WP529 (v1.0)

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

The Kria K26 SOM is built to enable millions of developers in their preferred design environment to deploy their smart vision applications faster with an out-of-the-box ready, low-cost development kit to get started.

ABSTRACT

With various advancements in artificial intelligence (AI) and machine learning (ML) algorithms, many high-compute applications are now getting deployed on edge devices. So, there is a need for an efficient hardware that can execute complex algorithms efficiently as well as adapt to rapid enhancements in this technology. Xilinx's® Kria™ K26 SOM is designed to address the requirements of executing ML applications efficiently on edge devices. In this white paper, the performance of various ML models and real-time applications is investigated and compared to the Nvidia Jetson Nano and the Nvidia Jetson TX2. Xilinx's results show that the K26 SOM offers approximately a 3X performance advantage over the Nvidia Jetson Nano. It also offers greater than a 2X performance/watt advantage over the Nvidia Jetson TX2. The K26 SOM's low latency and high-performance deep learning processing unit (DPU) provides a 4X or greater advantage over Nano, with a network such as SSD MobileNet-v1, making the Kria SOM an ideal choice for developing ML edge applications.

Computer Vision Growth

Over the past decade, significant advancements have been made in artificial intelligence and machine learning algorithms. The majority of these are applicable in vision-related applications. In 2012, AlexNet (Krizhevsky et al. [Ref 1]), became the first deep neural network trained with backpropagation, to win the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), demonstrating an unprecedented step improvement of approximately 10% in Top 5 prediction accuracy versus historic shallow networks. This key development was an inflection point, and since that time, deep neural networks have continued to make tremendous strides in both performance and accuracy. In 2015, ResNet (Kaiming et al, [Ref 2]) surpassed the classification performance of a single human. See Figure 1. For all intents and purposes, the problem of image classification in computer vision applications was deemed to have been solved with today's networks achieving results rivaling Bayes error.

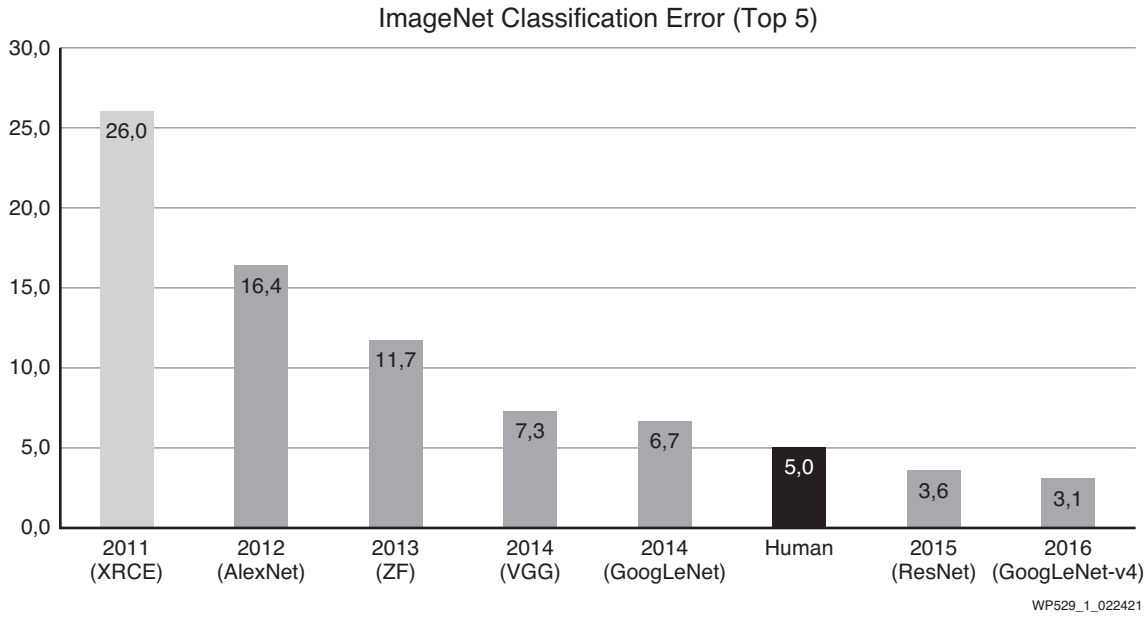

Figure 1: **Top 5 Prediction Accuracy (Source: https://devopedia.org/imagenet)**

With such exciting results, semiconductor manufacturers are racing to develop devices that support the high computational complexity requirements of deep-learning deployments. Over the coming decade, commercialization of inference, most notably computer vision with deep learning, will continue to grow at a rapid pace. The market for computer vision is anticipated to grow at a CAGR of 30%, reaching $26.2 billion by 2025 [Ref 3]. Adoption of deep-learning for computer vision has broad application for industrial markets, including retail analytics, smart cities, law enforcement, border security, robotics, medical imaging, crop harvest optimization, sign-language recognition, and much more. In the immediate future, AI will become an integral part of daily life, and some would argue that the future is upon us today!

As novel networks, layers, and quantization techniques are developed, a hardened tensor accelerator with a fixed memory architecture is not future proof. Such was evidenced following the development of MobileNets [Howard et al. [Ref 4]. The use of depthwise-convolution in the MobileNet backbone significantly reduces the compute-to-memory ratio. This is ideal for devices such as general-purpose CPUs, which are typically compute bound, but limits efficiency on architectures that are memory bound, wasting valuable compute cycles in a way that is analogous

to idle servers consuming power in a data center. This compute-to-memory ratio impacts the overall efficiency of a specific tensor accelerator architecture when deploying the network [Ref 5] Unfortunately, ASIC and ASSP tensor accelerator architectures are frozen in time a year or more prior to device tape-out. This is not the case with Xilinx inference technologies, which can be adapted and optimized over time to support the rapid evolution of neural network architectures, ensuring optimal power consumption and high efficiency.

This white paper introduces the Xilinx portfolio's newest portfolio member, the Kria™ K26 SOM, and presents key advantages of the Kria K26 SOM in embedded vision applications.

Kria K26 SOM

The Kria K26 SOM, is designed to cater to current and future market demands for vision AI and video analytics. Small enough to fit in the palm of a hand, the Kria SOM marries an adaptive SoC based on the Zynq® UltraScale+™ MPSoC architecture with all of the fundamental components required to support the SoC (such as memory and power).

Customization for production deployments is simple. The Kria SOM is mated with a simple end-user-designed carrier card that incorporates the connectivity and additional components specific to that user's end system.

For evaluation and development, Xilinx offers a starter kit, that includes a Kria K26 SOM mated to a vision-centric carrier card. The combination of this pre-defined vision hardware platform, a robust and comprehensive software stack built on Yocto or Ubuntu, together with pre-built vision-enabled accelerated applications provides an unprecedented path for developers to leverage Xilinx technologies to build systems. For additional details, refer to the Xilinx companion white paper, *Achieving Embedded Design Simplicity with Kria SOMs* [Ref 6] and *Kria KV260 Vision AI Starter Kit User Guide* [Ref 7]. This white paper's findings are based on the KV260 Vision AI Starter Kit. See Figure 2.

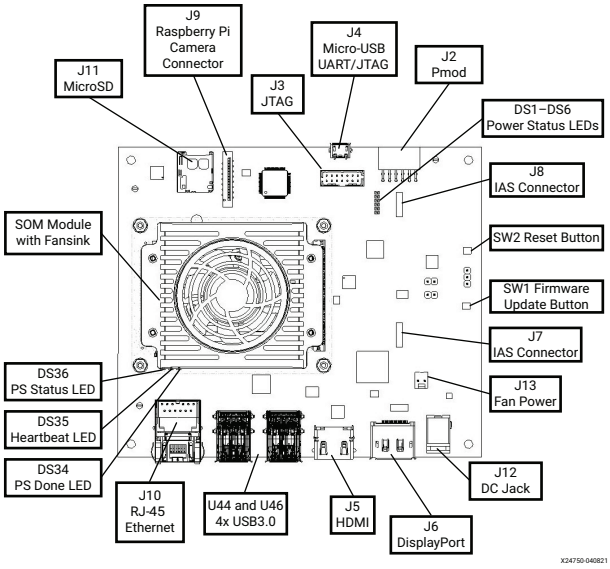

Figure 2: **KV260 Vision AI Starter Kit**

K26 SOM as an Edge Device

In addition to sub-millisecond latency requirements, intelligent applications also need privacy, low power, security, and low cost. With the Zynq MPSoC architecture at its base, the Kria K26 SOM provides best-in-class performance/watt and lower total cost of ownership, making the K26 SOM an ideal candidate for edge devices. Kria SOMs are hardware configurable, which means the solutions on the K26 are scalable and future proof.

Raw Compute Power

When it comes to deploying a solution on edge devices, the hardware must have sufficient compute capacity to handle workloads of advanced ML algorithms. The Kria K26 SOM can be configured with various deep learning processing unit (DPU) configurations, and based on performance requirements, the most applicable configuration can be integrated into the design. As an example, DPU B3136 at 300MHz has peak performance of 0.94TOPS. DPU B4096 at 300MHz offers a peak performance of 1.2TOPS, which is almost 3X the published peak performance of the Jetson Nano, which is 472GFLOPs [Ref 8].

Lower Precision Data Type Support

Deep-learning algorithms are evolving rapidly, lower precision data types such as INT8, binary, ternary, and custom data are being used. It is difficult for GPU vendors to meet the current market needs because they must modify/tweak their architecture to accommodate custom or lower precision data type support. The Kria K26 SOM supports a full range of data type precisions such as FP32, INT8, binary, and other custom data types. Also, data points provided by Mark Horowitz, Yahoo! Founders Professor In The School Of Engineering And Professor Of Computer Science at Stanford University [Ref 9], show that operations on lower precision data types consume much less power, e.g., an operation on INT8 requires one order of magnitude less energy than an FP32 operation. See Figure 3.

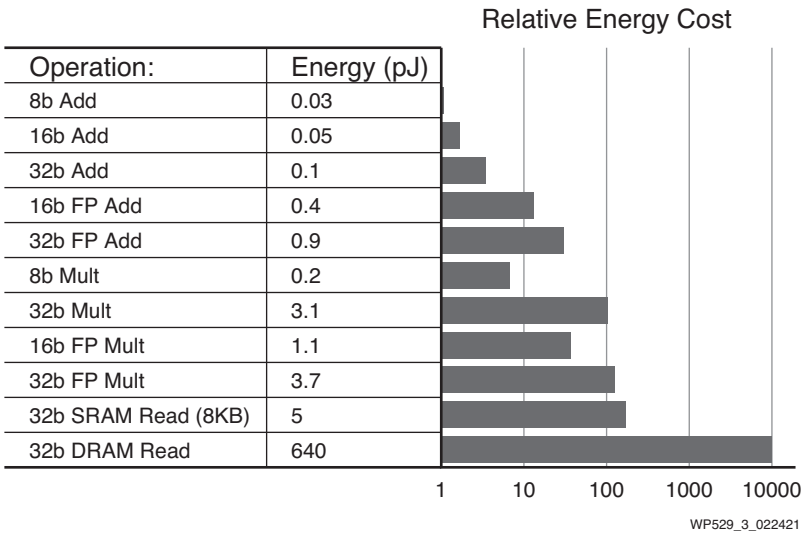

Figure 3: **Energy Cost of Operations**

The numbers in Figure 3 are based on TSMC's 45nm process and have been proven to scale well on smaller process geometries. Therefore, the re-configurability of Kria SOMs is a big advantage by accommodating any data type.

Low Latency and Power

In general, for any application design implemented on a multicore CPU, GPU, or any SoC, the overall breakup of power consumption is roughly estimated [Ref 9] as:

- Cores = 30%
- Internal memory (L1, L2, L3) = 30%
- External memory (DDR) = 40%

This is the main reason why GPUs are power hungry. To facilitate software programmability, the architecture of the GPU needs to access external DDR frequently, which is very inefficient and can sometimes become the bottleneck for high bandwidth design requirements. In contrast, the Zynq MPSoC architecture is power efficient. Its reconfigurability allows the developer to design their application with reduced or no external memory accesses, which not only helps to reduce the overall power consumption of the application, but also increases the responsiveness with lower end-to-end latencies. Figure 4 depicts the architecture of a typical automotive application wherein the GPU's communication among the modules happens via DDR while the Zynq MPSoC devices have an efficient pipeline that can be designed to avoid any DDR accesses.

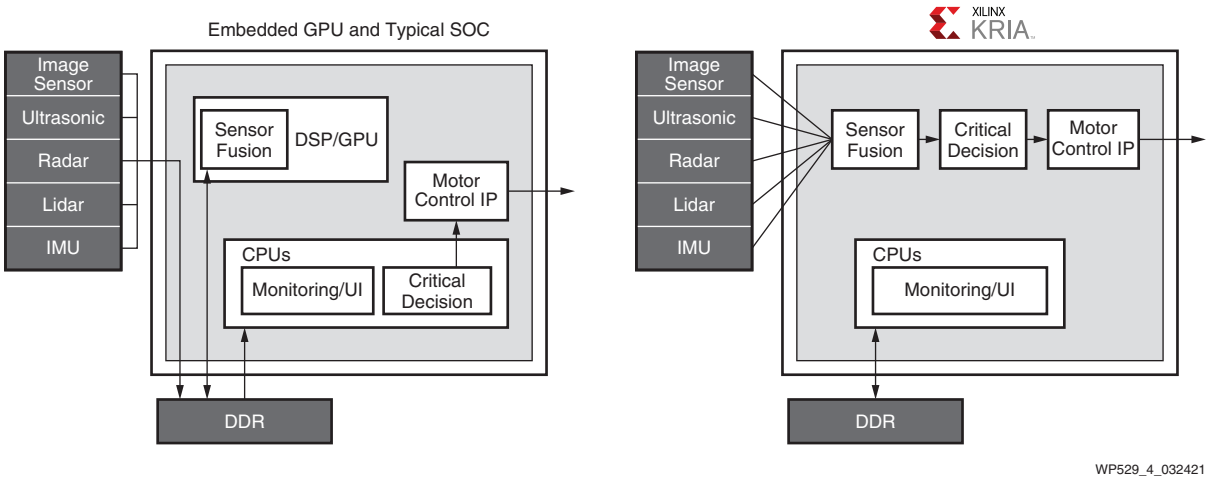

Figure 4: **Typical GPU vs. Zynq MPSoC Architecture**

Flexibility

Unlike GPUs, where the data flow is fixed, Xilinx hardware offers flexibility to uniquely reconfigure the datapath to achieve maximum throughput and lower latencies. Also, the programmable datapath reduces the need for batching, which is a major drawback in GPUs and becomes a trade-off between lower latencies or higher throughput. The Kria SOM's flexible architecture has shown great potential in sparse networks, which is one of the hot trends in current ML applications. Another important feature—and one that makes Kria SOMs even more flexible—is the any-to-any I/O connection. This enables the K26 SOM to connect to any device, network, or storage device without the need for a host CPU.

Feature Comparison

Table 1 shows a feature comparison of the Kria K26 SOM versus Nvidia Jetson Nano and Nvidia Jetson TX2.

Table 1: **K26 SOM vs. Nvidia Jetson Features Comparison**

No.	Feature	Xilinx's K26 SOM	Nvidia Jetson Nano(1)	Nvidia Jetson TX2(1)
1	Application Processor	Quad-core Arm® Cortex®-A53 MPCore™ up to 1.5GHz	Quad-Core Arm Cortex-A57 MPCore processor	Dual-Core Nvidia Denver 2 64-Bit CPU and Quad-Core Arm Cortex-A57 MPCore processor
	Real-time Processors	Dual-core Arm Cortex-R5F MPCore up to 600MHz		
2	GPU	Mali™-400 MP2 up to 667MHz (primarily used for graphics rendering)	128-core Nvidia Maxwell GPU	256-core Nvidia Pascal GPU
3	Machine Learning Throughput	1.36TOPS(2)	472GFLOPs (FP16)	1.33TFLOPs (FP16) 1.26TFLOPs for TX2i
4	Camera Interfaces	MIPI: Up to 44 DPHY2.0 lanes, Up to 11 Cameras, Max BW 10Gb/s BW per interface, Up to 16 virtual channels per interface	12 DPHY1.1 x 4 lanes Up to 4 Cameras Max BW 6Gb/s per interface	12 DPHY 1.1 lanes Up to 6 cameras Max BW 6Gb/s per camera
		SLVS, LVDS: 11 x4 SLVS or LVDS cameras interfaces	NA	NA
		SLVS-EC: 4 lane, 5Gb/s /lane	NA	NA
5	Display Interface	DP1.2 x 2 lane	2x HDMI 2.0, DP 1.2, eDP 1.2, 2x MIPI DSI x2 lanes (1.5Gb/s /lane)	2x HDMI 2.0, DP 1.2, eDP 1.2, 2x MIPI DSI
		Additional HDMI 2.0 (GTs), DisplayPort 1.4 (GTs) with Soft IPs, MIPI DSI x4 lane (2.5Gb/s /lane) w/ Soft IPs	NA	NA
6	Video Encode H.264/H.265	Up to 32 simultaneous streams, Max resolution 4K @60FPS Color format: 422 8/10 bpc and 420 8/10bpc	Up to 9 streams; Max resolution 4K @30 FPS 420 8bpc	Up to 8 streams H.265, 14 streams of H.264, Max resolution 4K @60FPS 420 8bpc
7	Video Decode H.264/H.265	Up to 32 simultaneous streams, Max resolution 4K @60FPS Color format: 422 8/10bpc and 420 8/10bpc	Up to 9 streams; Max resolution 4K @60 FPS 420 8bpc	Up to 32 streams H.265, 16 streams of H.264, Max resolution 4K @60FPS 420 8bpc

Table 1: **K26 SOM vs. Nvidia Jetson Features Comparison *(Cont'd)***

No.	Feature	Xilinx's K26 SOM	Nvidia Jetson Nano(1)	Nvidia Jetson TX2(1)
8	Wireless	GTR M.2/SATA	M.2 Key-E site on carrier	802.11a/b/g/n/ac 2×2 867Mb/s | Bluetooth 4.1 BCM4354 on module
9	Ethernet	4x 10/100/1000 Base-T Ethernet Additional Ethernet ports with Soft IP in HP I/Os	10/100/1000 Base-T Ethernet	10/100/1000 Base-T Ethernet
10	USB	2xUSB3.0, 2x USB2.0	4x USB 3.0 + Micro-USB 2.0	USB 3.0 + USB 2.0
11	PCIe®	PCIe Gen2 x 4 PCIe Gen3 x4 in with Soft IP on GTs	PCIe Gen2 x4 lanes	PCIe Gen2 x5 lanes
12	High-speed I/O (GTs) Provides Additional Interface Support	4x GTH transceivers in programmable logic can be configured to support a plethora of high-speed protocols such as SLVS - EC, PCIe Gen 3, HDMI, 10GE and many more ….	NA	NA
13	I/O Flexibility	69 3.3V I/Os, 116 1.8V I/Os allows users to create highly flexible and configurable I/O interfaces in programmable logic	NA	NA
14	Programmable Logic	256K System Logic Cells, 1248 DSPs, 26.6Mb on-chip memory allows users to implement custom accelerators for vision and ML functions	NA	NA
15	DRAM	4GB 64-bit DDR4	4GB 64-bit DDR4	8GB 128-bit LPDDR4?
16	eMMC	16GB	16GB	32GB
17	Flash	512MB QSPI	NA	NA
18	Socket Carrier Card Interface	Two 240-pin connectors	260-pin edge connector	400-pin board-to-board connector

Notes:

1. Source: https://developer.nvidia.com/embedded/jetson-modules
2. The Deep learning processor's performance is based on Xilinx DPU configurations B4096 @ 333MHz, which is implemented in the programmable logic.

The K26 SOM is based of Zynq UltraScale+ MPSoC architecture. It includes a 64-bit quad-core Arm® Cortex™-A53 application processor complex with a 32-bit dual-core Arm Cortex-R5F real-time processor and an Arm Mali™-400MP2 3D Graphics Processor. Included on the SOM is 4GB of 64-bit DDR4 memory along with QSPI and eMMC memory. The inherent secure boot capabilities with built-in hardware root-of-trust of the Zynq UltraScale+ architecture is extended via an external TPM2.0 for measured boot and is IEC 62443 capable.

The SOM offers 256K System Logic Cells, 1248DSPs, 26.6Mb of on-chip memory, which provides users with a flexible canvas to implement application-specific designs, maximizing the

performance at optimal power. Users can implement many variations of DPU, up to B4096-sized (1.4TOPs), for machine learning needs, and additional hardware accelerators for vision functions and ML pre- and post-processing in programmable logic. Additionally, the SOM has a built-in video codec for H.264/H.265, which can support up to 32 streams of encode and decode simultaneously as long as the total video pixels does not exceed 3840x2160 at 60FPS. The codecs support high quality 4:2:2 8/10bpc and 4:2:0 8/10 bpc.

One of the key advantages of K26 SOM is its unparalleled I/O flexibility. It has a plentiful amount of 1.8V, 3.3V single-ended, and differential I/Os with a quad of 6Gb/s transceivers and a quad of 12.5Gb/s transceivers. This allows users to support higher numbers of image sensors per SOM and many variations of sensor interfaces such as MIPI, LVDS, SLVS, and SLVS-EC, which are not commonly supported by ASSPs or GPUs. Additionally, users can implement DisplayPort, HDMI, PCIe®, USB2.0/3.0, and much more, including user-defined standards in programmable logic.

The size of the K26 SOM is 77mm x 60mm x 11mm supporting ruggedized applications with future SOMs planning aggressive size reductions. The K26 SOM supports a commercial temperature rating of 0°C to +85°C junction temperature as reported by the internal temperature sensor to the application processor, with all other devices on the SOM falling in line accordingly to that single measurement. Similarly, there is support for –40°C to +100°C in the industrial grade version.

Apart from features, it is also important to understand and analyze how the deep learning networks perform on each of the devices. Following is a detailed study, including data that compares similar devices on networks with similar complexity.

Comparison of Deep Learning Models

This white paper includes latency-optimized and throughput-optimized performance numbers published by Nvidia for the Jetson Nano and Jetson TX2 [Ref 10] and has measured the performance of equivalent models having similar complexity from Xilinx AI Model Zoo. The performance numbers are captured by executing the models on the KV260 Starter Kit, configured with B3136 DPU and B4096 DPU. All models for the Xilinx platform are quantized at INT8 for better power and bandwidth efficiency. Performance numbers for Nvidia Jetson Nano and Nvidia Jetson TX2 are reported with FP16 precision, as these Nvidia devices do not have lower precision support of INT8 [Ref 11]. However, both Xilinx and Nvidia performance applications use synthetic data as input and do not include pre- and post-processing time in reporting. See Table 2.

Table 2: **Deep Learning Models Performance Comparison**

No.	Model	Image Size	Xilinx K26 B3136 DPU		Xilinx K26 B4096 DPU		Nvidia Jetson Nano		Nvidia Jetson TX2	
			FPS (Latency Optimized)(1)	FPS (Throughput Optimized)(2)	FPS (Latency Optimized)	FPS (Throughput Optimized)	FPS (Latency Optimized)	FPS (Throughput Optimized)	FPS (Latency Optimized)	FPS (Throughput Optimized)
1	Inception V4	299x299	19	19.1	30.3	30.4	11	13	24	32
2	VGG-19	224x224	17.9	17.9	17.4	17.4	10	12	23	29
3	Tiny Yolo V3	416x416	88.2	92.6	148.0	161.3	48	49	107	112
4	ResNet-50	224x224	49	49.1	75.6	75.9	37	47	84	112
5	SSD Mobilenet-V1	300x300	129.6	133.4	192.1	200.4	43	48	92	109
6	SSD ResNet34	1200x1200	1.6	1.6	2.5	2.5	1	1	3	2

Notes:

1. Latency optimized on K26 SOM means, executing with 1 thread.
2. Throughput optimized on K26 SOM means, executing with 2 threads.
3. Contact your local Xilinx sales representative for Xilinx ML performance package information.

Per the information in Table 2, the performance numbers on the K26 SOM of all the models are better than the Nvidia Jetson Nano, and some models, like the SSD Mobilenet-V1, outperform Nvidia with throughput by more than 4X against Jetson Nano and around 2X higher against Jetson TX2. See Figure 5, where significant throughput gain can be easily visualized.

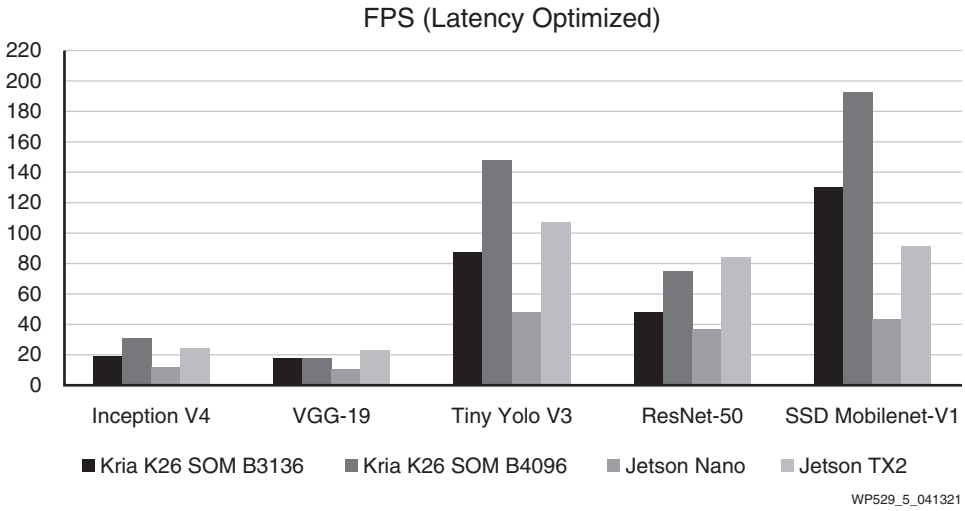

Figure 5: **FPS Measurement (Latency Optimized)**

Power Measurement

It is important for edge devices to produce optimum performance, but at the same time, they must consume less power. In this white paper, the peak power consumed by Nvidia and Xilinx's SOM modules have been measured at the time of executing a particular model from Table 2. Also the power measurement of the SSD ResNet34 model is not included because the Jetson benchmark repository does not include this model. [Ref 10]

The peak power numbers were captured while executing the comparisons in latency-optimized mode on all the three devices. For Nvidia Jetson Nano and TX2, the readings were taken every

30 seconds from the sysfs node of the INA3221x driver and at every 10 seconds from the ina260-adc driver for Xilinx's K26 SOM. See Table 3.

Table 3: **Peak Power Measurement (In Watts)**

No.	Model	Xilinx K26 SOM B3136 DPU	Xilinx K26 SOM B4096 DPU	Nvidia Jetson Nano	Nvidia Jetson TX2
1	Inception V4	8.09	10.10	7.40	11.20
2	VGG-19	8.55	11.28	8.10	13.10
3	Tiny Yolo V3	8.26	11.08	7.80	12.30
4	ResNet-50	7.47	9.28	7.70	11.70
5	SSD Mobilenet-V1	7.67	9.29	7.30	10.80

To get a better picture of the power advantage, see Figure 6, which showcases performance per watt. The K26 SOM clearly has 3.5X advantage over Jetson Nano and 2.4X advantage over Jetson TX2.

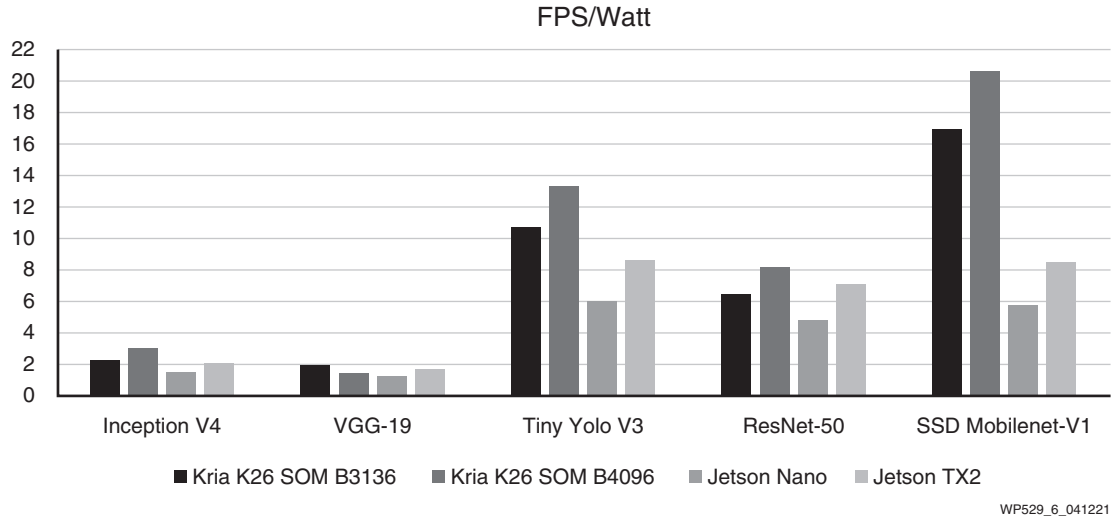

Figure 6: **FPS/Watt**

Advantages of Pruning

Xilinx provides an AI Optimization tool to further enhance the performance of various neural networks on the K26 SOM. The comparison data presented in this white paper, so far, is conducted on raw models with no optimization or pruning. Most neural networks are typically over-parameterized, and there is a significant level of redundancy that can be optimized. Xilinx's AI Optimizer is an industry-leading model compression technology. The tool can help in reducing the model complexity up to 50X with minimal impact on accuracy.

In this white paper, an example was taken from the case study done by Xilinx Research [Ref 12], where a very complex SSD + VGG model having 117 Giga Operations (Gops) was refined over multiple iterations using Xilinx's AI Optimizer tool. Figure 7 shows the benefit of pruning the model using the AI Optimizer tool. As a baseline, the model having Operations 117Gops, ran at a max FPS of 18 on a Zynq UltraScale+ MPSoC configured with 2*B4096 DPU. Over a few iterations of pruning, the data shows there is a significant reduction in complexity, with a corresponding increase in the FPS without any impact on the accuracy (mAP). At the 11th iteration, there is a 10X reduction in the

complexity, i.e., the complexity dropped from 117Gops to 11.6Gops, and had more than a 5X gain in performance, i.e., an increase from 18FPS to 103FPS, with only 1.1% drop in accuracy, i.e., an increase from 61.55mAP to 60.4mAP.

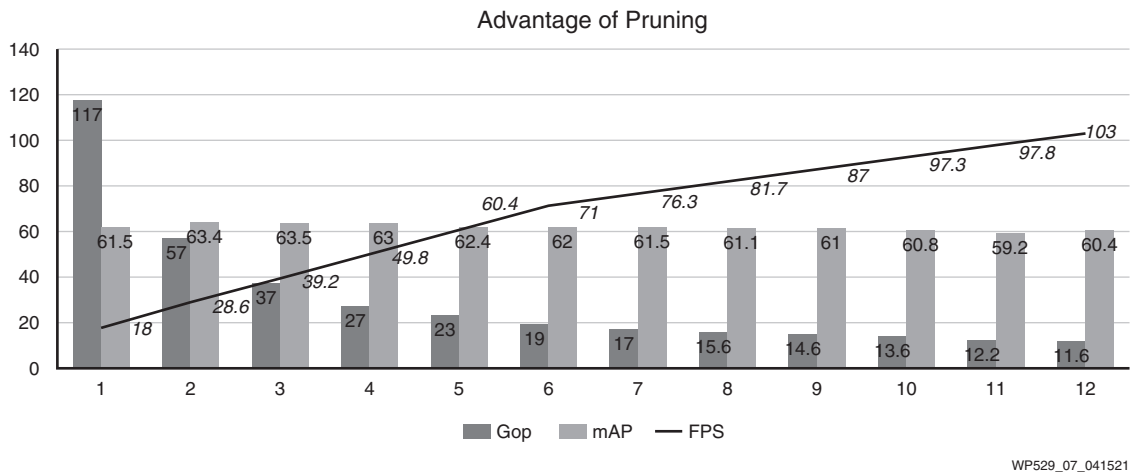

Figure 7: **Pruning Results**

To this point, only the raw performance of the Kria K26 SOM vs. GPUs has been examined. It is important to understand how this raw performance translates in real use cases. Real use cases are complex and involve other modules in the pipeline like pre- and post-processing components for any AI-ML application. In such applications, maximum throughput is determined by the least performing component in the pipeline.

The performance comparison in the following section is based on a real use case for both devices.

Performance Comparison of a Real World Application

To analyze a real-world use case, a machine learning based application that accurately detects and recognizes vehicle license plates was chosen. Xilinx has partnered with Uncanny Vision, an industry leader in video analytics solutions for smart cities, with the goal to provide a world-class auto number plate (license plate) recognition (ANPR) solution to the market. The application is being adopted widely across many cities in the world as part of the smart city buildout. Key applications of ANPR include automated toll management, highway monitoring, parking access, and secure gate access. The ANPR application is an AI-based pipeline that includes video decode, image pre-processing, machine learning (detection), and OCR character recognition. See Figure 8.

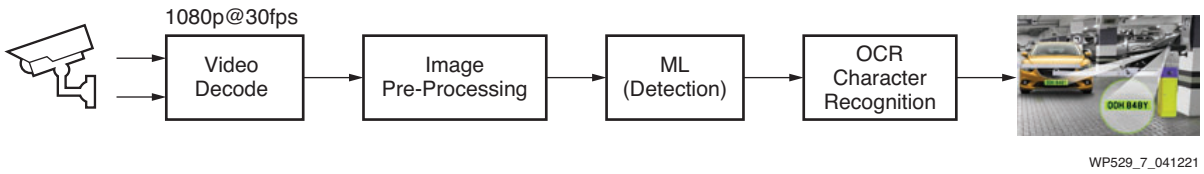

Figure 8: **Processing Blocks of ALPR Application**

ANPR AI box applications typically ingest one to many H.264 or H.265 encoded RTSP streams from off-the-shelf IP cameras and are decoded (uncompressed). The decoded video frames are pre-processed (typically scaled, cropped, color space converted, and normalized) prior to ingress to the

machine learning algorithm. In the case of high-performance commercial ANPR implementations, a multi-stage AI pipeline is typically required. It is the role of the first network to detect and localize a vehicle in the frame. This is coupled with additional algorithms that serve to track the vehicle trajectory across multiple frames, as well as algorithms to optimize image quality for OCR by selecting the ideal frame exposure. The vehicle ROI is typically cropped and scaled, and then fed to a secondary detection network that localizes the number plate. The pixels associated with the number plate ROI are cropped and scaled, and finally ingested by the last neural network that implements the OCR prediction. The meta-data predictions from this final stage are the alphanumeric characters embossed and painted, or otherwise visible on the number plate. For purposes of comparison, Uncanny Vision's ANPR application, which had been deployed commercially on GPUs and CPUs, was optimized for deployment on the Kria KV260 Vision AI Starter Kit. The result was that Uncanny Vision's algorithm, deployed on the Kria SOM, broke the $100 per-stream barrier, while achieving greater than 2–3X the performance of the commercially available, competing SOMs historically leveraged by Uncanny Vision. While no specific effort was made by Xilinx to benchmark Uncanny Vision's algorithms on Nvidia SOMs, Table 4 illustrates the performance of Uncanny Vision's industry-leading ANPR algorithm, deployed on the Kria SOM, side-by-side with public data published for "License Plate Recognition," built on Nvidia's Deepstream-SDK [Ref 13].

Table 4: **Performance Comparison of ANPR Application**

Hardware (Module)	Nvidia Jetson Nano	Nvidia Jetson TX2 4GB	Nvidia Jetson TX2 NX	Xilinx Kria K26C SOM
Price	$129	$299	$199	$250
Fps	8(1)	23(1)	Not available, expected to be ≤ TX2	33
Number of streams (assuming 10fps per stream)	~1	2	Not available, expected to be ≤ TX2	3
$/streams	$129	$150	$100	$83
Max Power (SOM only)	10W(2)	15W(2)	Not available, expected to be ≤ TX2	15W(3)
Watts/stream	10	7.5	7.5	5

Notes:

1. Source: https://developer.nvidia.com/deepstream-sdk
2. Power numbers for Nvidia are maximum power ratings for the SOM. Source: https://developer.nvidia.com/embedded/jetson-modules
3. Maximum power rating for K26 SOM only.

The data shows that Uncanny Vision's ANPR pipeline, optimized for the KV260 Starter Kit, achieves throughput greater than 33fps, which is significantly higher than Nvidia benchmarks of 8fps for Jetson Nano and 23fps for Jetson TX2. This unprecedented level of performance provides ANPR integrators and OEMs with greater flexibility in buildouts than the competition. Each additional AI box that is installed has a direct impact on installation costs, not the least of which is the associated cost of wiring and conduit. Depending on the specifics of the installation, designers can trade-off higher frame rates for the processing of a higher number of streams per box. Parking lot installations (i.e., stop-and-go, boom-barrier, free-flow) can typically be supported by inference and capture frame rates of 10fps while reliably capturing number plate metadata. This allows designers to aggregate multiple camera streams into a single box, often reducing the overall

CAPEX and OPEX per gate. In high-speed applications, such as highway toll and law-enforcement, higher frame rates ensure reliable detection and recognition for vehicles operating at high speeds. At 33fps, the K26 SOM supports more robust recognition and evidence collection than comparable, competing solutions available in the market today.

Most ANPR systems are required to operate in environmentally harsh conditions with high reliability. The I-grade version of the K26 SOM is built for rugged environments, supporting an operating temperature of –40°C to 100°C and an industry-leading three-year warranty. The result is that the total cost of ownership for ANPR systems based on the K26I SOM is significantly lower than what was previously available on the market.

Uncanny Vision's ANPR application illustrates that not only does the K26 SOM perform extremely well in standard performance comparisons, but also that it is more efficient in providing developers with access to the raw performance necessary to accelerate the entire AI and vision pipeline. By contrast, competing solutions tend to offer low levels of efficiency, outside of the realm of standard benchmarks, with much higher power consumption.

Ease-of-Use

Xilinx has put considerable effort into ease-of-use by developing various tools, libraries, frameworks, and reference examples mentioned below, which have made application development very easy for software developers on the Xilinx platform.

Vitis AI Development Kit

The Vitis™ AI development environment consists of the Vitis AI development kit for AI inference on Xilinx hardware platforms, and supports Xilinx SoCs, Alveo™ accelerator cards, and now, Kria SOMs. It consists of optimized tools, libraries, and pre-trained models that provide the developer with a stepping-stone to custom model deployment. It is designed with high efficiency and ease-of-use in mind, unleashing the full potential of AI acceleration on Xilinx devices. The Vitis AI development environment makes it simple for users without FPGA knowledge to deploy deep-learning models on the Kria SOM, abstracting away the intricacies of the underlying silicon devices See the Vitis AI User Guide [Ref 14].

Furthermore, Vitis AI integration with the Apache open-source TVM project [Ref 15] expands the possibilities for model deployment on Xilinx platforms. TVM provides diverse, up-to-date, and extensible machine learning framework support, leveraging contributions from the open-source community as well as a variety of large commercial users. TVM automates graph segmentation and compilation, ensuring that almost any network operator from any framework can be deployed, even if that operator is not natively supported by the Vitis AI development environment. Virtually, any operator can be compiled to x86 and Arm® targets in the TVM environment. The implication is that developers can quickly deploy their models, accelerating Vitis-AI development environment supported sub-graphs on the Xilinx DPU, while the remainder of the graph is deployed on the CPU [Ref 16].

PetaLinux SDK

PetaLinux is an embedded Linux Software Development Kit (SDK) targeting Xilinx SoCs. Based on Yocto, PetaLinux is a Xilinx development tool that provides everything necessary to build, develop, test, and deploy embedded Linux systems.

The PetaLinux tool contains:

- Yocto Extensible SDK (eSDK)
- XSCT (Xilinx Software Command-Line Tool) and toolchains
- PetaLinux CLI tools

Vitis Vision Libraries

Vitis Vision Libraries enable the user to develop and deploy accelerated computer vision and image processing applications on Xilinx platforms, while working at an application level. These open-source library functions offer a familiar interface, based on OpenCV, but have been optimized for high performance and low resource utilization on Xilinx platforms. They also offer flexibility to meet the adaptable throughput requirements of the user's vision systems. Table 5 lists some of the major functions that have been pre-accelerated on Xilinx platforms. For an exhaustive list, please visit the Vitis libraries landing page on Xilinx.com [Ref 17].

Table 5: **Accelerated Functions on Xilinx Platforms**

Basic Functionally	ISP	Morphology	ML
absolute difference	2dnr-nlm	dilate (max)	normalization
accumulate	3dlut	distanceTransform	standardization
arithmetic addition	3a	erode (min)	**Edge Detector**
arithmetic subtraction	ccm	houghlines	canny
bitwise: AND, OR, XOR, NOT	crop	standard deviation	difference of gaussian (DOG)
bounding box	cvt	**Thresholding**	laplacian
color convert	demosaic-cfa-mono	binary threshold	scharr
crop	demosaic-hlqi	binary inverse threshold	connected components analysis
custom convolution	dpc-single	color threshold	sobel
duplicate image	gamma correction	to zero threshold	**Tracking**
histogram	gain control	to zero inverse threshold	Lucas-Kanade dense optical flow
magnitude	gtm	truncate threshold	extended Kalman filter
paint mask	hdr-decompanding	**GeoTransform**	Kalman filter
pixel-wise multiplication	hdr-multi-expo-fusion	affine	mean shift
lookup table	hist_equalize	flip	**Misc**
zero	lsc	perspective	stereobm
Smooth	ltm-clahe	pyrdown	stereorectify
bilateral	ob	pyrup	stereosgbm
box (mean)	quantize-dither	remap	feature matcher
gaussian	resize-area	resize	feature detector
median	resize-bilinear	rotate	feature descriptor
mode	resize-nn	translation	–

Video Analytics SDK

The Video Analytics SDK is an application framework that is built on the open-source and widely adopted GStreamer framework. This SDK is designed to support seamless development across all Xilinx platforms, including Xilinx FPGAs, SoCs, Alveo cards, and of course, Kria SOMs. With this SDK, developers can assemble vision and video analytics pipelines without the necessity of possessing deep knowledge of the underlying complexities of an FPGA. In addition, this SDK provides APIs that enable users to quickly develop efficient, custom, acceleration kernels for integration into the SDK framework in the form of GStreamer plug-ins. Custom acceleration pipelines, with or without custom acceleration kernels, can be easily assembled with minimal effort by a typical embedded developer.

Accelerated Applications

Accelerated applications are a foundational building block of the Kria SOM offering. These applications are complete, production-enabling, end-to-end solutions that are specifically targeted to support popular vision use cases. Xilinx accelerated applications include a pre-optimized vision pipeline accelerator in the programmable logic region that developers can use as-is, or further optimize to meet application specific requirements. A robust software stack enables users to easily customize and enhance the functionality of the solution by modifying only the firmware or swapping AI models. Furthermore, via the new Kria Xilinx App Store, developers can explore, evaluate, or purchase industry-leading applications from domain specialists such as Uncanny Vision.

Conclusion

This white paper started with a detailed feature comparison of Xilinx's newest Kria K26 SOM against Nvidia Jetson Nano and Nvidia Jetson TX2. Then the performance was compared on a few of the industry's leading models like Tiny Yolo V3, SSD_Mobilenet_V1, ResNet50, etc., on the K26 SOM with two DPU configurations, i.e., B3136 and 4096 DPU.

The results were compared with published benchmarking numbers on Nvidia’s website. The results found that the K26 SOM outperforms both Jetson Nano and Jetson TX2 devices. Not only did the Kria SOM achieve a massive throughput gain, i.e., greater than 4X over Jetson Nano, but it also had a significant advantage in performance/watt, i.e., more than 2X over Jetson TX2.

The analysis was further expanded to compare the performance of a real application, which automatically detects and recognizes the license plates on vehicles. The K26 SOM works very well for real applications and delivers a significant performance advantage. The underlying hardware in the K26 SOM is adaptable and helps in accelerating complete applications with lower latencies and lower power consumption. For ANPR applications, the K26 SOM performs 4X better than Jetson Nano, and with higher throughput, customers can accommodate more streams per device, resulting in a lower total cost of ownership. With advantages like higher throughput, lower latencies and power consumption, while being flexible and scalable for future needs, the Kria K26 SOM is a perfect choice for Vision AI based applications on edge devices.

Adaptive acceleration starts here: https://www.xilinx.com/kria

References

1. Krizhevsky, Alex, 2012, "ImageNet Classification with Deep Convolutional Neural Networks," https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
2. Ke, Haiming, 2015, "Deep Residual Learning for Image Recognition," https://arxiv.org/pdf/1512.03385.pdf
3. Dialani, Priya, 2019, "Computer Vision: The future of artificial intelligence," https://www.analyticsinsight.net/computer-vision-the-future-of-artificial-intelligence/
4. Howard, Andrew, 2017, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," https://arxiv.org/pdf/1704.04861.pdf
5. Yao, Song et al., Hotchips Symposium 2018, Session 8, "The Evolution of Accelerators upon Deep Learning Algorithms," https://www.youtube.com/watch?v=zbtZ-rALqF4
6. Xilinx White Paper, WP528, "Achieving Embedded Design Simplicity with Kria SOMs," https://www.xilinx.com/products/som/achieving-embedded-design-simplicity-with-kria-soms.html
7. Xilinx User Guide, UG1089, Kria KV260 Vision AI Starter Kit User Guide: https://www.xilinx.com/cgi-bin/rdoc?t=som-doc;v=latest;d=ug1089-kv260-starter-kit.pdf
8. Nvidia Website: https://developer.nvidia.com/EMBEDDED/Jetson-modules
9. Horowitz, Mark, ISSCC Presentation, 2014: "Computing's Energy Problem and What We Can Do About It," https://vimeo.com/99757212
10. Nvidia Website: https://developer.nvidia.com/embedded/jetson-benchmarks
11. Nvidia Website: https://forums.developer.nvidia.com/t/how-to-use-int8-inference-with-jetsontx2-tensorrt-2-1-2/54890
12. Clark, Alvin, "Xilinx Machine Learning Strategies For Edge", 2018. https://www.xilinx.com/publications/events/machine-learning-live/sandiego/xilinx_machine_learning_strategies_for_edge.pdf
13. Nvidia Website: "License Plate Recognition" by Nvidia's Deepstream-SDK: https://developer.nvidia.com/deepstream-sdk
14. Xilinx User Guide, UG1414 Vitis AI User Guide: https://www.xilinx.com/cgi-bin/docs/rdoc?t=vitis_ai;v=1.3;d=ug1414-vitis-ai.pdf
15. Apache TVM website, https://tvm.apache.org/
16. Xilinx TVM webinar, https://event.on24.com/eventRegistration/EventLobbyServlet?target=reg20.jsp&partnerref=EDM1&eventid=2746831&sessionid=1&key=597DBD491AEF87E1AB3FBAA582A3EAEE&regTag=&sourcepage=register
17. Xilinx Website: https://www.xilinx.com/products/design-tools/vitis/vitis-libraries/vitis-vision.html

Additional Reading

1. Xilinx White Paper, Cathal Murphy and Yao Fu, WP492, "Xilinx All Programmable Devices: A Superior Platform for Compute-Intensive Systems". June 13, 2017. https://www.xilinx.com/support/documentation/white_papers/wp492-compute-intensive-sys.pdf
2. Jason Cong, Zhenman Fang, Michael Lo, Hanrui Wang, Jingxian Xu, Shaochong Zhang, Center for Domain-Specific Computing, UCLA, "Understanding Performance Differences of FPGAs and GPUs." https://vast.cs.ucla.edu/sites/default/files/publications/FCCM2018-paper120-final.pdf
3. Gordon Cooper, "New Vision Technologies For Real-World Applications", 3 Oct 2019. https://semiengineering.com/new-vision-technologies-for-real-world-applications/
4. Xilinx White Paper, "FPGAs in the Emerging DNN Inference Landscape", 28 Oct, 2019. https://www.xilinx.com/support/documentation/white_papers/wp514-emerging-dnn.pdf
5. Linus Pettersson, "Convolutional Neural Networks on FPGA and GPU on the Edge: A Comparison", Jun 2020. https://www.diva-portal.org/smash/get/diva2:1447576/FULLTEXT01.pdf
6. Alvin Clark, "Xilinx Machine Learning Strategies For Edge", 2018. https://www.xilinx.com/publications/events/machine-learning-live/san-diego/xilinx_machine_learning_strategies_for_edge.pdf
7. Ihwoo, "Nvidia Jetson Nano / Xavier power monitoring", Jul, 2020. https://inaj012.medium.com/nvidia-jetson-nano-xavier-power-monitoring-62d374e9dc81
8. Intel, "FPGA chips are coming on fast in the race to accelerate AI", 10 Dec, 2020. https://venturebeat.com/2020/12/10/fpga-chips-are-coming-on-fast-in-the-race-to-accelerate-ai/
9. Mehul Ved, "Artificial Intelligence (AI) Solutions on Edge Devices", 8 May 2019. https://stratahive.com/artificial-intelligence-ai-solutions-on-edge-devices/
10. Hewelett Packard Enterprise, "7 Reasons Why We Need to Compute at the Edge". https://www.hpe.com/us/en/newsroom/blog-post/2017/03/7-reasons-why-we-need-to-compute-at-the-edge.html
11. https://github.com/jkjung-avt/tensorrt_demos
12. https://github.com/NVIDIA-AI-IOT/deepstream_reference_apps/tree/restructure/yolo#trt-yolo-app
13. https://elinux.org/Jetson_Zoo

Acknowledgment

The following Xilinx employee has authored or contributed to this white paper:

Jasvinder Khurana, Staff System Design Engineer

Quenton Hall, AI Systems Architect

Girish Malipeddi, Director of Marketing, Multimedia and Board Solutions

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
04/20/2021	1.0	Initial Xilinx release.

Disclaimer

Automotive Applications Disclaimer