Under the Digital Hood

Adaptive Computing and AI for Autonomous Vehicles

ElectronicDesign | SPONSORED BY AVNET Reach Further™ XILINX AVNET SILICA

Under the Digital Hood

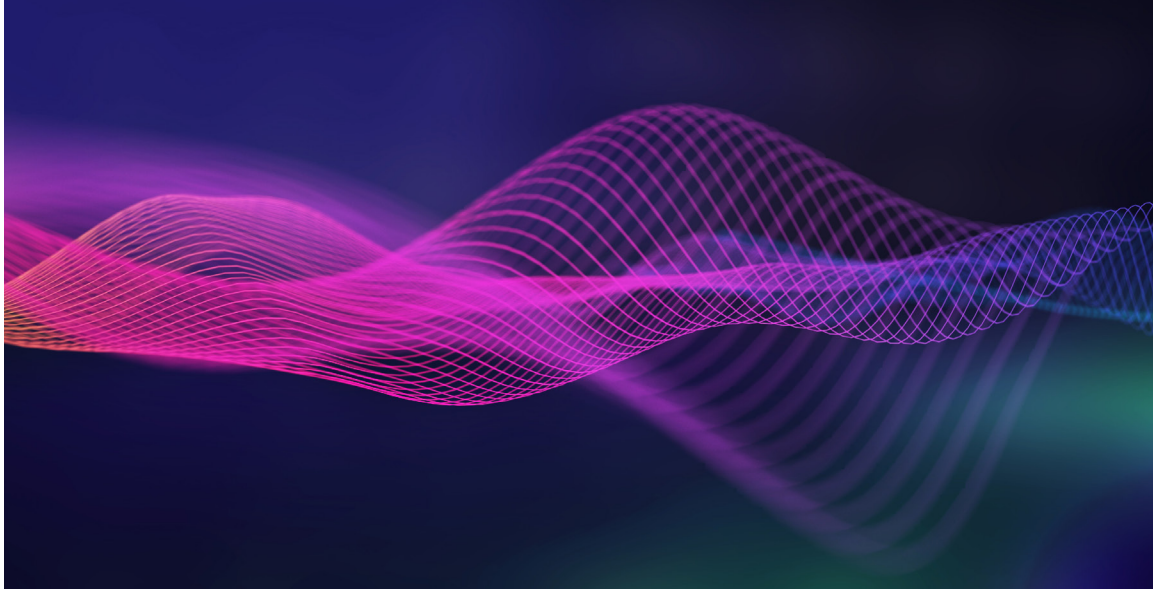
Adaptive Computing and AI for Autonomous Vehicles

INTRODUCTION

AV TECHNOLOGY is continuing to advance at a swift clip. Although roadblocks still exist, automotive leaders and big tech are making significant progress in achieving higher levels of autonomy. Today we are seeing new system architectures and software tools emerge to support the compute-intensive AV environment. Whole-application acceleration for advanced processing and AI inference spanning the vehicle, edge and cloud is kicking into high gear. And the lines separating hardware and software development are getting thin. This ebook reviews these and other key advances pushing today's driver-assist technologies faster down the AV road—so buckle up. We're in for an exciting ride.

Bill Wong
Editor,
Senior Content
Director

CONTENTS



Credit: Dmitriy Razinkov | Dreamstime

CHAPTER 1:

The AV from Algorithm to Acceleration

PETER JENCOE, Industrial Technology Writer

The latest advancements in AV technology are starting to accelerate past Level 2.

To the casual observer not steeped in the nuance and complexity of AI and automotive technology, the twilight zone we find ourselves betwixt and between today's hands-on SAE Level 2 ADAS technology and tomorrow's autonomous vehicle (AV) may seem perplexing.

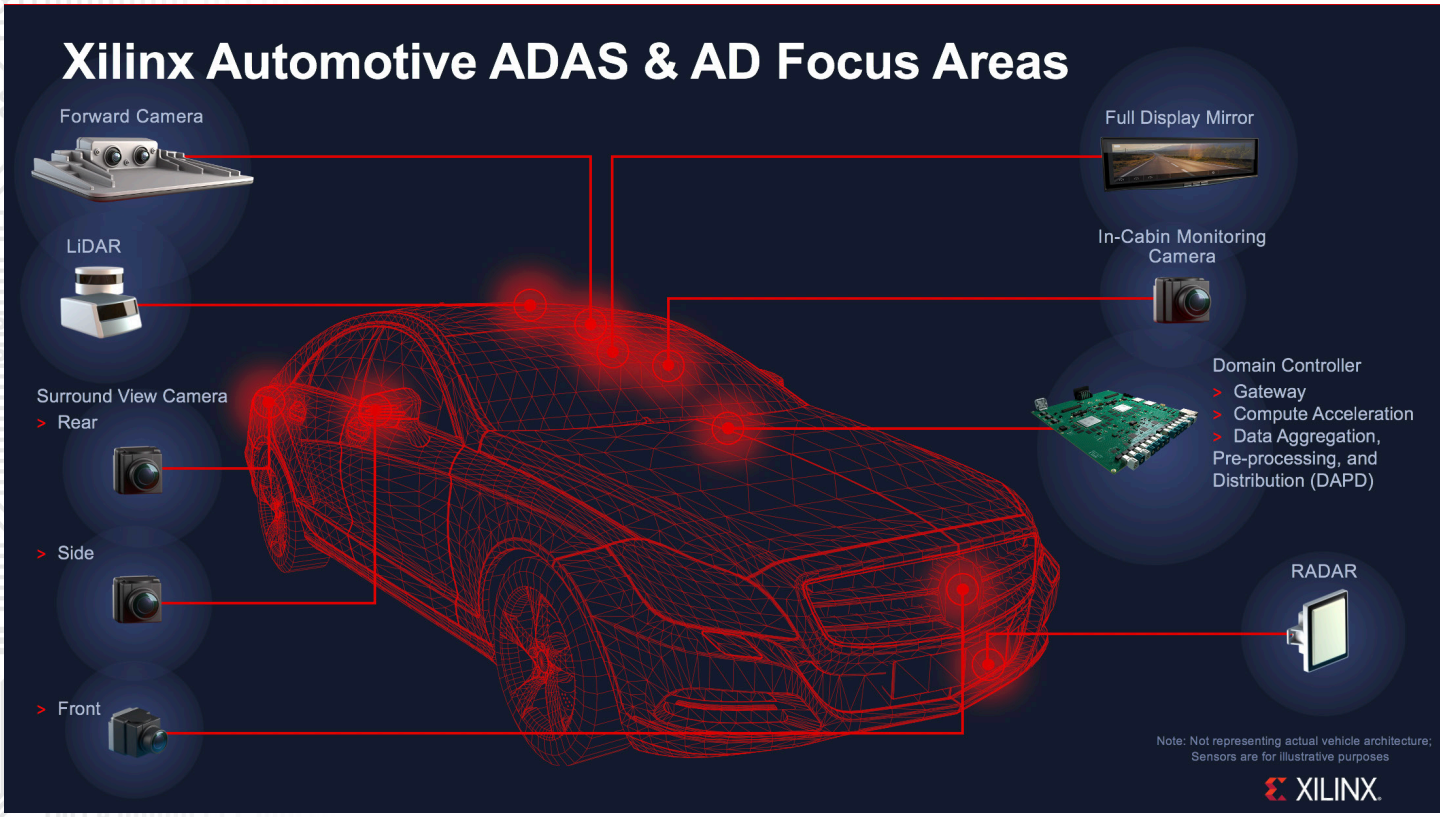
Nevertheless, the industry at large and design engineers worldwide are indeed busy sorting it all out. In recent years, the automotive industry has started to push beyond the first generation of ADAS solutions to what's commonly pitched as L2+, or conditional autonomy. These vehicles are going beyond basic adaptive cruise control, lane-keeping, and automatic emergency braking, to begin the process of breaking from full dependency on the driver (albeit ever so slight).

Surround perception capabilities enabled by deep learning-based vision are allowing vehicles to handle situations where lanes split or merge and safely perform lane changes. AI-centric system-on-chips (SoCs) and FPGAs process deep neural networks (DNNs) for perception as well as surround camera sensor data from outside the vehicle and inside the cabin.

Inside the cabin, features such as occupant monitoring and in-cabin visualization are rapidly evolving. In fact, vehicles featuring intelligent cockpit assistance and advanced visualization of the surrounding environment already surpass today's L2 ADAS offerings in performance, functionality, and road safety (**Fig. 1**). That's still a far cry from SAE J3016's highest levels of autonomy (**Fig 2**), but represents a significant jump in capability from just a few years ago.

Toward Higher Autonomy

Besides providing a cockpit rich in AI capabilities, deep learning-based algorithms employ vision processing to execute complex functions in urban environments and harsh weather. A set of advanced DNN technologies enables the vehicle to perceive a wide range



1. Vehicles integrating ADAS with sensors and processors enabling conditional autonomy (L2+) are nearing production. Source: Xilinx

of objects and driving situations.

All of these and other intelligent capabilities are powered, of course, by our friend the algorithm, which comes in all shapes and sizes in ADAS and AV applications to perform the tasks they are intended to serve. The continuous rendering, classification, and prediction of changes to everything in a vehicle's surrounding environment require algorithms to carry out the major tasks of regression analysis, pattern recognition, clustering analysis, and decision making (**Fig. 3**).

Automotive-grade processors for highly automated vehicles must run

SAE Levels of Vehicle Autonomy	
Level 0: No Automation	System has no vehicle control, but may issue warnings.
Level 1: Driver Assistance	Driver must be ready to take control at any time. Automated system may include features such as Adaptive Cruise Control (ACC), Parking Assistance with automated steering, and Lane Keeping Assistance (LKA) Type II in any combination.
Level 2: Partial Automation	The driver is obliged to detect objects and events and respond if the automated system fails to respond properly. The automated system executes accelerating, braking, and steering, and can deactivate immediately upon takeover by the driver.
Level 3: Conditional Automation	Within known, limited environments (such as freeways), the driver can safely turn attention away from driving tasks but must be prepared to respond if alerted to intervene.
Level 4: High Automation	The automated system can control the vehicle in all but a few environments such as severe weather. The driver must enable the automated system only when it is safe to do so. When enabled, driver attention is not required.
Level 5: Full Automation	Other than setting the destination and starting the system, no human intervention is required. The automatic system can drive to any location where it is legal to drive.

2. Summary of SAE levels of vehicle autonomy. See full description and graphic of the SAE J3016 standard.

ALGORITHMS IN ADAS AND AV SYSTEMS
REGRESSION
MAJOR TYPES: Bayesian, Neural Network, and Decision Forest • Used to predict events based on repetition in an environment. • Forms a statistical model of the relationship between an image and the position of a specific object within it. • Analysis is dependent on the number of independent variables, the type of dependent variables, and the shape of the regression line.
PATTERN RECOGNITION
MAJOR TYPES: Support for Vector Machines with Histogram of Oriented Gradients (HOG), Principal Component Analysis (PCA), Bayes Decision Rule, and K-Nearest Neighbor (KNN) • Used for data reduction and classification. • Sensor data is filtered by detecting object edges. • Line segments and arcs are applied to fit all object edges. • Segments and arcs are recombined until the features match a known object.
CLUSTERING
MAJOR TYPES: K-Means and Multi-Class Neural Networks • Used to predict events based on repetition in an environment. • Forms a statistical model of the relationship between an image and the position of a specific object within it. • Analysis is dependent on the number of independent variables, the type of dependent variables, and the shape of the regression line.
DECISION MATRIX
MAJOR TYPES: Gradient Boosting (GDM) and AdaBoosting • Determines the vehicle's actions, e.g. turn right, turn left, brake and accelerate. • Analyzes and rates the performance of relationships between datasets and their information. • Action depends on classification, recognition, and prediction of necessary next movement. • Predictions of multiple decision models are synthesized to create a final prediction of the condition and minimize error.

3. The most common categories of algorithms used in automotive applications.

different neural network algorithms on multiple compute engines. These processors must also accommodate fast-changing AI algorithms and provide flexibility in data pipelines to reduce AI latency.

Clearing the Processing Hurdle

It's important to note that the earliest ADAS designs had discrete architectures with modest processing power and limited sensor suites. As a result, most L2 ADAS systems have offered inconsistent vehicle detection and limited ability to stay within lanes on curvy or hilly roads. Even adaptive cruise control systems haven't lived up entirely to consumer expectations.

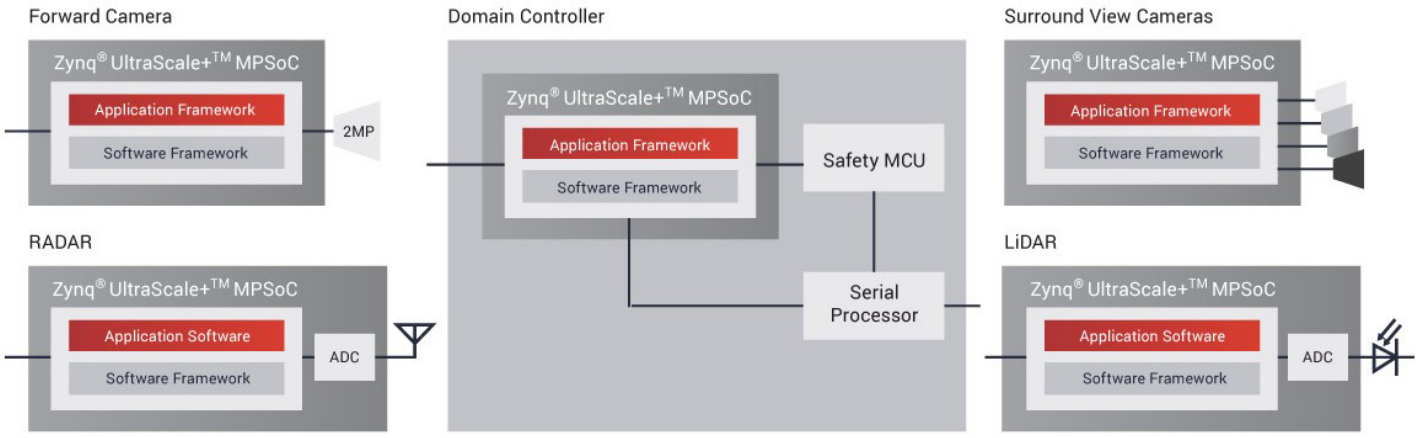
The limitations of earlier ADAS designs led to a high occurrence of system disengagements requiring the driver to take control abruptly. Today's designs are increasingly employing radar and LiDAR (albeit still rudimentary), both of which generate massive amounts of data, further increasing the processing requirements of the sensor modules. Inevitably, highly integrated chips will be instrumental in processing complex sensory data coming from a variety of sensors, including image sensors, radar, LiDAR, ultrasound and others. They must handle all of the data from the car's sensors with far greater speed and efficient processing than most of today's off-the-shelf AI chips.

In effect, this has created a mounting challenge requiring faster, more efficient processing architectures to handle increased penetration of not just one but many sensor types. Take the example of a camera-only subsystem for object detection that must run as much as six different algorithms. Four different algorithms are then added to make the jump from 2D to 3D imaging. Add 12 cameras creating a complete 360-degree hemispheric view of the vehicle and its surroundings, and it's easy to see the challenge facing the industry on just the processing front.

Beyond Confusion to Fusion

While ADAS and tomorrow's full-blown AV designs have parallels, they generally have different development paths and design teams working on the necessary sensor suites, processors, and design architectures.

Take the case of multiple sensory modalities—intelligent vision, radar, and LiDAR—and



4. The data aggregation, pre-processing and distribution (DAPD) capability improves AI processing by fusing sensor data and preparing it for processing by the performance modules. Source: Xilinx

sensor fusion, which will enable the giant leap from Level 2 drive-assist to the highly automated Level 4. True AV designs like this might feature 30 (and likely more) sensors across all sensor modalities to perceive the surrounding environment. Diving deeper into the processing functions and architectures required for true AVs provides a glimpse into what developers are preoccupied with right now: data aggregation, pre-processing and distribution (DAPD), and compute acceleration (**Fig. 4**).

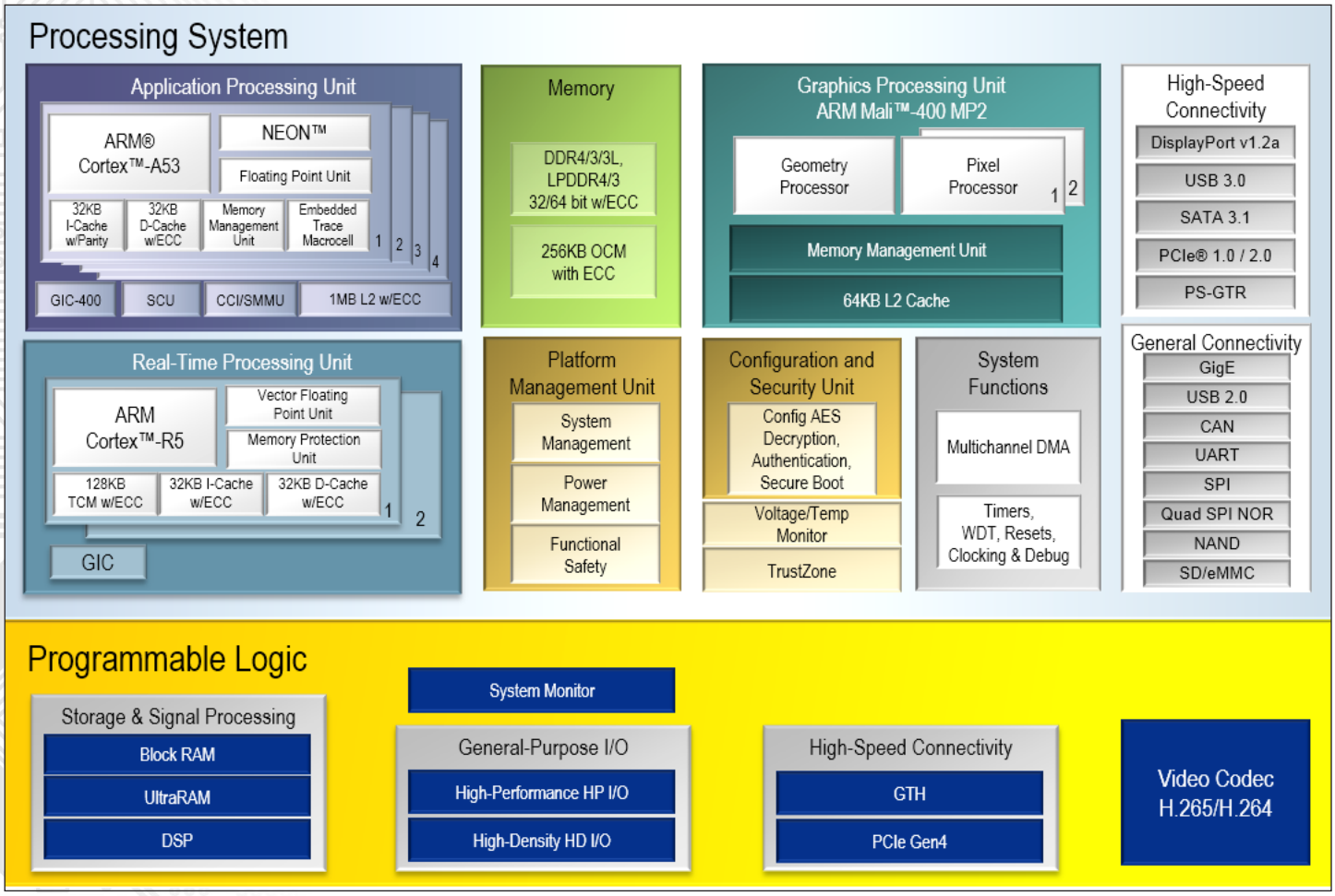
The XA Zynq UltraScale+ MPSoC 7EV and 11EG chips from Xilinx are a case in point. These ASIL-C certified 16-nm chips, targeted at L2+ ADAS to L4 AV applications, integrate programmable logic as well as 64-bit quad-core Arm Cortex-A53 and a dual-core Arm Cortex-R5 based processing system. The 504,000 logic cells and 1,728 DSP slices in the 7EV, and more than 650,000 logic cells and 2,928 DSP slices in the 11 EV raise programmability to a new level in the automotive application space. The other automotive-qualified devices in the XA portfolio (2EG, 3EG, 4EV, and 5EV) provide a complete range of options to fit every need in today's automotive applications.

The XA 7EV device contains a video codec unit for H.264/H.265 encode and decode, while the XA 11EG device provides 32 12.5 Gbps transceivers and four PCIe Gen3x16 blocks (**Fig. 5**). With these highly integrated chips, automotive developers have begun eyeing supervised self-driving on the highway, from on-ramp to off-ramp, as a ripe opportunity.

In addition to highway merge, notable capabilities such as lane change, lane splits, and path planning are all part of it. Here, the job of AI algorithms is to help vehicles understand where other vehicles are, read lane markings, detect pedestrians and cyclists, distinguish different types of lights and their colors, recognize traffic signs, and understand complex scenes.

GPUs, SoCs and FPGAs

In the world of automated driving technology, two current design trends are clearly apparent. First and foremost is computational horsepower to support more sophisticated AI algorithms. Not surprisingly, today's automotive system designers rely on highly integrated chips to manage complex software applications, real-time data processing, and functional safety.



5. The XA Zynq UltraScale+ MPSoC 7EV platform offers diverse processing engines to support features such as sensor fusion, AI compute acceleration, and functional safety. Source: Xilinx

On the higher end, there are powerful SoCs and MPSoCs that incorporate GPU cores and offer whopping teraflops. These GPUs come integrated with large AI models for faster acceleration, lower latency, and higher resolution. Then there are tightly integrated, purpose-built ASICs to handle all the data from vehicle sensors and cater to unique processing requirements. On the lower end, there are AI chips that run tiny machine learning models, but they often pose accuracy trade-offs.

Somewhere in the middle are FPGAs that perform batchless inference to ensure low and deterministic latency and higher throughput. On the other hand, powerful GPUs carrying out deep-learning inferences require batches of massively parallel data to go through single-instruction multiple data (SIMD) for doing more computing and less fetching. That, however, makes register files wide. Also, unlike ASICs, which are hardened in an instruction set, FPGAs allow designers to apply proprietary instruction sets on a compute-efficient platform and even enable engineers to tweak instruction sets to try new things.

It's also worth mentioning that FPGAs, like GPUs, have been used for AI acceleration in data center environments. So, for L2+ ADAS and AV designs, their DSPs and parallel architectures make FPGAs well suited for neural network acceleration.

AI acceleration is going to be crucial in enhancing image quality for AVs, especially in

low-light conditions. Here, the AI-enabled FPGAs can perform a lot of demanding video capture and processing tasks without changing the camera hardware.

Take, for instance, Baidu's production-ready Automated Valet Parking (AVP) feature, which is part of the company's in-vehicle computing platform for autonomous driving. The AVP system, an ingredient of Baidu's Apollo Computing Unit (ACU), employs Xilinx's XA Zynq UltraScale+ MPSoC for sensor fusion and AI processing for five cameras and 12 ultrasonic radars. Baidu claims its Apollo Project is the world's first open AV platform.

The second notable trend involves combining modular hardware with open software architecture, simply because traditional compute models with a fixed combination of hardware and software is reaching the end of its useful life (more on that in chapter 3). So, while SoCs bake-in AI algorithms for tasks like vision processing into the chip, FPGAs allow automotive OEMs and Tier 1 suppliers to update and tweak processing requirements for new AI algorithms.

FPGA-based camera solutions are a case in point; they allow developers to add new AI algorithms months and years after the camera is installed in the vehicle. That shows how an open platform can facilitate the customized integration of new software algorithms over time—which is a smart way to future-proof a solution.

On Track for Arrival

Given the highly consumer-centric nature of the automotive industry, it's apparent that implementing technology for technology's sake won't fly in the trek toward the fully autonomous vehicle. Nevertheless, a closer look at the AV development track shows that the automotive industry at large has clearly set course for the AV. While the true AV is still some years away, the faint rumble of development has become a steady drumbeat of significant breakthroughs recently.

At the system level, it's also becoming apparent that general-purpose CPUs, GPUs, off-the-shelf AI chips and the like is not where things are headed in the highly specialized world of AV designs. Some chipmakers already provide specialized solutions for ADAS and AV designs with a full hardware and software stack along with software development kits.

While these and other recent advances in the AV design realm may be viewed as incremental, they are nonetheless significant. In fact, we may be at the tipping point we've seen play out time and time again in other technology areas: Not here today … and everywhere tomorrow.

With the support of adaptable Xilinx solutions, and Avnet's integration expertise, you can be sure that you're able to keep up with the quickly evolving demands of the industry. Contact us to learn more.

BACK TO TABLE OF CONTENTS

Credit: Yongnian Gui | Dreamstime

CHAPTER 2:

Data Centers: Accelerating Into the AV Curve

JACK BROWNE, Citadel Engineering

As cars get smarter and more connected, data centers prepare for a driverless future.

While today's most advanced vehicles still require the driver to control the vast majority of driving tasks, advanced driver assistance systems (ADAS) are making their way into more driving functions with each new model year. It will be some time before a majority of cars on the road represent SAE Level 5 autonomous vehicles (AVs), but the technology needed to get us there is developing rapidly. Artificial intelligence (AI) and machine learning (ML) are already guiding vehicles along their routes at automotive proving grounds and other test routes, acting on hazardous conditions with increasing accuracy, and capturing "memories" to learn from each trip. But with all the excitement swirling around on-board AV tech, you have to wonder ... what does all this mean for data centers?

The Digital Traffic Jam

The amount of data generated by multiple cameras, radar, LiDAR system modules, in-cabin monitoring systems, GPS, and other sensor types can be overwhelming for even the best-equipped vehicle electronics system. The amount of data generated by today's ADAS test vehicles is in the trillions of bytes per day. Although on-board processing is among the largest hurdles to clear, data centers also have a heavy lift ahead to prepare for what's coming.

ADAS vehicles have often been called data centers on wheels, with many different sensor subsystems. Data captured by ADAS vehicles will be networked to other ADAS vehicles by means of a cloud-based networking environment. Enormous amounts of data will be collected for analysis and updated regularly through the cloud. The data will also be transferred for universal use by the entire vehicle-to-everything (V2X) infrastructure,

while other data will be shared directly between vehicles within range using vehicle-to-vehicle (V2V) communications.

It has been estimated that as much as 4 TB of data will be captured during a typical day of city driving—and three times as much for robo-taxis since they operate continuously. Properly managing the data is essential to achieving a safe, ADAS-guided roadway–which will require a new category of data center dedicated to handling and communicating enormous amounts of streaming data reliably. A lot of the processing needed to execute the most crucial real-time system responses will occur right on the vehicle's central processing modules, but enormous amounts of data will still stream beyond the vehicle.

But to where? In recent years, various communications giants have teamed up with automakers to answer that question. For now, the idea is to rethink the current network topology of data center deployment globally to better support the IoT generally—and specifically connected cars since they will be among the largest data generators of all.

Rethinking the Data Center

Today's data centers were originally designed to serve the needs of consumers and enterprises. Supporting millions of 2-ton wheel-mounted IoT devices was not in scope. Sure, they can provide cloud and Internet access, but not at the data transfer requirements projected for ADAS-equipped vehicles and future AVs. Add the expected explosion in specialized AVs delivering everything from packages to pizza, and it becomes clear how large the need is. In effect, it means building an entirely new category of data center for AVs and V2X at large.

The data processing and networking needs of an ADAS-equipped vehicle are both enormous and unique, requiring low-latency, wide-bandwidth data access to minimize data transfer times—even with tons of on-board or "edge" processing. The AV will require extremely fast data access for parallel streams of video, 4D imaging radar, LiDAR, ultrasound, and sensor fusion processing. Within the vehicle's on-board computer system, the data will be used in combination with AI and ML algorithms to make split-second decisions—possibly faster and with more precision than a human driver—to ensure the correct and safest system response.

Data processing speeds may not be as critical for data backups and software updates, but access to multiple protocols (such as NFS, SMB, FTP, and HTTP) will be required. Because of the large amounts of data being processed and stored, the data center interconnections (DCIs) must be more reliable than typically required for "general purpose" applications.

AVs must also adapt to constantly changing conditions. Through AI, they will turn sensor data into vehicle control data, but will also need communications about the surrounding environment. This is integral to the entire rethink going on with data center topology. Adequate coverage may be partly facilitated through more nimble "near-edge" AV and IoT compute centers rather than large, traditional data centers. These smaller, highly distributed data centers could effectively address the impact of distance to the data center as it relates to latency, in addition to reducing the processing load on all nodes. In terms of form factor, brick and mortar won't play a big part in these distributed data centers, if at all. (Think big metal boxes on rooftops, along interstates, and the like.)

The ANSI/TIA-942 standard created in 2005 by the American National Standards Institute (ANSI) and Telecommunications Industries Association (TIA) provide guidelines for

the location, architecture, security, and telecom requirements of new data centers. Data centers supporting tomorrow's vehicles will have heightened requirements for many of the same areas, and new ones as well. They must support high-density architectures that keep processing speed and performance up, heat down, and space requirements low.

Accelerating Dynamic Workloads

Regardless of how the world's data center topology shakes out, today's data center teams are already grappling with the growing data load and need for more speed due to the IoT and ADAS systems that already stream data to the cloud.

The Xilinx Alveo portfolio of accelerators is one well-known data center acceleration solution. As the industry's first comprehensive smartNIC offering true convergence of network, storage, and compute acceleration functions on a single platform, Alveo addresses the ever-changing scaling needs of cloud data centers to support the intense workloads produced by AVs, such as pre-acceleration and core CPU compute offload.

Xilinx's Alveo Data Center accelerators are compact SmartNICs that accelerate dynamic workloads, adapting to constant algorithm optimizations faster than fixed-function accelerators.

These modular accelerators already serve as data-processing compute engines in traditional data centers, providing significant increases in data processing compared to traditional CPUs—especially for ML, video transcoding, database search and analytic functions. Xilinx's unified software platform, Vitis, also makes accelerating dynamic workloads end-to-end easier than ever with its integrated development environment for programming, profiling, and debugging accelerated applications with powerful domain-specific libraries.

Unlike fixed-function computing engines, modular accelerators are highly adaptable to changing operating conditions and requirements as is routine with ADAS vehicles. They enable data center operators to make programming and operating changes not possible with other integrated-circuit (IC) processing engines, including ASICs and even GPUs.

Alveo accelerators are based on Xilinx's 16-nm Zynq UltraScale+ silicon IC technology. As smartNICs, the Alveo U25, Alveo U50, Alveo U200, Alveo U250, and Alveo U280 are integrated programmable FPGAs compatible with all Ethernet standards and certified to FCC, UL, CE, and RoHS hardware requirements. Supplied with an application development tool, the accelerators enable direct access to the cloud to simplify the development of new ADAS algorithms, significantly increase real-time ML throughput, and accelerate vehicle camera data processing.

The large Alveo portfolio also meets the expanding scope of acceleration needs in today's data center. For example, the Alveo U25 boosts the speed of cloud-based applications with its low-latency kernel bypass capabilities. Data is highly synchronized by means of an on-board Stratum 3 clock oscillator. The Alveo U25 features 6GB DDR4 RAM and measures just 6.60 x 2.54 in. (167.65 x 64.4 mm). When networking speed is important but power consumption is a concern, the PCIe Gen4-compatible Alveo U50 features a 100-GbE network interface with 8 GB HBM2 memory, 316 GB/s HBM2 bandwidth, and 872,000 look-up tables (LUTs)—yet it consumes no more than 75 watts.

When more internal memory is needed, the Alveo U200 and U250 both boast a DDR memory bandwidth of 77 GB/s and a DDR capacity of 64GB to handle the large amounts of data generated by an ADAS vehicle's cameras, LiDAR, and radar systems. The Alveo U200 has an internal SRAM memory bandwidth of 31 TB/s with 892,000 LUTs, while the Alveo U250 accelerator card features internal SRAM bandwidth of 38 TB/s with 1,341,000 LUTs for large amounts of digitized camera images.

For data center developers seeking the highest-speed solutions for compute-intensive applications, the Alveo U280 packs 8 GB of HBM2 460 GB/s bandwidth, a DDR capacity of 32 GB with 38 GB/sec bandwidth, and 2x PCIe Gen4 x8 compatibility for advanced server interconnects.

Getting Ready to Roll

As AV technology evolves to higher SAE levels of autonomous driving, more data centers will be designed to meet very different requirements. Designed to support advances in machine learning, autonomous vehicles, and the IoT at large, their storage capacity and computing power will dwarf today's requirements.

Much of that power and capacity will be deployed at the network edge, scattered in and around everywhere from densely populated areas to the boondocks, receiving and processing the data from millions of connected cars. That—combined with a full suite of intelligent sensor systems and end-to-end adaptive compute acceleration—could get us rolling down the driverless road sooner than we expected.

With the support of adaptable Xilinx solutions, and Avnet's integration expertise, you can be sure that you're able to keep up with the quickly evolving demands of the industry. Contact us to learn more.

☞ **BACK TO TABLE OF CONTENTS**

CHAPTER 3:

Evolving the Heterogeneous Model for Adaptive Compute

TERENCE POJE, Contextas Communications, LLC

As heterogeneous architectures take hold everywhere, the silicon roadmap is being redrawn.

Today's automotive industry is in the throes of an extremely disruptive period in its history, largely driven by autonomous systems and functionalities that futurists first envisioned more than a century ago. Today's automobile is at the crossroads of countless technology trends—from artificial intelligence, cloud computing and IoT, to next-generation processors and computing architectures.

Arguably, modern vehicles are confronted by the most challenging operational environment of any existing application space. Reducing bandwidth requirements, accelerating system response times, and adding more logic are never-ending goals. Fail-proof interoperability with external systems while maintaining core functionality is another giant order to fill—be it a self-guiding warehouse forklift or a self-driving car navigating the perils of a congested city.

Computing's Seismic Shift

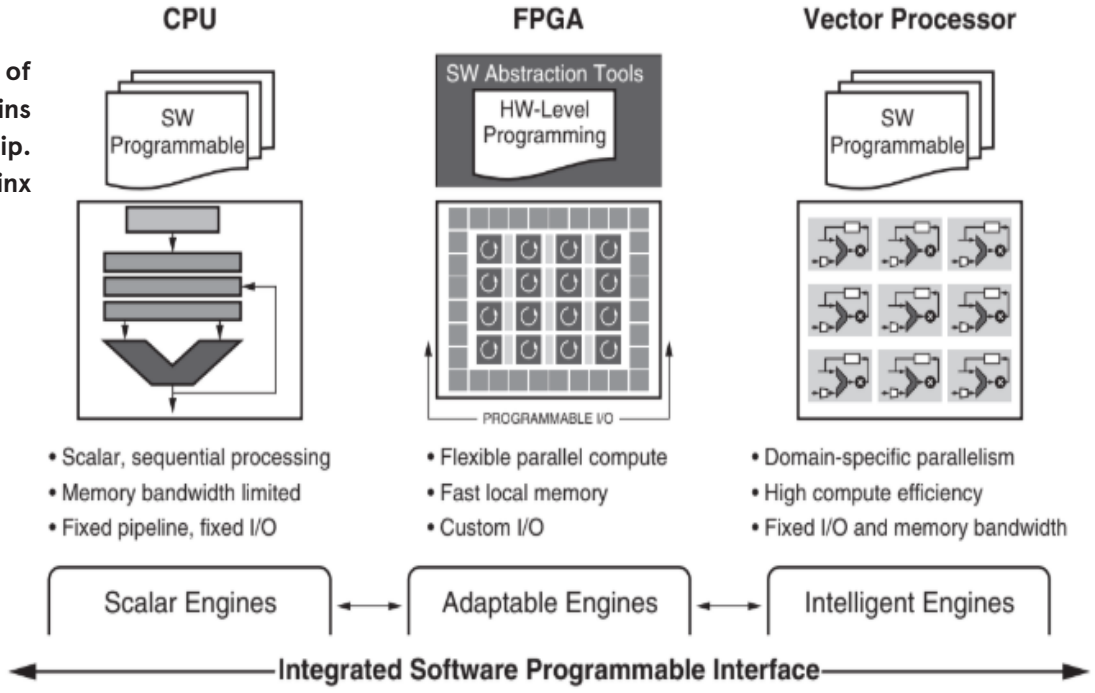
For more than a decade, the rise of AI and big data has been fueling a seismic shift in semiconductor technology. As the traditional "one-size-fits-all" scalar compute engine inches ever-closer to the outer limits of Moore's Law and Dennard Scaling, the industry is pursuing new heterogeneous system architectures (HSAs) that address the inherent limitations and disadvantages of traditional compute engines and processing methods.

Due to the latency and response issues of cloud-based self-driving systems, cars still must have all the processing power and logic needed for all critical operations on board at all times. Advanced levels of heterogeneous computing leverages the advantages of various domain-specific architectures to create more powerful, efficient, highly adaptable silicon solutions that simplify programmability, acceleration, and algorithmic

innovation. All compute-intensive industries realize this is where they need to head, but most agree it's where things *must* head to ensure the AV's continued progress toward higher levels of autonomous driving.

Heterogeneous Computing Defined

Heterogeneous system architecture employs parallel processing rather than increasing clock frequency in a concurrent processing model. This architecture can leverage any combination of scalar, vector and programmable logic processors specified for the application to create more powerful, efficient, highly adaptable silicon solutions (**Fig. 1**).



1. Key characteristics of various customized domains on a heterogeneous chip. Courtesy: Xilinx

By executing increasingly complex, compute-intensive system responses driven by intelligent camera systems, radar, LiDAR, ultrasound and other technologies onto a single multi-core processor with mixed safety-critical and non-critical aspects, tighter software integration can be achieved while creating significant cost, weight, size and power savings.

By leveraging heterogeneous computing and virtualization, multiple operating systems with mixed-criticality requirements can share the same hardware, managed by a host software hypervisor. This evolutionary progression to include high-throughput in-vehicle networking and high data-rate connectivity to the cloud will certainly continue. This in turn will increase the demand for high-performance smart gateways and domain-controller Electronic Control Units (ECUs). Such a solution can deliver powerful on-board processing capability while handling firewall functionalities, predictive maintenance, Over-The-Air (OTA) software upgrades, and high data-rate communication among the different ECUs and the cloud.

MPSoCs: Building on a Strong Foundation

For obvious reasons, hardware acceleration is a major driver of the heterogeneous compute trend. One way to address the challenge of system performance has been to

offload functions from the CPU to other types of processors more suited for the task. By definition, a multicore chip has the different processor cores needed for task- and functionality-sharing without needing any external processors.

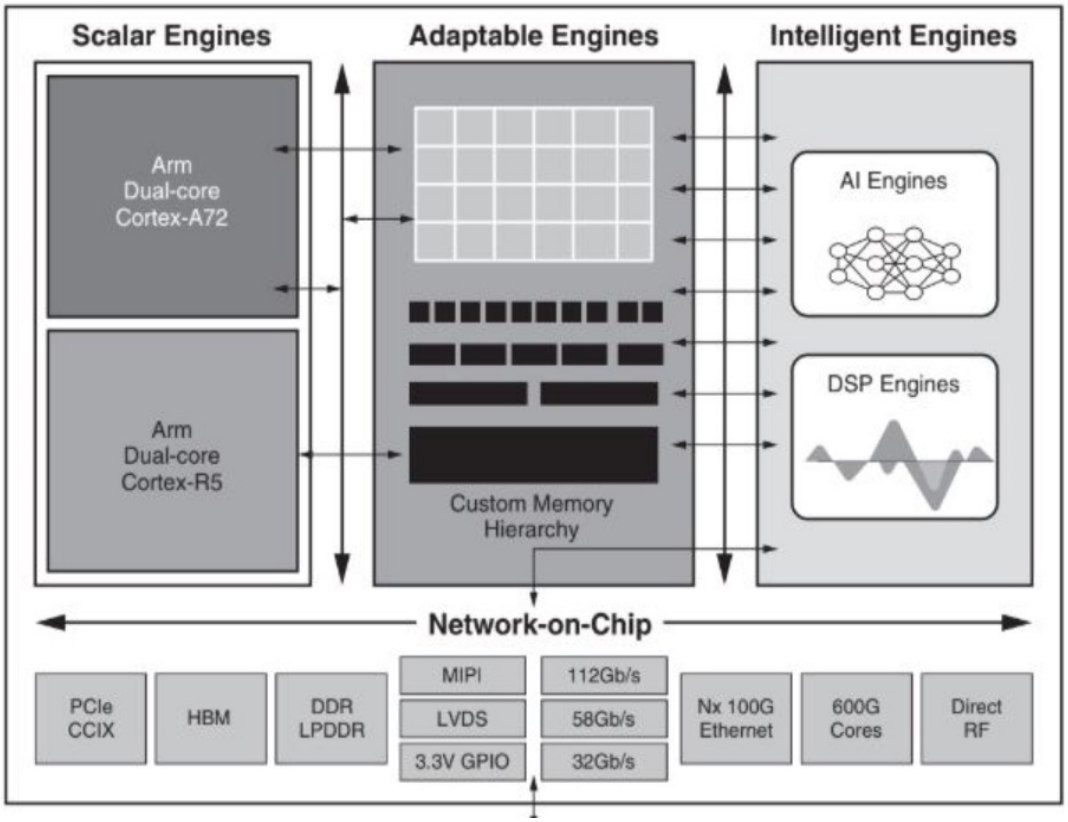
In recent years, multi-processor SOCs (MPSoCs) have shown their superiority in accelerating both processing speed and execution speed up. Usually targeted for advanced embedded applications, MPSoCs contain multiple heterogeneous processing elements with specific functionalities, a memory hierarchy, and I/O components. Everything is linked to each other by an on-chip interconnect.

The first all-programmable MPSoC, launched in 2014 by Xilinx, redefined what could be achieved with a heterogeneous chip. The new UltraScale+ MPSoC architecture extended the company's ASIC-class UltraScale+ FPGA and 3D IC architecture to implement heterogeneous multi-processing with the right engines for the right tasks, the industry's fastest fin field-effect transistors (FinFets), and tools to support programmability and design abstraction not seen before.

In 2018, Xilinx expanded its UltraScale+ portfolio to include the automotive-qualified XA Zynq UltraScale+ MPSoC family for safety-critical ADAS and autonomous driving systems. Baidu's automated valet parking platform is among the best-known applications of the XA Zynq, which is used for sensor fusion and AI processing in the company's Apollo Computing Unit.

Versal ACAP: Beyond the Multiprocessor

Today, next-generation computing solutions focus on evolving heterogeneous chips beyond their limits on every performance front. In 2019, a new category of heterogeneous



2. Functional diagram of the Versal ACAP. Courtesy: Xilinx

compute platform was introduced that brought that roadmap into focus: the Adaptive Compute Acceleration Platform, or ACAP.

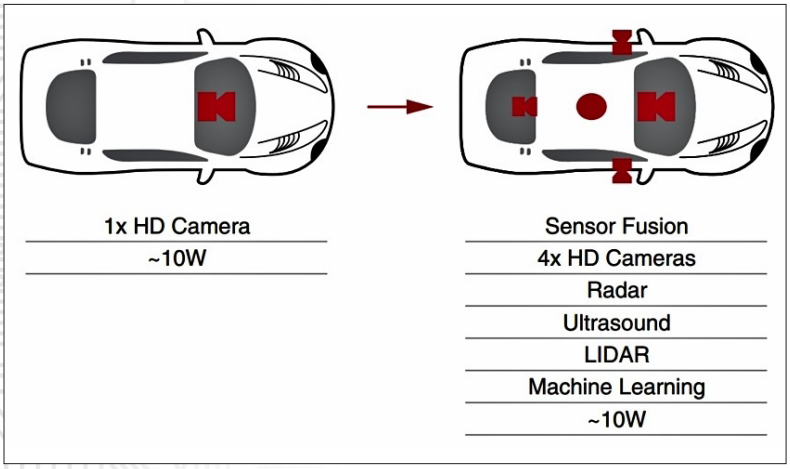
The Versal ACAP, developed by Xilinx, is a fully software-programmable solution that supports customized domain-specific architectures (DSAs). As the industry's first ACAP, the solution integrates all processing elements while dramatically simplifying the programming with a unified toolchain supporting a variety of abstractions, from framework to C to RTL-level coding (**Fig. 2**).

The Versal architecture combines scalar, adaptable, and intelligent engines—tied together as a network-on-chip (NoC) via a memory-mapped interface and hardened domain-specific interfaces—allowing more dramatic customization and performance. The power improvements gained from the 7nm process yield a 2X improvement in DMIPs/watt over 16nm implementations. The NoC makes each hardware element and the soft-IP modules easily accessible to each other, further reducing latency by accelerating communication between the engines and the interface logic.

The massive memory bandwidth enabled by the programmable logic and integrated RAM blocks enables programmable memory hierarchies optimized for individual compute tasks. This prevents the high latency and latency uncertainty inherent in other cache-based compute units.

The multi-core CPU provides comprehensive embedded compute resources for the remaining application needs, and the system is designed to be easily programmable without requiring hardware expertise. In addition to Versal ACAP's advanced programmability and acceleration, this new product category also takes the FPGA concept of in-field programmable logic a step further by enabling dynamic reconfiguration—accelerating configuration time by swapping partial bitstreams in milliseconds.

In order for tomorrow's AV to perform its multitude of roles safely and reliably, it has to integrate the input from various sensor types. Known as sensor fusion, this involves more than just combining the input from multiple cameras to stitch together a bigger picture. In addition to performing as desired, AVs must do so with a high throughput efficiency while conforming to the highest ASIL standards (**Fig. 3**).



3. Xilinx ACAP devices enable sensor fusion with high throughput efficiency in small power envelopes.

Data from all on-board sensing technologies must be integrated in such a way to provide increased situational awareness and operational redundancy. In a vehicle equipped with a wide array of sensor systems, what the radar picks up is augmented with feedback from what is seen by the surround view cameras, what is detected by the LiDAR system, and what is heard by the ultrasound system—as well what the V2X network adds to the mix. Many believe that achieving this level of redundancy and data integration is the only viable way for Level 4 and 5 AVs to meet or surpass the capabilities of the human driver while achieving the highest Automotive Safety Integrity Levels (ASIL).

The First True 5G Radio-On-Chip

The expectations for silicon in terms of wireless capability have become as challenging as the traditional concerns of size, weight, power and cost (SWaP-C). As 5G takes hold in automotive and other compute-intensive applications, the drive to wider spectrum at lower cost and the addition of machine learning inference technologies in the radio have taken on immense importance. The base station poses one of the greatest challenges in this regard, because it will have to support dense environments and the heavy data loads of future V2X networks.

With the advent of 5G, vector DSP-based ASICs have seen success in reducing cost at wider spectrums. The recent trend away from FPGAs back to ASICs was provoked by 5G's heightened demands. While FPGAs could deliver the performance required for many of 5G's digital components, they wouldn't always meet the low-power and economics test, and might not have the logic or on-chip memory required for AV applications.

Again, Versal ACAP may represent an attractive detour for automotive applications. For the first time, four key technologies have been implemented on a single chip to enhance beam steering algorithms and subscriber hand-off algorithms:

- Direct-RF sampling ADC and DAC
- Integrated SD-FEC codes
- High-density vector-based DSP
- Framework-programmable machine learning inference engines

The intelligent engine in Versal ACAPs largely addresses the cost issues of ASICs by delivering 5X more single-chip TMACs. Versal ACAPS are also designed to increase beam steering and subscriber hand-off algorithms by an additional factor of two over traditional programmatically defined algorithms, approaching 85% of the theoretical limit.

Riding Into the Future

Today's explosion in ADAS, more advanced AV technology, and the requirements of other compute-intensive applications have touched off a highly disruptive yet exciting period in the evolution of silicon and computing. In automotive specifically, on-board data from an array of intelligent systems, the emerging V2X infrastructure, and the need for real-time system responses have all contributed to a new breed of heterogeneous compute architectures powered by hardware that adapts to constantly changing conditions.

While mainstream adoption of the fully autonomous vehicle is still years away, its evolution from a completely closed, self-contained machine on wheels to a smart node in a larger intelligent system is becoming more of a reality every day. From new breakthroughs in silicon and AI to the connected infrastructures for V2X, the right technologies are beginning to coalesce into the viable, safe and commercially feasible solutions needed for tomorrow's AV.

With the support of adaptable Xilinx solutions, and Avnet's integration expertise, you can be sure that you're able to keep up with the quickly evolving demands of the industry. Contact us to learn more.

BACK TO TABLE OF CONTENTS

CHAPTER 4:

Unified Software Development from Vehicle to Cloud

TERENCE POJE, Contextas Communications, LLC

New software development flows are easing the journey toward the AV.

The disruptive trends we are seeing in the automotive industry are reaching far beyond the traditional forms of upheaval we've seen over the decades. Much of today's tumult is technology driven—from Advanced Driver Assistance Systems (ADAS) and V2X connectivity to vehicle electrification. As the industry takes a hard turn toward all things digital, automotive systems are undergoing a transformation with new compute models and architectures connecting the vehicle to everything around it and out to the cloud.

Combined, the advancements in machine learning, AI, sensor fusion, perception, and cloud computing are expanding the concept of vehicle "performance" as it has been understood for more than a hundred years. Whole electronic control units packed with silicon "engines" of all types are being powered by a new fuel—data and software.

Whole-Application Acceleration

Acceleration is among the most critical of today's performance factors in vehicle system design, but faster processing and execution speedup on selected components or chips is virtually useless in today's architectures. The key to understanding why this is rests in understanding the trends that are making whole-application acceleration a necessity—and what it means for software and AI developers.

The first is the clear trend toward heterogenous compute. Traditional CPUs have already become obsolete in today's more advanced vehicles. These and similar processing platforms cannot be scaled to effectively handle the intensive workloads created by enormous amounts of sensor data already being processed on ADAS-equipped vehicles. Parallel processing and domain specific architectures (DSAs) are

taking over these larger workloads, which are themselves in a constant state of flux. All of this can only be managed by algorithms that are trained to take on increasingly complex scenarios and workloads.

FPGAs and adaptive compute acceleration platforms (ACAPs) are gaining momentum in DSAs because they can be adapted across domains both on and off the vehicle and can reconfigure on the fly. The challenge lies in the programming and integration of all these heterogenous "actuators" across the domains of a vehicle and its network. Developing applications that span from the vehicle to the edge server and cloud is no simple task. And there is no "template" for it. Some application processes at the cloud level, such as analytics processing, might need to move to the edge or the vehicle subsystem to reduce latency, increase privacy, or meet other performance requirements. Other processes may typically reside on the vehicle and have to scale outward—be it for performance, cost, or both.

For the designer, this is never a once-and-done exercise. One project might require the analytics processing for an application to move from the cloud to an edge server, yet the next might require analytics to be retargeted onto the vehicle subsystem. And since AV technology involves a multitude of sensor technologies, the possible configurations from cloud to end point for each application are virtually endless. Short of qualified data and AI scientists suddenly becoming commonplace, retargeting applications again and again to meet new standards, core tech, and customer needs can get insurmountable.

Folding in the AI piece into heterogeneous compute deployments from cloud to edge is where things get even harder. Machine learning and AI developers need a way to implement their algorithms using machine learning frameworks like TensorFlow and Caffe, but more importantly they need an efficient way to integrate and deploy it all on adaptable devices throughout architectures to build pipelines of deep neural network (DNN) intelligence from vehicle to cloud, application by application.

Development Flows for Adaptable Compute

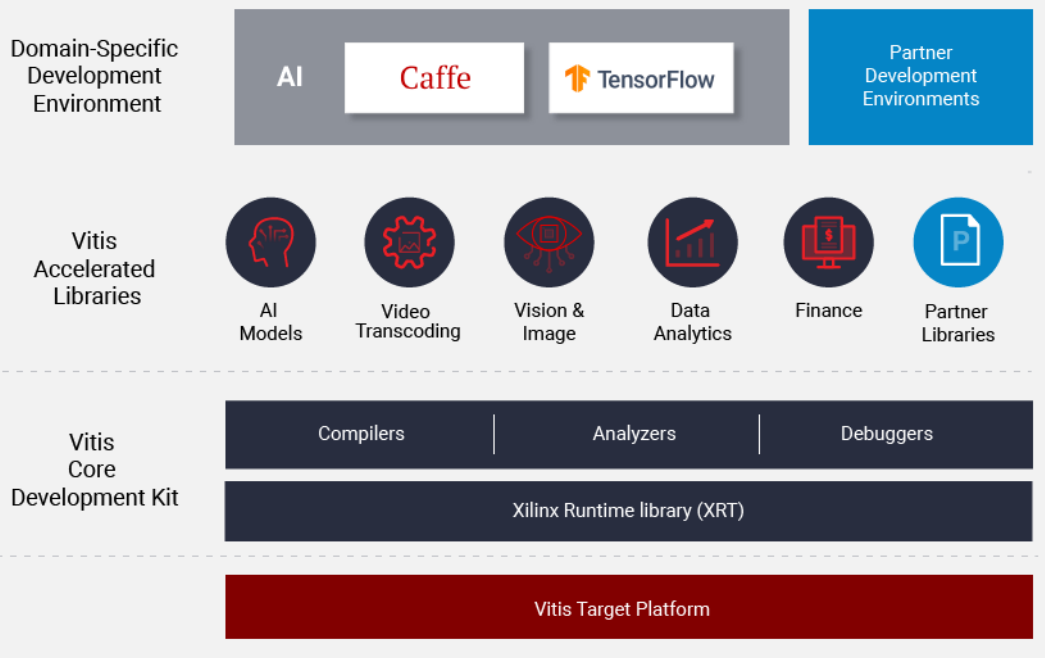
Unified software development environments, per se, are nothing new. But most were developed in a different time—which today means anything beyond a few years ago. Until recently, the heterogeneous compute train had barely left the station in automotive and other industrial markets. Domain-specific acceleration? Still quite new. As for automotive AI, what was state-the-art a couple years ago is starting to seem "quaint" today.

Clearly, new development flows are needed to support today's ADAS system architectures and the more advanced AV technologies we are starting to see. At issue is the increasing complexity of software development and cross-domain deployment in order to exploit the immense processing and acceleration power of today's best adaptable silicon solutions. In automotive and other compute-intensive applications, the emergence of adaptive compute acceleration platforms (ACAPs) means that many applications requiring programmability will be built on more than "simple" FPGAs. The difference is that ACAPs incorporate an array of adaptable engines—from programmable logic and intelligent engines (AI and DSP), to scalar engines (CPUs and Real-Time Processing Units or RPUs). This best-of-all-worlds approach not only addresses the scaling challenges of vector processing, but also significantly extends the capabilities of traditional FPGA fabric.

Admittedly, the programmable logic piece causes more than a few groans. While FPGAs have always offered great configurability, they're notorious for their programming challenges since they use hardware descriptive language (HDL), which is not as intuitive as C++ or Python. Thankfully, we are beginning to see that hurdle come down. Today, new development environments are emerging to help application teams harness the power of ACAPs and other complex hardware without having to understand its underlying details or hire specialized programmers to design the hardware acceleration.

The Rise of Software-Defined Hardware

One of the latest solutions to deal with the development flow issue is Vitis, Xilinx's answer to unified software development for intelligent, heterogeneous compute systems (**Fig 1**). Vitis offers a single design methodology and programming model for deploying accelerated applications on all Xilinx platforms. Vitis and Vitis AI accelerated libraries allow end-to-end application acceleration using a purely software-defined flow, with no hardware expertise required since all hardware is composed in software.

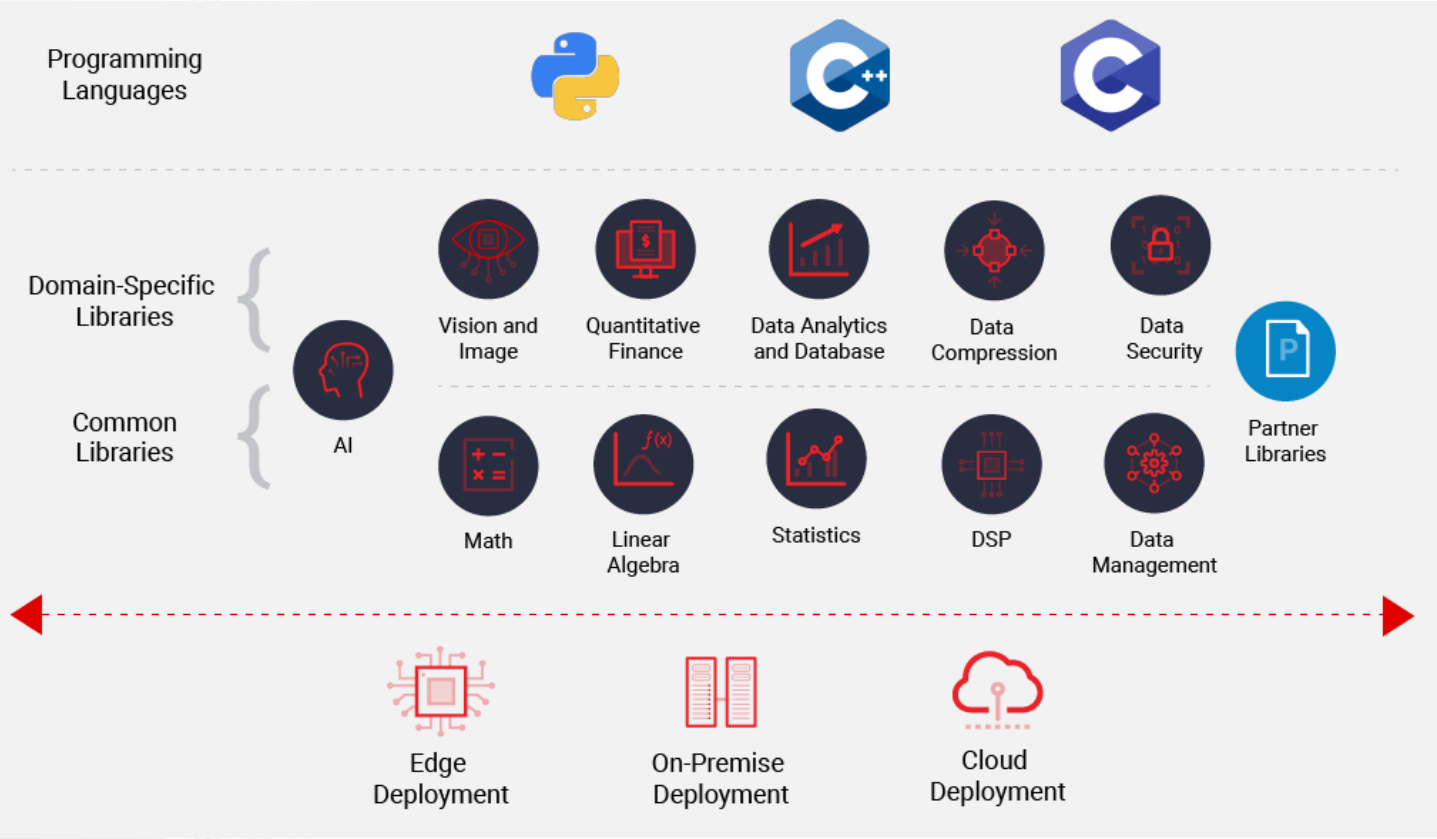


1. The Vitis platform enables end-to-end application acceleration using a purely software-defined development flow.

The Vitis core development kit contains an open-source runtime library that manages the data movement between various computing domains and subsystems. Designers can start with predefined target platform definitions for Xilinx SoC, MPSoC, RFSoC, and ACAP embedded devices, or simply import them if created in the Vivado Design Suite. For on-premise or cloud accelerator cards, the PCIe communication interfaces are automatically configured, eliminating the drudgery of implementing the connections. The comprehensive set of compilers for building the host program and kernel code, analyzers for profiling and performance analysis, and debuggers enable applications to be built, finetuned, targeted, retargeted, deployed, and scaled out as needed.

Going far beyond the norm in terms of available resources typically found in similar solutions, Xilinx and its partners have developed more than 400 pre-optimized, open-source libraries enabling developers to accelerate their applications with little or no reprogramming (**Fig. 2**).

In addition to cutting application prototyping to a fraction of what it normally takes, the libraries provide a roadmap for partitioning between computation, memory, and data movement resources. Then in the blink of an eye, Vitis automatically distributes the application functions as C/C++ blocks for the high-level-synthesis (HLS) compiler—eliminating the tedious work of RTL, synthesis, place and route and other not-so-fun tasks.

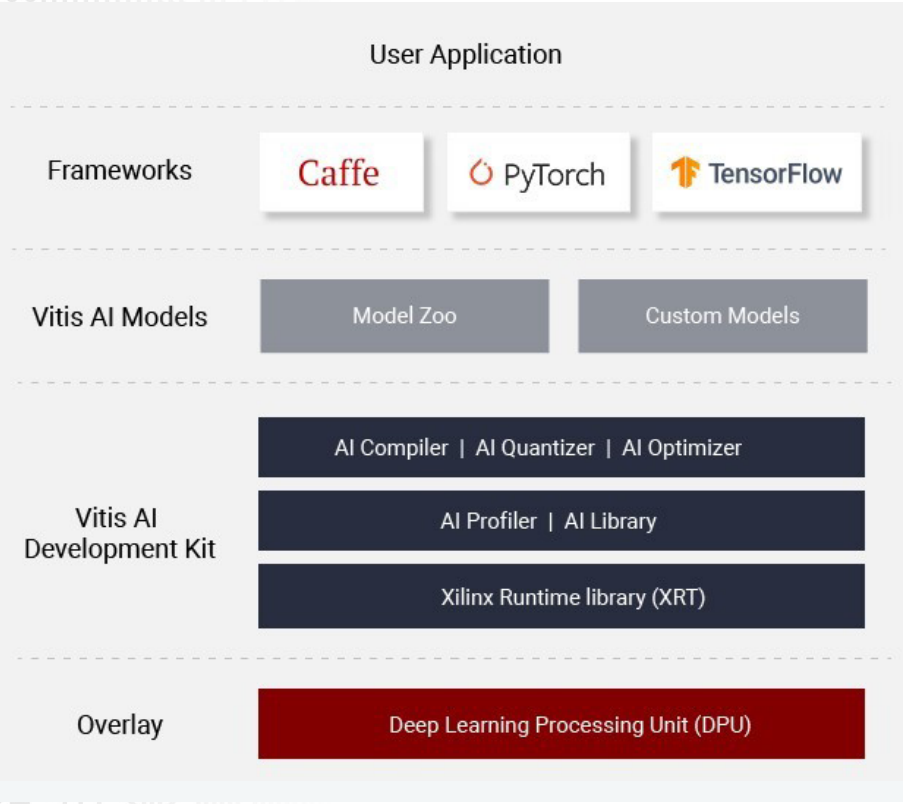


2. Vitis offers hundreds of domain-specific and general-purpose libraries that greatly simplify acceleration and partitioning.

AI Inference Acceleration

Vitis AI is one of the platform's biggest payloads, because it fills the critical need for AI developers to implement algorithms on FPGAs and ACAPs using mainstream frameworks and easy-to-use APIs (**Fig. 3**). After an AI model is initially trained, the tool's performance benefits really kick in during the deployment phase. In fact, this is where complexity usually derails the best of applications, keeping them in a perpetual state of "almost done." Once deployed, it's time to parallelize, quantize, and prune the network to improve execution speed, reduce latency, trim power consumption, and meet other performance requirements from inference to cost.

Vitis AI addresses this issue with its AI Quantizer, converting a model's original 32-bit floating-point weights and activations to fixed-point INT8 or INT4, preserving prediction integrity and saving gobs of memory bandwidth. This is supported by



3. Vitis AI drastically simplifies algorithmic innovation and deployment on hardware while streamlining all inference optimization processes.

Vitis AI's aptly named "Model Zoo," which contains optimized deep learning models for inference acceleration in compute-intensive applications, including ADAS and Autonomous Drive (AD), video surveillance, robotics, data centers, and others. The AI library contains a set of high-level deep-learning neural network libraries and APIs built for efficient AI inference.

The libraries and APIs found in the Vitis AI Library also help developers achieve high inference efficiency with little experience in deep-learning or FPGAs. Widely adopted algorithms used in ADAS and AV designs are included for image and video, classification, semantic segmentation, and object detection and tracking. The Deep-Learning Processor Unit (DPU) is designed to accelerate the computing workloads of deep learning inference and fully supports the Xilinx Runtime library (XRT), which runs on the Xilinx target platform's ARM processor in edge applications, or on the x86-based CPU in the case of Alveo Accelerator cards in the cloud.

A Smoother Ride Ahead?

Clearly, today's rapidly evolving AV technologies have raised the ante for a single design methodology and programming model for deploying accelerated, intelligent applications from vehicle to cloud. Much like integrated hardware development platforms revolutionized automotive subsystem design, it seems like software's take on the same idea is finally materializing in a comprehensive way. Sure, there's more work to do and wrinkles to iron out, but the bumpy ride toward the autonomous vehicle is undoubtedly getting a lot smoother.

With the support of adaptable Xilinx solutions, and Avnet's integration expertise, you can be sure that you're able to keep up with the quickly evolving demands of the industry. Contact us to learn more.

☞ BACK TO TABLE OF CONTENTS