2018 Ushers in a Renewed Push to the Edge

achronix
SEMICONDUCTOR CORPORATION

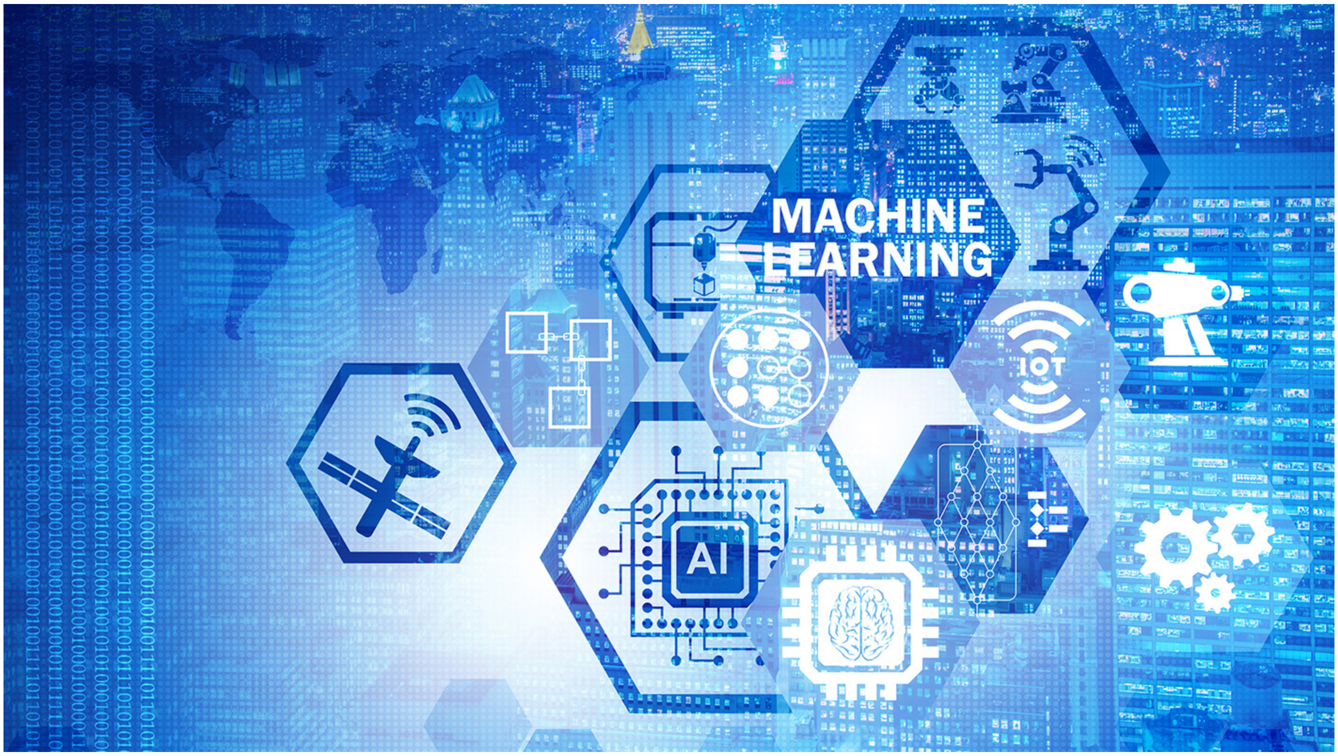
January 08, 2018 **WP012**

The past decade has seen massive growth in centralized computing, with data processing flowing to the cloud to take advantage of low-cost dedicated data centers. It was a trend that seemed at odds with the general trend in computing — a trend that started with the mainframe but moved progressively towards ambient intelligence and the internet of things (IoT). As we move into 2018, this centralization is reaching its limit. The volume of data that will be needed to drive the next wave of applications is beginning to force a change in direction.

Currently, just 10% of enterprise-generated data is created and processed outside centralized data centers. Industry analyst Gartner predicts this figure will reach 50% by 2022. This necessary reversal is driven by the shift towards hyper-connected cyber-physical systems, enabled by the arrival of technologies such as 5G wireless communications and a new wave of application-focused computing hardware.

The first wave of the IoT devices generated small individual amounts of data that were most efficiently aggregated and processed *en masse* in large data centers. But as IoT devices and, increasingly, cyber-physical systems come to rely on the ability to interpret much more substantial data streams, the center of gravity needs to shift towards the edge.

The automotive sector provides a microcosm of the changes that will be seen across many sectors. For example, the aggregation of GPS data from automobiles enables the collection of information on traffic congestion. Reflected back to the in-car units, it guides drivers to freer-flowing routes. Weeks of repeated data capture have had the longer-term benefit of allowing systems running in data centers to learn where lanes run on freeways from the passage of thousands of vehicles. These systems can then react to changes in traffic flow caused by construction work and adapt their maps dynamically. The result is much more accurate, live digital maps — all achieved without having to send out large numbers of survey vehicles.



Automobile makers have realized that maintaining data in centralized servers can only go so far. As a result, they are now moving quickly to build greater levels of autonomy into their vehicles. Today, the intelligence in automated driver assistance systems (ADAS) is largely self-contained, with scenes captured by the built-in cameras and radar systems processed purely within the vehicle. Only a tiny proportion of this data is relayed to the carmaker's servers where it may be used to update databases to help with predictive maintenance and collect statistics on the performance of the ADAS software.

Every mile an ADAS-equipped car travels generates gigabytes of data per mile, but bandwidth and processing limitations hinder its use. This information is processed once and then quickly discarded, being too dense to send to centralized cloud servers; however, this data contains insights that could be valuable to many systems. But systems much closer to the vehicle can make use of high-speed, cost-effective wireless networks, such as the IEEE 802.11p protocol devised for V2X communications, to capture the data and make informed decisions based on it.

Roadside beacons and smart traffic signals will cooperate with moving vehicles to optimize road usage as autonomy becomes increasingly pervasive. In a simple case, transmissions from passing vehicles could use V2X to relay data to roadside beacons about road surface conditions they have encountered. These beacons themselves may be in isolated locations with only a low-speed connection to the cloud. Rather than discard much of the data before delivery and processing in the cloud, the beacons could use their own local compute capability to learn about road conditions and send that information to vehicles passing in the other direction.

Similarly, smart traffic signals could capture data and images from vehicles to help determine the locations of pedestrians and other vulnerable road users, as well as the automobiles themselves. This data could help their software make intelligent decisions as to when they should change for optimum traffic flow. Roadside beacons and vehicles could begin to cooperate to provide *see-through* functions, assisting maneuvers such as overtaking and lane changes by determining when it is safest to do so.

Autonomy Needs Real-Time Performance

As the level of autonomy increases, the need for low-latency, real-time response becomes more vital — milliseconds matter when vehicles pass at speed. Today's wireless networks can impose round-trip latencies on the order of hundreds of milliseconds. 5G has been re-architected to respond to messages in as little as a single millisecond. But that impressive latency improvement needs to be compared to the delays incurred in moving data to the cloud.

A fundamental limitation of today's cloud-centric computing model lies in the laws of physics. The transmission of data by photons through fiber-optic cables is restricted to 70% of the speed of light in a vacuum. A distance of 1,000 km adds 10 ms of round-trip delay. There are also the delays incurred by multiple switch-routers and other networking infrastructure that needs to be added. To support the millisecond-scale response times, the compute resource needs to be brought much closer to the point of delivery — to *cloudlets* at the network's edge.

Smart transportation is only one example of the need for such edge computing. There will be other battery-powered robots emerging in the coming decade — machines that will move freely around the factory environment, around homes and, as drones, fly through the sky to survey land for farmers or to deliver packages to consumers. Their limits on power consumption will be a major obstacle to bringing onboard the sophisticated algorithms they need to run. 5G wireless networks could allow this data to be gathered from the various end-points (the *things* in IoT), but the fundamental limits of propagation time through the network mean that this data must be analyzed at the network's edge rather than in a central location. Thus processing will increasingly occur colocated with the wireless base station or nearby in cloudlets.

Another application where both real-time responsiveness and energy consumption impose severe limitations on design is in augmented reality (AR) and virtual reality (VR) systems. A fundamental problems of immersive virtual reality is the delay between when a user shifts their gaze and when the system provides the updated view in their headset. If the *motion-to-photon* delay increases beyond 20 ms, there is a high risk of the user suffering the effects of motion sickness. This restriction forces a difficult trade-off between battery power and the performance of modeling and rendering software running locally.

By offloading VR processing to nearby grid-powered servers, the energy consumed by the body-worn device drops significantly as does its heat output. The combination of 5G and nearby cloudlet processing can deliver high-performance scene generation and keep latencies well within the motion-to-photon limit. The same approach can be used to offload the energy-intensive AI operations needed by advanced drones and robots.

The bandwidth demands of AR and VR illustrate another dimension in which edge computing can service a growing need. The delivery of 4K and higher resolution video across the global Internet is already a costly exercise. In many cases, the content can serve multiple viewers in a particular locale. Therefore, using cloudlets to act as a further layer of smart caches in content-delivery networks can provide much greater capacity for high-priority communications and more efficient services on the network backbone.

Cloudlets and other edge computers can help extend the delivery of services to users who would otherwise be beyond the reach of cloud-based high-performance computing. Companies in fields such as resource extraction often need to work in inaccessible locations. By deploying micro data centers, mining specialists can harness high-performance simulation and AR tools to optimize extraction. The local compute resource can, at the same time, manage autonomous vehicles as they move material around the site. Such systems can use their own compute capacity to process, filter and compress data for upload to cloud servers overnight over slow connections such as satellite uplinks.

Computing at the Edge

The processing that cloudlets perform will include extensive data analytics, often based on machine-learning techniques. This trend to push AI technology to the edge of the network is already occurring with initiatives such as Amazon Web Services' Greengrass. This service was developed to deal with situations where IoT services are likely to encounter problems shipping data to the cloud at times but still require AI support whether server resources are available or not.

Initially, the training of these machine-learning algorithms will take place in the core cloud while the edge computing systems provide devices with the ability to offload inferencing so they are not burdened with its energy demands. But even training is likely to move to the edge, again because of the gravity of data. It will often prove impractical to upload enough data to enable good training even with high-ratio data compression.

Local training also enables systems to modify their behavior for the conditions they see rather than a national or global average. Smart traffic signals may learn local congestion patterns or use the fuel-efficiency traces of vehicles that pass through to optimize flow around them based on local conditions.

To properly serve these varied, highly responsive applications, a cloudlet or edge computer cannot simply be a scaled-down cloud server. High performance is necessary but it has to be delivered in a form that is compact, reliable, and energy efficient. These systems are likely to sit alongside communications equipment that may be as small as a roadside cabinet. Even those deployed on a campus may not have access to the same level of support as the blades in a core data center.

To maximize compute efficiency, hardware-acceleration technologies will play a key role in edge computers and cloudlets. Multicore processors on their own will be too slow and power-hungry to take care of tasks such as real-time machine learning. One option for acceleration may be to augment the multicore CPUs with a general-purpose graphics processing unit (GPGPU) or a vision processing unit (VPU). GPGPUs and VPUs are being used in some embedded systems to run machine-learning and data-analytics algorithms because of their highly parallelized floating-point arithmetic units. These units can sum the inputs for many neurons in parallel, and do so much faster than clusters of CPUs, although vendors of high-end CPUs have tried to close the gap thanks to their incorporation of massively parallel single-instruction, multiple-data (SIMD) units such as AVX512 from Intel and ARM's NEON.

The emphasis on peak floating-point performance in the GPGPU tends to make the architecture unsuitable for energy-constrained systems, often requiring sustained power levels that exceed 150W. Even within machine learning, there are drawbacks to using GPGPUs and VPUs as they are devices designed primarily for accelerating 2D and 3D graphics and imaging applications that involve operations such as convolutions. They lend themselves to the convolutional layers of deep neural networks, but other types of operations can prove troublesome in terms of memory access. Max pooling and fully connected layers place a greater emphasis on data transfers between the virtual neurons, using access patterns that do not fit their memory architecture. A further problem is that their emphasis on floating-point and matrix arithmetic makes GPGPUs and VPUs a poor fit for other applications that will need acceleration in edge computers. Solutions based around programmable hardware rather than processors provide the freedom to optimize data transfers between virtual neurons. Programmable hardware also provides the freedom to accommodate the wider range of tasks that edge computers will be called on to perform.

Distributed Communications

The edge-computing revolution will go hand-in-hand with the rollout of 5G. The architecture needed for the efficient delivery of 5G services will also rely on a cloudlet-based approach. The promise of millisecond-level response latency and data rates of up to 10Gb/s call for much more intelligence to be deployed within the network. Data from individual users could be switched quickly across different RF technologies, from sub-GHz protocols to millimeter-wave to WiFi as transmission conditions change. The cloud-RAN concept makes it possible to spread the workload of coordinating multiple RF transceivers across multiple servers and manage the resources dynamically as needs change.

There are numerous ways in which cloud-RAN implementations will be deployed. But the problems of latency and backhaul availability are likely to push cloud-RAN towards an edge-focused model. In principle, the most substantial efficiency gains achievable with cloud-RAN come from the aggregation of digital resources into a centralized location. Only the RF subsystems and conversion to baseband are performed at the transceiver location. All digitized data from layer one upwards is sent to the central unit.

The use of massive multiple input/multiple output (MIMO) and beam-forming techniques to focus transmissions on individual users, who may be moving at high speed, calls for highly accurate real-time estimation. This requirement, in turn, points to the deployment of signal-processing capacity to the edge of the network rather than the center. Multiple small servers could coordinate RF transceivers and their active antennas in concert.

The cloud-RAN concept itself fits into a larger software-defined networking (SDN) infrastructure. Backhaul connections to the core network as well as high-speed local access through fiber will use network function virtualization (NFV) applications to deliver advanced communications services to users and optimize routing. For example, an edge server may use virtualized network functions (VNFs) to analyze packets and route them based on their contents. Low-priority training data sent to the cloud may travel over links with best-effort delivery guarantees, reserving connections with high quality of service for traffic that supports key real-time applications.

In addition, as with large-scale cloud computing, edge computers and cloudlets are likely to make extensive use of containers and virtualization technologies. Together they enable easy migration between processor nodes for load balancing. Virtualized containers also maintain secure separation between the applications that will be uploaded and executed on behalf of different users.

As with those running in the core cloud, applications will communicate using standard network protocols. As cloud server-farm operators have found, this inter-application communication leads to a significant proportion of network traffic never encountering a physical cable, with much of the traffic being from virtual machine to virtual machine. This east-west traffic is readily supported by network functions but imposes an overhead on conventional CPUs that need to route packets between virtual machines or out onto the wider network.

In the cloud environment, the solution is to offload much of the packet handling to smart network interface cards (smart-NICs) that can move workload back off the server processors. Smart-NICs are usually built around low-power multicore CPUs assisted by hardware accelerators that employ a combination of fixed and programmable logic. Within the space-constrained environment of edge computing, programmable logic provides a solution for the acceleration requirements of data analytics, cloud-RAN and NFV services.

Data-Centric Processing

Programmable logic provides the ability to make computing much more data-centric. While traditional processors demand data be fed to their pipelines through a complex hierarchy of caches, programmable logic makes it possible to construct data pipelines. Data can flow seamlessly from node to node, with a combination of custom logic circuits and digital signal processing (DSP) engines manipulating the data elements as they pass through. Each element is readily forwarded to the next node that requires it. As needs change, the fabric can easily be rewired with a new configuration. This fabric provides better support for data-centric applications than the control-intensive code that better suits microprocessors.

However, standalone FPGAs generally incur a power penalty with the need to move data on and off chip frequently to more specialized ASICs. Embedded FPGA (eFPGA) technology provides a way to satisfy the constraints of energy efficiency, performance, and size in one package.

With eFPGA technology, it possible to take commonly used functions that would otherwise be deployed in standalone ASICs and implement them in custom hardware for higher performance and density. For machine-learning applications those functions may be dedicated processor arrays for convolution kernels or max-pooling calculations. By combining programmable logic and custom logic on the same IC, large power savings can be made by avoiding the need to transfer data off chip.

EFPGA technology has one other key advantage in the cloudlet or edge-computing environment — containers and virtualization provide effective support for secure operation in the core cloud because these systems can take advantage of good physical security. Devices on the edge of the network require greater levels of hardware protection as it is far easier for attackers to break into the enclosure and tamper with systems sitting in roadside cabinets or service rooms. The edge-computing systems will also have less support from administrators who can monitor server behavior and look for evidence of network-based intrusion, such as the uploading of malicious workloads that try to perform side-channel attacks.

Integrating security functions into the hardwired logic that surrounds the eFPGA cores makes it possible not just to support encrypted uploads of virtual circuits into the fabric but continually monitor them for potential breaches. The hardwired logic can ensure separation of programmable functions that may be uploaded by different users and prevent them from spying on each other.

Having both security and programmable logic integrated on-chip makes it extremely difficult if not impossible for an attacker with physical access to the system to eavesdrop on communications. With integrated CPUs, the compute functionality of entire services can be isolated to the eFPGA to limit the amount of information sent off chip. Communications with other services can be performed using strong encryption facilities baked into the hardwired logic. As a result, the embedded FPGA concept supports a strong security architecture suitable for the needs of edge computing.

Conclusion

The mixture of hardware flexibility with the security and performance guarantees of hardwired custom circuits makes the eFPGA a vital technology for edge computing and cloudlets. The journey towards edge computing is just beginning — the direction of these waves can not be ignored as they turn in 2018. The pull of data towards the growing base of real-time applications on the edge makes outflow from the cloud inevitable.

Achronix Semiconductor Corporation

2953 Bunker Hill Lane, Suite 101
Santa Clara, CA 95054
USA

Phone : 855.GHZ.FPGA (855.449.3742)
Fax : 408.286.3645
E-mail : info@achronix.com

NOTICE of DISCLAIMER: The information given in this document is believed to be accurate and reliable. However, Achronix Semiconductor Corporation does not give any representations or warranties as to the completeness or accuracy of such information and shall have no liability for the use of the information contained herein. Achronix Semiconductor Corporation reserves the right to make changes to this document and the information contained herein at any time and without notice. All Achronix trademarks, registered trademarks, disclaimers and patents are listed at http://www.achronix.com/legal.