted on the chart are normalized transistor and metal delays for the same quad routing resource on 28nm that are then scaled to 20nm, 16nm, 10nm, and 7nm. While transistor delays are continuing to decrease at a modest rate, metal delays increase nearly quadratically [Ref 3]. This trend is partially mitigated in the 7nm Versal ACAP fabric because Xilinx added metal layers to get thicker, less resistive metal tracks. Another way Xilinx addresses this is by delivering more ASIC hard blocks, which stand to provide much more in performance gains as compared to metal-dominated programmable logic.

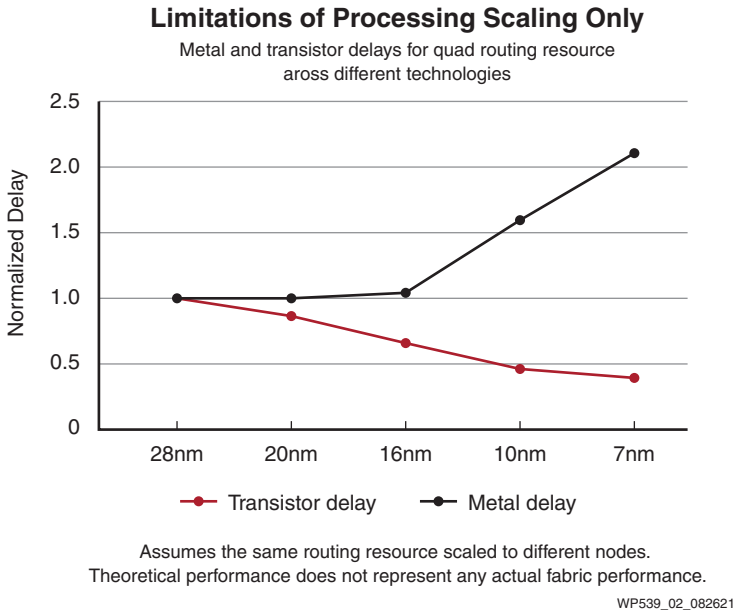


Figure 2: **Limitations of Process Scaling Only**

Historically, FPGAs have been benchmarked based solely on fabric QoR. And partially due to these challenges around metal delays, today's programmable logic fabric performance is similar to previous generations. As an example, Figure 3 shows the Geomean F_{MAX} performance across a collection of 24 RTL designs, comparing the fabric performance of Xilinx's previous-generation Virtex® UltraScale+™ FPGAs to Intel's Agilex devices.

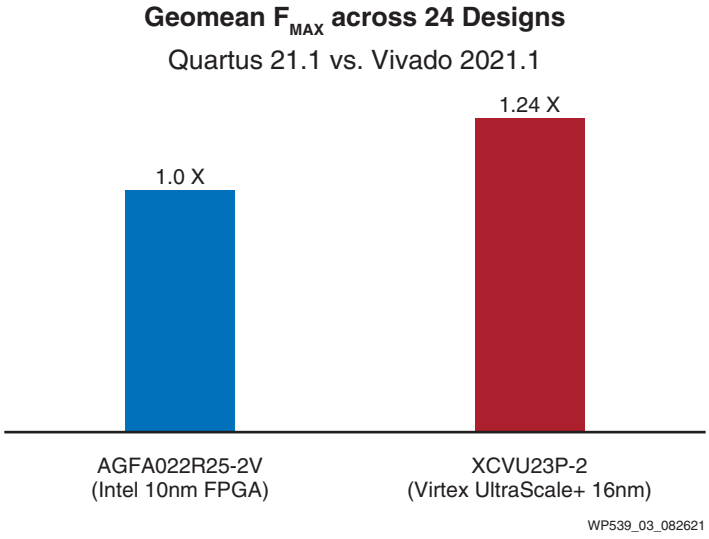


Figure 3: **Geomean F_{MAX} Performance Comparison**

Vivado, Vitis, and Vitis AI for a Software Programmable Architecture

Versal's heterogeneous processing architecture enables developers to target workloads to the right types of processors to allow for optimal performance. To realize Versal system-level value, a set of tools and workflows are required.

Xilinx provides software development platforms to support this. Vitis and Vitis AI are specialized development platforms that remove the traditional barriers for software developers while Vivado ML offers a traditional flow for hardware developers.

Developing for Versal ACAP requires a system design methodology. In an initial phase, the system architect makes key decisions about the application architecture by determining which software functions should be mapped to device kernels, how much parallelism is needed, and how it should be delivered.

Choosing the right mapping for the application tasks allows users to create the ideal balance of bandwidth, throughput, and latency – maximizing system-level performance by mapping each task to the hardware that can most efficiently execute it. For example, complex algorithms and control are a good fit for scalar processing while irregular data structures requiring low and predictable latency are mapped onto the adaptable hardware. Then heavy regular computational tasks can be coded onto the vector SIMD processing units with native support for floating-point and complex numbers arithmetic. The transport of data between these functional blocks is facilitated via a dedicated routing backbone. See Figure 4.

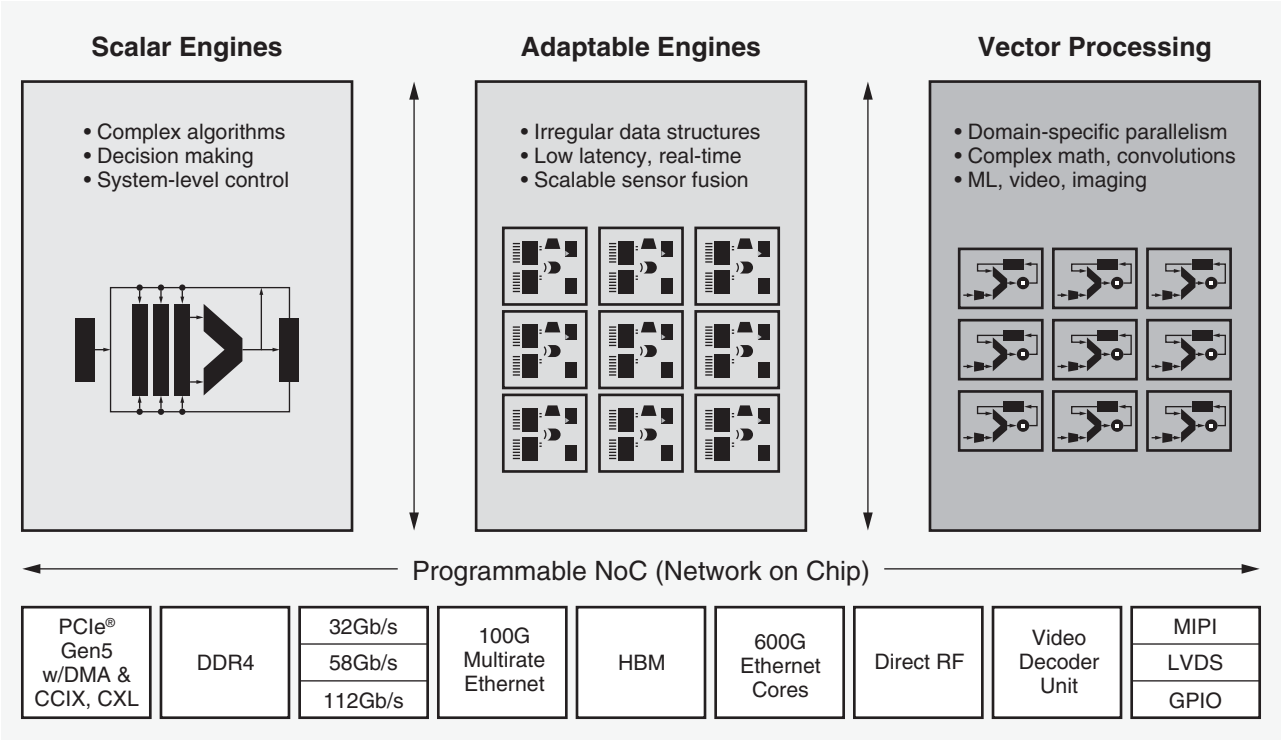


Figure 4: **Hardware Engines for Application Tasks**

Depending on these choices, the development tasks are assigned to different developers who can code and tweak the kernels concurrently. This effort primarily involves structuring C++ source code and applying the adequate compiler options and directives to create the effective implementation that meets the performance target. The kernel code can then also reuse one or several of the ~1,000 Vitis open source library functions that are available for either the adaptable hardware or the AI Engine accelerators.

All along, developers at different levels are able to work in design environments familiar to them – maximizing productivity. Emulation is also available at all these phases of development. In the final steps, the Vitis flow validates, links, and packages the kernels with the underlying platform. The NoC routing backbone transports data to and from some of these kernels as well as external memories via dedicated controllers.

AI and Video Analytics Developers

In addition to Vitis, Xilinx offers the Vitis™ AI development and Vitis Video Analytics SDKs. These kits are designed with high efficiency and ease-of-use in mind, unleashing the full potential of AI acceleration and turning computer vision into powerful insights for the datacenter or at the edge.

An acceleration card such as the Xilinx VCK5000 delivers outstanding compute power and ML perf inference performance using AI Engines, with standardized development flows. With Vitis AI, designers can run Tensorflow, PyTorch, Caffe models using Python or C++ APIs in minutes without any prior hardware knowledge.

Versal ACAPs vs. Competing FPGA

The following sections describe the system-level performance across a set of domain applications delivered by the Versal architecture, and compare it to competing programmable-logic based devices.

CNN-based Image Detection

Applied ML techniques have now become pervasive across a wide range of application domains. In fact, it will soon be difficult to find an industry that is not transformed by machine learning. One area of applied ML that has seen tremendous growth is in the field of vision and video processing. Video content on the Internet has grown rapidly over the past few years, and the need for improved methods of sorting and classifying imagery has grown commensurately.

One of the cornerstones of the Versal architecture is AI Engine technology. To keep up with 5G wireless and ML workload signal processing requirements, the Versal architecture needed to focus on how to deliver scaling compute functionality. While fabric-based DSP offers very flexible fine-grain programmability, the bit-level interconnect and programmability adds overhead that limits the scaling of compute density.

AI Engine technology is a 2D array of VLIW vector-vector, matrix-matrix compute engines, with a word-level programmable interconnect. AI Engine delivers much greater compute density while retaining Xilinx's heritage in data flow, deterministic, and high-compute efficiency processing. Below is a chart comparing the compute density progression from previous-generation DSP to AI Engines (AIE) and AIE-ML Engines. As shown, AI Engines enable Xilinx devices to deliver order-of-

magnitude increases in compute, which is fundamental in applications like 5G wireless and applied ML. See Figure 5.

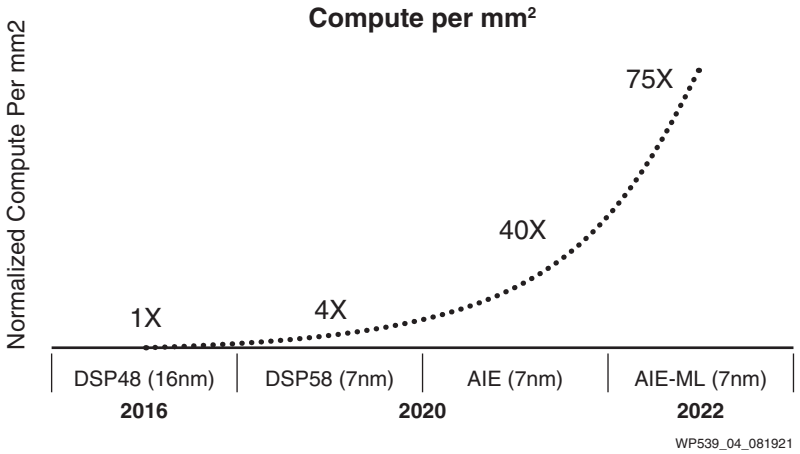


Figure 5: **Versal ACAP: Order of Magnitude Increase in Compute**

For more on AI Engines, refer to the Xilinx AI Engines and Applications white paper [Ref 4].

Showcasing the inference throughput performance on Versal ACAPs, Xilinx submitted results in the ML Perf Data Center Inference v1.0, showing industry-leading performance among hardware-programmable platforms for the ResNet50 v1.5 image detection benchmark on the VC1902, Xilinx's first Versal AI Core series device on the Xilinx VCK5000 development card for AI Inference.

Figure 6 shows a comparison of measured results (VCK5000) on Versal devices, with projected performance of competing programmable devices from Intel and the Versal AI Edge VE2802 device. Versal architectural features (AIE, NoC, and CPM(1)) drive a 2.7X–8.2X better performance per watt vs. Intel's 10nm FPGA device.

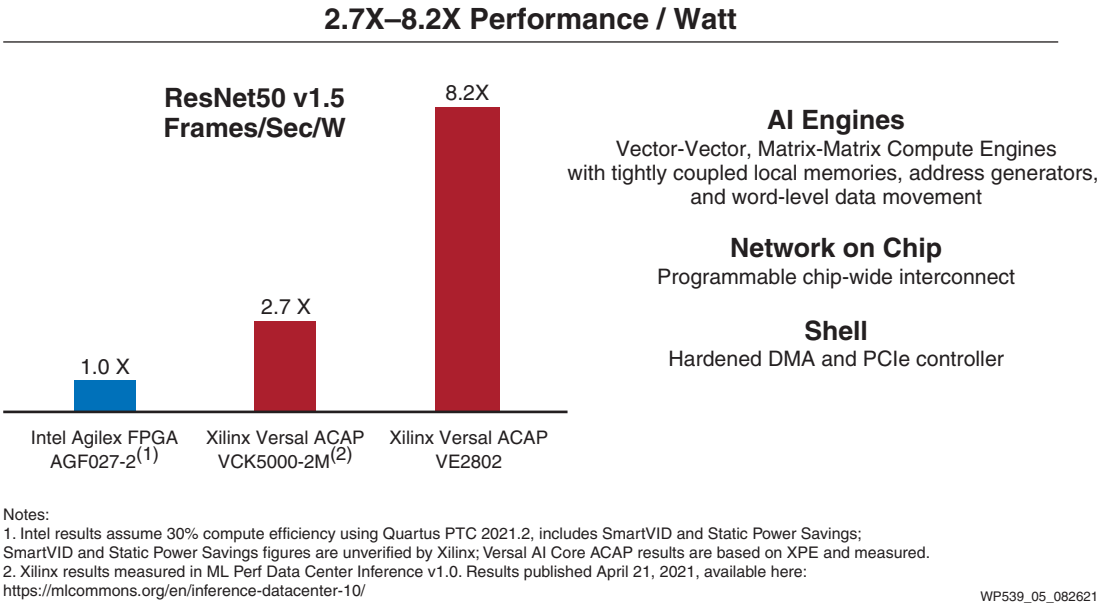


Figure 6: **Versal ACAP Delivers 2.7X Performance per Watt**

For more on the Versal AI Edge series, the VE2802 ACAP, and AIE-ML, go to Xilinx's website: https://www.xilinx.com/products/silicon-devices/acap/versal-ai-edge.html

1. CPM is an integrated block for PCIe® with DMA and Cache Coherent Interconnect designs.

5G Wireless Beamforming

Massive MIMO radio is the leading form factor for 5G commercial deployments across the globe. The physical and higher layer procedures and control signaling is defined in 5G New Radio (5G NR) to support beamforming. Massive MIMO radio uses 32 or 64 antenna elements to form beams steered towards one or multiple users while using the same spectral resources in time and frequency to significantly increase the cell capacity while reducing intra-cell and inter-cell interference.

A typical radio configuration for a 64-antenna 200MHz system, e.g., beamforming device in radio, needs to perform more than 1.5 TMAC operations per second for downlink. Additional compute is needed to perform beamforming in the uplink direction.

The Versal architecture provides adaptive compute flexibility and performance to meet the challenging and evolving 5G NR design requirements. Specifically, the Versal AI Engine technology increases desired compute density while reducing power when compared to the traditional FPGA fabric, comprising multipliers, memory, and interconnect [Ref 5].

Figure 7 shows a comparison of projected results for a wireless 5G application on a Versal AI Core VC1902 production ACAP [Ref 6], compared with the projected performance of competing programmable devices from Intel, which lack the hardened features needed to improve energy efficiency. Versal architectural features (NoC, AIE, and CPM) drive 2.14X better performance per watt vs. Intel 10nm FPGA designs.(1)

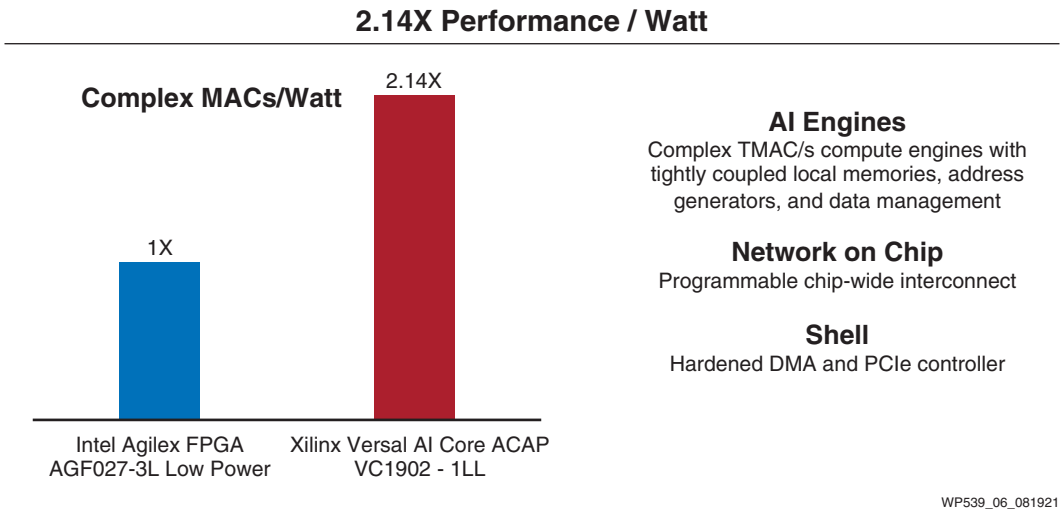


Figure 7: **Versal ACAP Delivers 2.14X Performance per Watt on Wireless 5G Beamformer Application[1]**

1. Intel results based on Quartus PTC 21.2 power estimation, assuming 75% 18x19 multiplier utilization, includes SmartVID and Static Power Savings; SmartVID and Static Power Savings figures are unverified by Xilinx; Xilinx power estimated in XPE 2020.3, Worst Case, Max Process, assuming similar compute efficiency.

Network Acceleration

In cloud provider and enterprise data centers, there is a growing need to offload a broad range of critical applications from CPUs, specifically around the area of network acceleration. A new class of hardware accelerators has emerged in the market to help offload CPU-intensive application processing.

Xilinx network accelerators revolutionize the effective use of the CPU by offloading computationally expensive network processes (e.g., IPsec and NVMeoF), while also providing composable and extensible data plane programmability. The Versal architecture stands to benefit network accelerator applications by adding many fundamental functions such as hard IP, while allowing for custom data plane processing.

These features include:

- Host PCIe interface and DMA
- 400Gb full-duplex High-Speed Crypto Engines
- Programmable NoC for on-chip data movement and hardened memory controllers
- 100GbE and 400/600GbE MACs
- Arm® Cortex®-A72 application APUs and Cortex-R5F real-time RPU cores
- Best-in-class transceiver technology

These key hard IP features save power, reduce footprint, and open up more device resources for other functions, such as in-line ML or custom packet processing functions.

Figure 8 shows a comparison of estimated power consumption of a network accelerator application on a Versal device with projected performance on competing programmable devices. Versal architectural features help drive a 16W power headroom in a 75W PCIe form factor when compared with a competing Intel 10nm FPGA(1), which exceeds the PCIe card power budget.

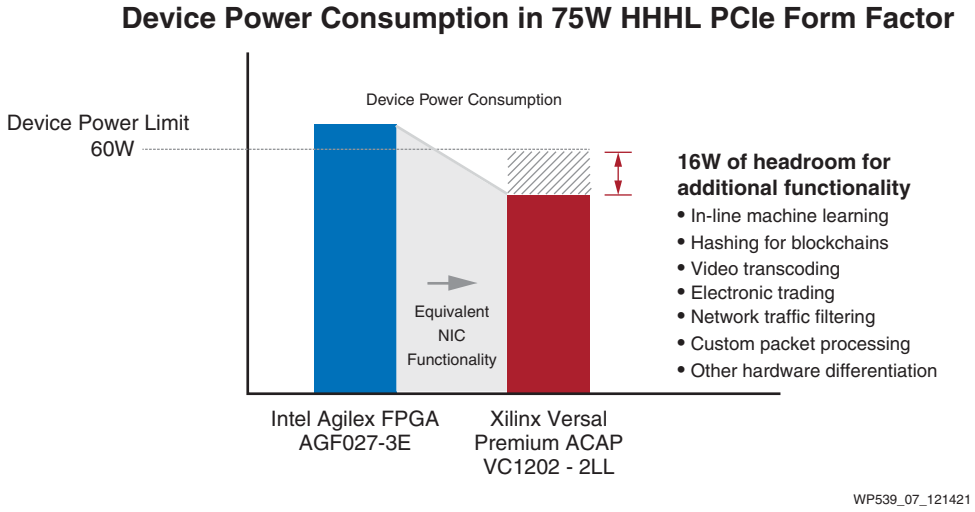


Figure 8: **Versal ACAP Provides over 16W of Additional Headroom vs. the Competing Device for a Network Acceleration Application[1]**

1. Network Accelerator Design: 540k LUTs, PCIe Gen4x16, 2x 100GbE MAC, 2x DDR4 Interfaces, NoC enabled, Xilinx power estimated in 2020.3 XPE, Worst Case, Max Process. Intel power estimated with 21.2 Quartus PTC, includes SmartVID and Static Power Savings; SmartVID and Static Power Savings figures are unverified by Xilinx.

SmartPHY Solutions for DCI Bridging and Transport

As data centers move to 400G and eventually 800G, data center interconnect (DCI) equipment will continue to need flexibility. These ever-increasing networking loads require router/switch chips to use the latest SerDes rates (56G going to 112G) to provide full density operation to facilitate DCI bridging and transport functions. Conversely, the client interfaces are varied from 10G all the way up to 112G per lane. Bridging these client interfaces to the networking and transport chips requires adjusting the SerDes rates and changing the FEC as needed. In many transport applications, other functions like multiplexing multiple clients or inverse multiplexing a client over multiple line interfaces is needed. In addition, security functions such as inline encryption/decryption are often needed to prevent frequent cyberattacks on infrastructure equipment.

Xilinx Versal ACAP SmartPHY solutions can connect up to 2.4Tb/s of transport/network interfaces to faceplate optics and integrate up to 1.6Tb/s full-duplex encryption in a single device. This is the highest density per device in the industry by far, enabling differentiated products for OEM system providers.

Figure 9 shows a comparison of estimated power consumption on Versal devices of an equivalent DCI bridging design, with projected performance of competing programmable devices.

Versal architectural features (NoC, CPM, and HSC) and the high density of hardened Ethernet interfaces drive 2.2X better performance per watt vs. Intel 10nm FPGA(1)—and at a 70% smaller PCB footprint.

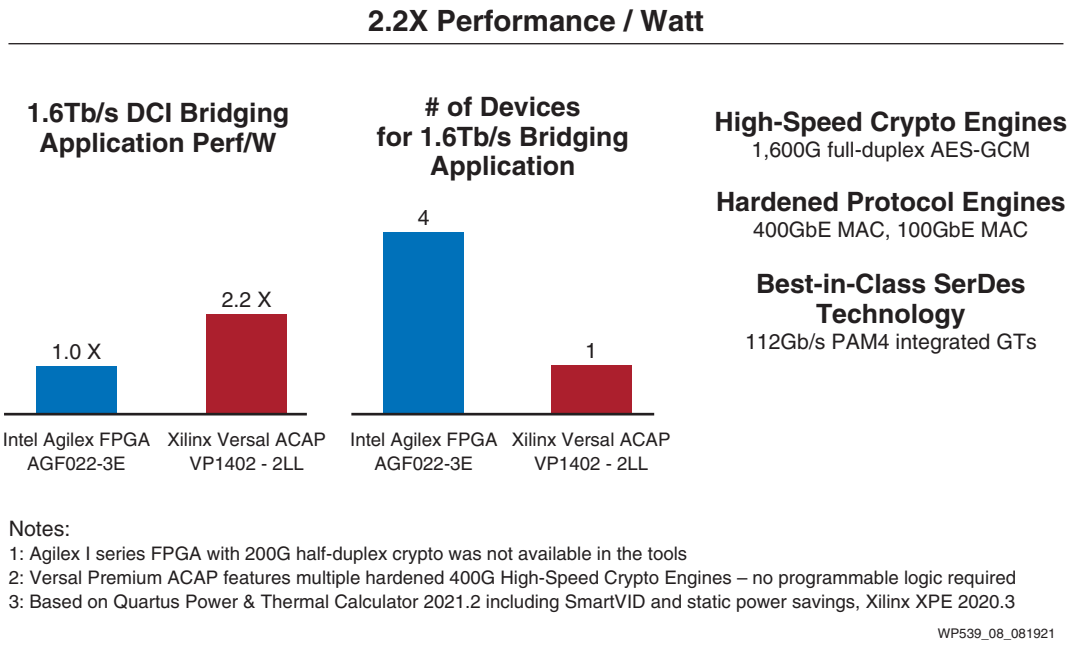


Figure 9: **Versal ACAP Provides 2.2X Better Performance per Watt vs. the Competing Device for a DCI Bridging Application[1]**

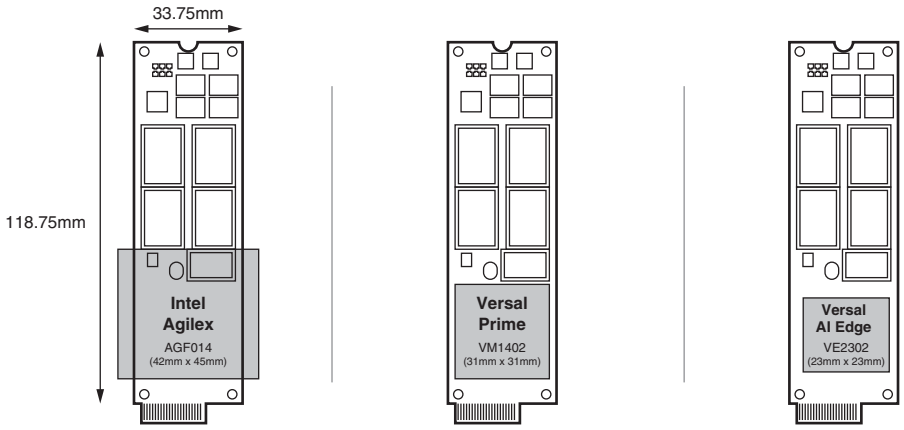
1. DCI Bridging Design: Xilinx power estimated in 2020.3 XPE, Worst Case, Max Process. Intel power estimated using 21.2 Quartus PTC includes SmartVID and Static Power Savings; SmartVID and Static Power Savings figures are unverified by Xilinx.

Storage Acceleration

Many applications are form-factor limited. Examples of these include inline enterprise storage acceleration applications, which require maximizing the capability of the adaptive hardware within a fixed power envelope. In U.2 form factor storage devices, there is a direct trade-off between the board area consumed by the adaptive hardware and the amount of media that can be included on the board. In addition, newer storage form factors, such as EDSFF E1, require packages that are 33.75mm or less in at least one dimension.

The Versal Prime VM1402 device offers 648 single-ended I/Os in a 35mm x 35mm package, and 324 single-ended I/Os in a 31x31mm package, with expected power consumption at 17W for a typical storage workload. Additionally, for computational storage applications, the Versal AI Edge VE2308 device offers up to 31 INT8 TOPs compute in a 23mm x 23mm package, delivering unprecedented performance per watt while conforming to storage hardware standards.

As shown in Figure 10, the closest competing Intel 10nm FPGA is not offered in package dimensions smaller than 42mm, preventing it from fitting into many enterprise DC storage form factors. Versal ACAP integrated hard IP (i.e., CPM DMA, NoC, and AI Engines) enable these devices to scale into smaller form factors and deliver differentiated features.



	Intel Agilex AGF014[2]	**Versal Prime** VM1402[3]	**Versal AI Edge** VE2302[5]
Logic Density[1]	487K ALMS + DDR	565K LUTs + CPM[4]+ NoC + DDR	328K LUTs + NoC + DDR
Device Package Size	42MM X 45MM	31mm x 31mm	23mm x 23mm
EDSFF Form Factor	NOT DEPLOYABLE	DEPLOYABLE	DEPLOYABLE
Compute Storage Functions			
Encryption	NOT DEPLOYABLE	✔	✔
Compression	NOT DEPLOYABLE	✔	✔
Hashing	NOT DEPLOYABLE	✔	✔
NVRAM Management	NOT DEPLOYABLE	✔	✔
ML Acceleration	NOT DEPLOYABLE	1,696 DSP Engines	34 AI Engine-ML Tiles[5]

Notes:
1: Storage acceleration functions at 6.9GB/s typically require ~300K LUTs or more
2: Agilex AGF014-2340A FPGA package
3: See Versal ACAP Prime Series Product Selection Guide for full product specifications
4: VM1402 features a CPM4, offering integrated PCIe® Gen4 with hardened DMA, eliminating the need to implement DMA in programmable logic
5: See Versal ACAP AI Edge Series Product Selection Guide for more details

WP539_09_082421

Figure 10: **Enterprise SC Storage Form Factor (EDSFF) Comparison**

Conclusion

Versal ACAP is an entirely new class of product with significantly increased capability and heterogeneous integration. By hardening many foundational IP in the Versal architecture (such as AI Engines, NoC, 100G MRMAC, 600G DCMAC, 400G High-Speed Crypto Engines, and 600G Interlaken), Versal ACAPs have significant performance and performance-per-watt superiority over competing FPGAs by offering much greater system-level performance across a wide range of applications, as shown throughout this white paper.

For additional information, including system-level benchmark comparisons, go to: www.xilinx.com/versal-performance-elevated.

To try the benchmarks, go to: https://www.xilinx.com/member/forms/registration/white-paper-539.html

References

1. Hennessy, Patterson (2019, February). A New Golden Age for Computer Architecture from https://cacm.acm.org/magazines/2019/2/234352-a-new-golden-age-for-computer-architecture/fulltext
2. 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). "Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product": https://ieeexplore.ieee.org/document/9499913
3. B. Gaide, D. Gaitonde, C. Ravishankar, and T. Bauer, "Xilinx Adaptive Compute Acceleration Platform: Versal Architecture," in proceedings of the 2019 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA'19). ACM, https://www.xilinx.com/support/documentation/white_papers/ACAP%20Paper.pdf
4. Xilinx White Paper (WP506): *Xilinx AI Engines and Their Applications*, https://www.xilinx.com/support/documentation/white_papers/wp506-ai-engine.pdf
5. Xilinx Application Note (XAPP1352): *Beamforming Implementation on AI Engines*, https://www.xilinx.com/support/documentation/application_notes/xapp1352-beamforming-ai-engine.pdf
6. Xilinx, Inc. (2021). "Xilinx Announces Full Production Shipments of 7nm Versal AI Core and Versal Prime Series Devices": https://www.xilinx.com/news/press/2021/xilinx-announces-full-production-shipments-of-7nm-versal-ai-core-and-versal-prime-series-devices.html

Acknowledgment

The following Xilinx employees have authored or contributed to this white paper:

Matthew Ouellette, Director, Silicon Product Planning

Mouli Chitta Venkata, Product Planning & Competitive Benchmarking

Brian Philofsky, Principal Technical Marketing Engineer - Power/Thermal

Harpinder Matharu, Senior Director, Technical Marketing

Faisal Dada, Principal Wired Architect

Ashwin Thiagarajan, Senior Manager - Technical Marketing

Nick Ni, Director of Product Marketing, AI - Vitis - Vivado - Ecosystem

Ryan Koehn, Product Line Manager, Mid-Range ACAPs & FPGAs

Frederic Rivoallon, Product Manager - Xilinx Software Development Flows and Competitive

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
02/15/2022	1.2	Updated Vivado, Vitis, and Vitis AI for a Software Programmable Architecture.
09/15/2021	1.1	Updated References.
08/26/2021	1.0.1	Typographical edit.
08/25/2021	1.0	Initial Xilinx release.

Disclaimer

Automotive Applications Disclaimer