Aitech

Embedded Computing *without* Compromise

Applying AI in Rugged Embedded Systems

Manage higher computation through parallel processing.

April 2020

Artificial intelligence (AI) has come a long way from merely carrying out automated tasks based on a data set. It is truly empowering embedded systems globally, providing 'rational' computer-based knowledge using data input. As the abilities of AI have matured, so have the methods to make data actionable as well as the amount of data and inputs that can be handled in any given system.

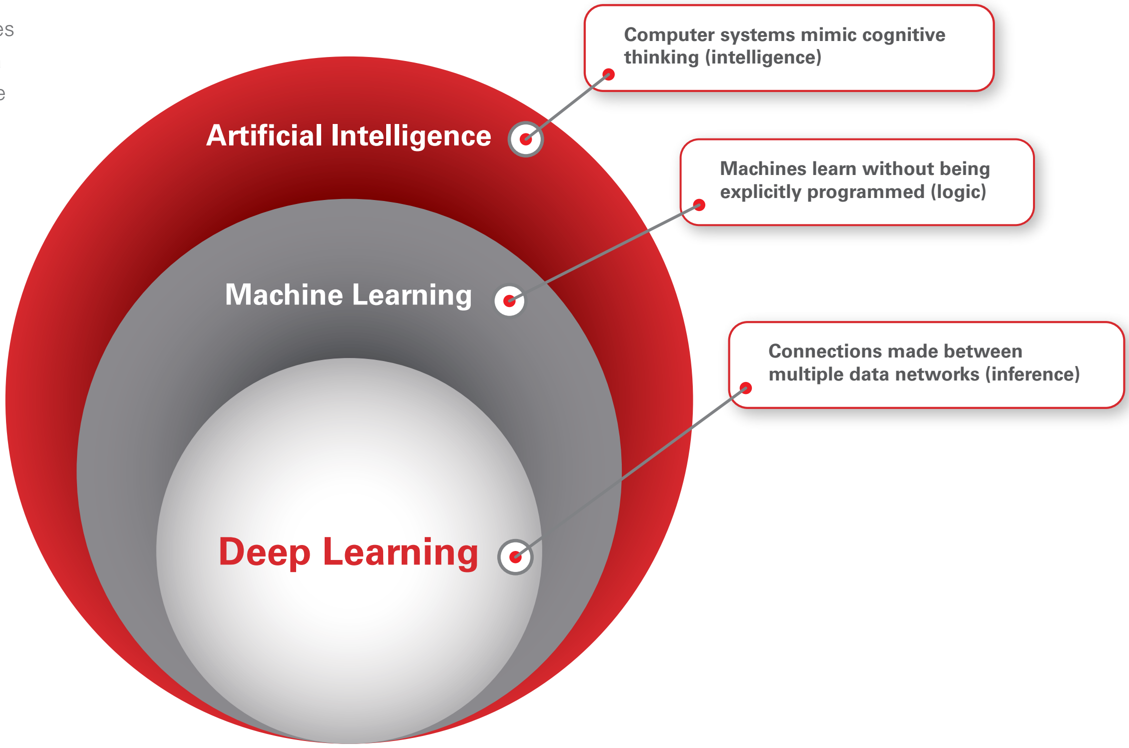
While shrinking geometries, improvements in ruggedization and general advancements in electronics have contributed to the expanded applications of today's embedded systems, it is the intuitive processing capabilities that have served as the catalyst to propel modern systems into this new realm of high intensity computing using real time data.

AI Now Models the Human Brain

As AI matured, it expanded into a new discipline called machine learning, where systems can learn from data inputs without being explicitly programmed. There is a logical component applied to actions the system deems appropriate and therefore executes. One step further is where we are today: deep learning...classified as a subset of machine learning.

FIGURE 1

Deep Learning, part of the broader family of Artificial Intelligence, makes connections between data to enable systems to make inferences that lead to credible, decision-based capabilities.


Computer systems mimic cognitive thinking (intelligence)
Artificial Intelligence
Machines learn without being explicitly programmed (logic)
Machine Learning
Connections made between multiple data networks (inference)
Deep Learning

Modeled after the brain's neural networks, deep learning makes connections between multiple data networks and optimizes those connections to enable a system to make realistic assumptions according to the data. Not only are logical choices made, but inferences across data paths can also be generated, leading to systems taking actions not previously prescribed, but rather that have been acquired through training and application of knowledge. Intelligence within the system continues to increase, with more accurate, quicker decisions made over time.

Data input, processing and clarity are all critical elements of properly applying this highly complex method of artificial intelligence. Leading this charge is GPGPU (general purpose calculation on graphics processing unit) technology, which is instrumental in managing the increased computational demands generated by this new paradigm of embedded processing. Nowhere is this more applicable than in mission-critical military operations.

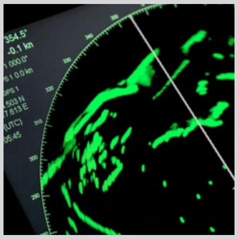
GPU Accelerated Computing Put into Practice

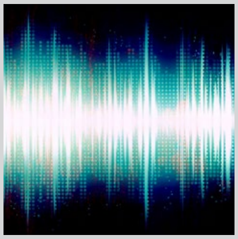
Technological improvements aimed at greater precision in weapon systems and military operations are highly regarded to ensure safety as well as more humane outcomes. When human lives will be directly impacted, completing a mission with fewer weapons expended and with less collateral damage is optimal. In addition, the use of AI-enabled, GPGPU-based HPEC systems to remotely-pilot vehicles can lessen the risk to military personnel by placing greater distance between them and danger.

Non-lethal defense-related activities can benefit from AI as well, including logistics, base operations, lifesaving battlefield medical assistance and casualty evacuation, navigation, communication, cyberdefense and intelligence analysis, to ensure military forces are safer and more effective. AI's role, and that of GPGPU technology, in these critical systems is growing to help protect people as well as prepare for and deter attacks.

The applications that benefit from GPU accelerated computing technology are numerous. In fact, any application involved with mathematical calculation can be a very good candidate for this technology, including:

- Image Processing – enemy detection, vehicle detection, missile guidance, obstacle detection, etc.
- Radar
- Sonar
- Video encoding and decoding (NTSC/PAL to H.264)
- Data encryption/decryption
- Database queries
- Motion Detection
- Video Stabilization





Optimizing High Performance Embedded Computing (HPEC)

For defense and military applications, AI has a unique opportunity to provide significant benefits across a range of activities. While industrial environments may garner financial and productivity benefits from implementing an AI-based strategy, mission-critical applications that protect human life and require extreme precision and accuracy are a different category altogether. Up-to-date operational intelligence is paramount to the success and safety of modern defense initiatives.

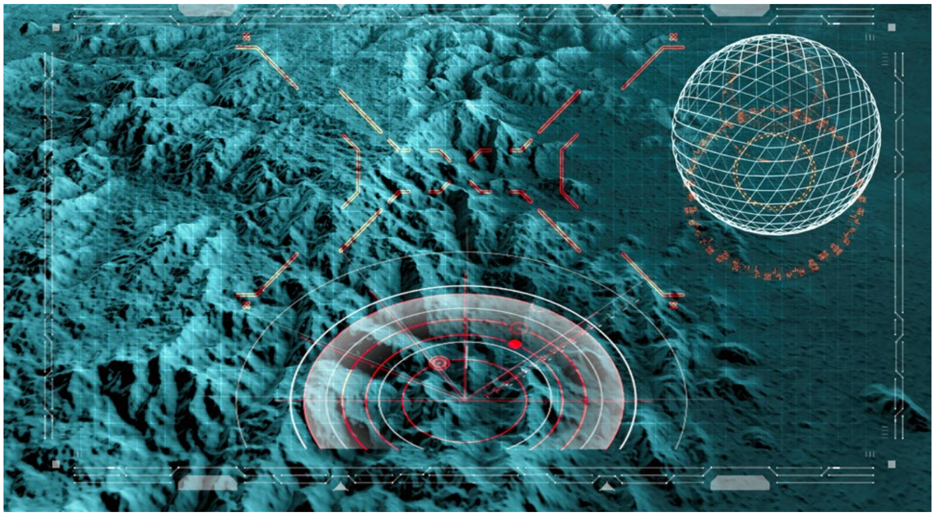


FIGURE 2

Up-to-date operational intelligence is paramount to the success and safety of modern defense initiatives.

The software applications used by complex embedded systems, in both consumer and military applications, is now far more intricate, putting added computation demands on the hardware. System architects, product managers and engineers must keep pace with the latest computing technology. When an engineer continues to reuse existing software applications, constantly adding new features and implementing new requirements, the code becomes increasingly complex, and the application grows CPU (central processing unit) “hungry”.

In addition to CPU choking that slows the operating systems response time, eventually you are faced with complex CPU load balancing – a constant struggle to satisfy conflicting software application demands – as well as other less-than-ideal methods of increasing computation power in an embedded system that can be costly (upgrading) or detrimental to component life (overclocking).

A Plan for Better Performance

Real time response applications are requiring systems that can perform AI processing at the sensors for "AI at the Edge" and for autonomous operations, exponentially increasing computing requirements. Using a GPU (graphics processing unit) instead of a CPU reduces development time and "squeezes" maximum performance per watt from the computation engine. The main reason this can happen is that GPUs use a parallel architecture, whereas CPUs are only serial in nature.

Thanks to the demands from the popular gaming industry, the GPU has not only increased system speed, but has evolved into an extremely flexible and powerful processor because of:

- **PROGRAMMABILITY**
- **PRECISION** (Floating Point Operations)
- **PERFORMANCE** - thousands of cores to process parallel workloads

GPU accelerated computing uses a GPU to accelerate the compute capabilities of a system by running compute intensive portions on the GPU, using less power and delivering higher performance over a CPU. Borrowed from the gaming industry, where graphics and data processing continue to set new limits, GPUs serve as the heart of these highly computation-intensive embedded systems. Through an increased power-to-performance ratio, GPU-based systems can meet the exorbitant calculation demands these applications now require.

As data needs continue to increase, modern embedded systems are faced with some serious performance issues: continuing to only use a CPU as a main computing engine would eventually choke the system. Divesting the highly demanding data calculations to the GPU, while allowing the rest of the application to run on the CPU, helps balance system abilities and resources more effectively.

With the end of Moore's Law in chip design edging closer to reality, expanding to a parallel processing architecture, where thousands of data streams are simultaneously captured, managed and made actionable, sets a path for the next generation of embedded data processing. Although some may try to apply Moore's Law to GPU accelerated computing, the recent development of a Multi-Chip Module GPU (MCM-GPU) architecture will enable continued GPU performance scaling, despite the slowdown of transistor scaling and photo reticle limitations, staving off the applicability of Moore's Law in GPGPU technologies for the foreseeable future.

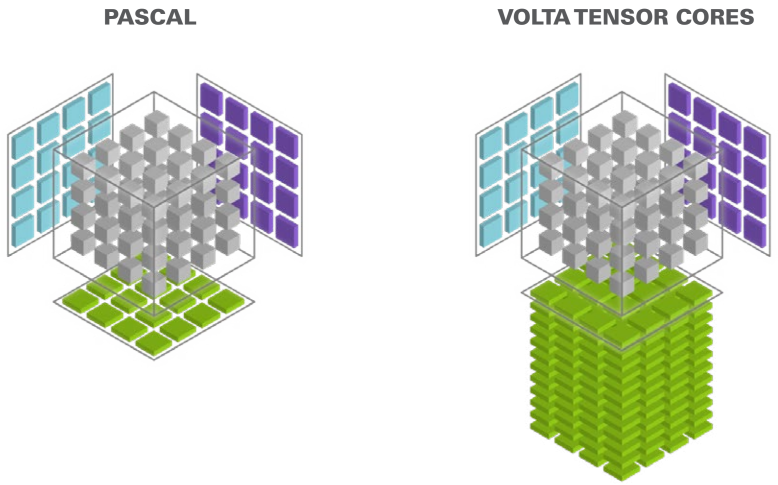


FIGURE 3

NVIDIA's new Volta architecture provides even more processing advantages than the groundbreaking Pascal architecture.

Graphic courtesy of NVIDIA

Parallel Processing Leading the Charge

Paradigms don't shift without cause. System architects, product managers and engineers are finding that traditional CPU-based systems can't keep pace with the growing amount of data streams within a given system or the processing requirements to manage the data. In addition to the influx of data sources, other culprits to system complexity include continued technology upgrades, shrinking system size and increasing densities within a system itself.

Balancing the load on a CPU can't always be accomplished with a simple board upgrade, as this may not be sufficient enough to manage the data processing a rugged embedded system will now require. The industry is adopting GPGPU capabilities for a reason: managing multiple streams of high definition graphics is literally what parallel processing was designed for. As the number of data inputs and image resolution continue to grow, the need for a parallel processing architecture will become the norm, not a luxury, especially for mission- and safety-critical industries that need to capture, compare, analyze and make decisions on several hundred images and data points simultaneously.

GPU accelerated computing has elevated beyond the gaming world and moved into other complex, highly sophisticated realms—and are enabling better intelligence across many industries by reliably managing higher data throughput and balancing system processing for more efficient computing operations.

Volta Amps Performance to New Levels

The application of NVIDIA's Compute Unified Device Architecture (CUDA) into non-gaming related embedded computing systems has redefined the parameters of data processing. And accelerating even beyond the trail-blazing processing capabilities of the initial Pascal architecture, Volta, NVIDIA's latest architecture, is providing even more compute power as data volumes continue to grow.

One of the biggest attributes of Volta is the redesigned streaming multiprocessor architecture that has been optimized for deep learning to enable more efficient workloads with a mix of computation and addressing calculations. Volta's new independent thread scheduling capability enables finer-grain synchronization and cooperation between parallel threads. And a 50% increase in energy efficiency means Volta can deliver the same FP32 and FP64 performance using the same power demands as Pascal. Tensor Cores designed specifically for deep learning deliver up to 12x higher peak TFLOPs for training.

Blazing Through TOPS of Data

With the introduction of NVIDIA's Volta architecture comes Tensor Cores, which amplify the matrix processing of large data sets, a critical function in AI environments, by enabling higher levels of computation with lower power consumption. For example, versions of the Volta are equipped with a full 640 Tensor Cores, each performing 64 floating-point fused-multiply-add (FMA) operations per clock. Up to 125 TFLOPS for training and inference applications are then delivered, enabling deep learning training using a mixed precision of FP16 compute with FP32 accumulate, achieving both a 3x speedup over the previous generation and convergence to a network's expected accuracy levels. The NVIDIA® Jetson AGX Xavier™ System on Module (SoM) combines the Volta GPU with a CPU, memory and many other processing elements, with an emphasis on inference over training, making it ideal for deep learning in embedded and edge based systems.

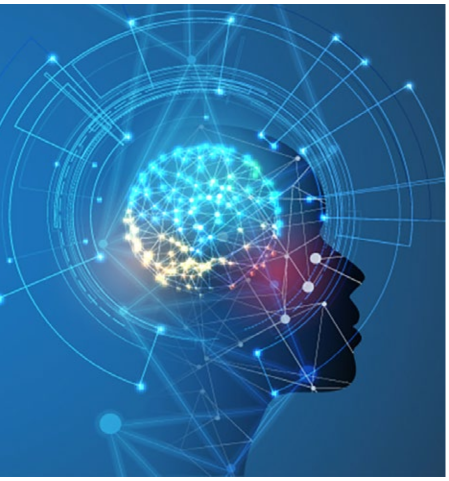
Xavier-based A178 Thunder

When the AGX Xavier SoM is coupled with Aitech's rugged computing expertise, the result is an AI supercomputer like the A178 Thunder, which handles up to 32 TOPS (trillion operations per second) to provide local processing of high volumes of data closest to the sensors, where it is needed. In addition, AGX Xavier-based systems typically feature twice as many CUDA cores as those using Jetson TX2 to offer some of the most powerful processing capabilities in an ultra-small form factor (SFF) system. The A178 system, for example, features 512 CUDA cores and 64 of NVIDIA's new Tensor Cores.

Using Open Source tools, developers can accomplish deep learning inference on the AGX Xavier, using the two NVIDIA deep learning accelerator (NVDLA) engines incorporated into the A178. This facilitates interoperability with modern deep learning networks and contributes to a unified growth of machine learning at scale. The system also features pre-installed Linux OS, which includes the bootloader, Linux kernel, NVIDIA drivers, an Aitech BSP and flash programming utilities.

Decreasing the time it takes a system to 'learn' a process within AI applications is a critical development area. Optimized software within the new Volta architecture is providing enhanced versions of deep learning frameworks (i.e. Caffe2, MXNet, CNTK, TensorFlow) to deliver dramatically faster training times and higher multi-node training performance. This is significantly advancing both deep learning and high-performance embedded computing (HPEC) applications. The new Multi-Process Service (MPS) feature provides hardware acceleration of critical components of the CUDA MPS server. The result is improved performance, isolation and better quality of service (QoS) for multiple compute-applications sharing the GPU, another attribute to facilitate improved deep learning.

Other areas where this new architecture is helping overall processing efficiency and resource management include more accurate migration of memory pages to the processor as well as the introduction of cooperative groups, a new programming model that organizes groups of communicating threads, so developers can express the granularity at which threads are communicating for richer, more efficient parallel decompositions.



Deep learning frameworks use multi-layered artificial neural networks to improve the rate of process training.

Potential Pitfalls in System Development

In order to reap the benefits of HPEC systems using GPU accelerated computing in military and defense operations, a system needs to be reliable. All that advanced processing won't mean a thing if the system is unable to withstand environmental factors and provide stable, long term operation. These systems are exposed to many of the challenges embedded designers face every day.

Ruggedizing the electronics is one, especially in military and mobile applications. Just like many parts and components used in a harsh environment, GPGPUs aren't rugged at manufacture. By applying the ruggedization expertise of board and system manufacturers to products based on GPU accelerated computing, advanced processing systems can reliably operate in remote, mobile and harsh environments, from industrial environments such as down-hole well monitoring and autonomous robotics systems to unmanned aircraft and ground vehicles as well as persistent video surveillance throughout military and defense operations.

This is where an understanding of how to design reliable systems for these environments becomes critical, including which techniques will best mitigate the effects of things like environmental hazards as well as ensure that systems meet specific application requirements. At Aitech, for example, our GPGPU-based boards and small form factor (SFF) systems are qualified for, and survive in, several avionics, naval, ground and mobile applications, thanks to the decades of expertise our team can apply to system development.

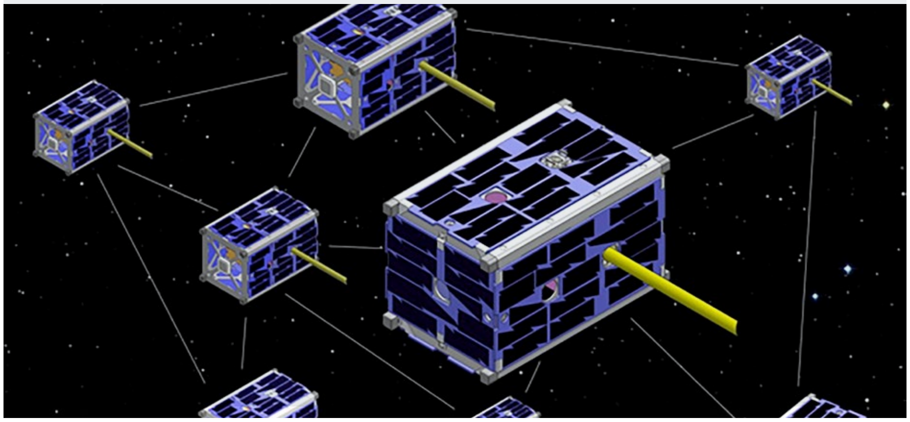
Managing power consumption is always a factor in a system's development, but because GPGPU boards process far more parallel data using thousands of CUDA cores, it's best to look at the positive impact of the power-to-performance ratio. In addition, GPGPU boards are very efficient, with some boards matching the power consumption of CPU boards. So, systems obtain more processing for the same, or slightly less, power.

Looking Beyond Land, Sea and Air

> "AI driven by Moore's Law and now super-fed by big data is in the midst of a true renaissance, becoming an integral part of our society, deeply transforming the way we work, operate, and live."
>
> *—European Space Agency, Towards a European AI for Earth Observation Research & Innovation (AI4EO R&I) Agenda, 2018*

Space is the newest frontier where GPU accelerated computing is being applied, as evidenced in several current undertakings by private and public space agencies alike. One of the biggest initiatives comes from the European Space Agency (ESA) to "promote the development of radically innovative technologies such as Artificial Intelligence (AI) capabilities onboard Earth Observation (EO) missions" to foster the use of AI technologies in space applications.

Dubbed Φ-Sat-2, the mission is a follow-on to the Φ-Sat, or PhiSat, experiment, which was the ESA's first demonstration of improved efficiency when reporting Earth observation data from space, using AI onboard a satellite. Φ-Sat-2 expands on the initial ESA project, launched in January 2020, by focusing on the disruptive potential of onboard AI using new, useful and innovative techniques. It specifically focuses on cubesat-based implementations...a perfect application for rugged, compact, space-rated systems using high data processing GPGPU technologies.



Balancing Power and Performance

And tradeoffs certainly still exist between performance and power consumption. Higher performance and faster throughputs require more power consumption. That's just a fact. But these are the same tradeoffs you find when using a CPU or any other processing unit.

As an example, take the "NVIDIA Optimus Technology" that Aitech is using, which is a compute GPU switching technology where the discrete GPU is handling all the rendering duties. The final image output to the display is still handled by the RISC CPU processor with its integrated graphics processor (IGP).

In effect, the RISC CPU's IGP is only being used as a simple display controller, resulting in a seamless, real time, flicker-free experience with no need to place the full burden of both image rendering and generation on the GPGPU or share CPU resources for image recognition across all of the RISC CPU. This load sharing or balancing is what makes these systems even more powerful.

When less critical or less demanding applications are running, the discrete GPU can be powered off and the Intel IGP handles both rendering and display calls to conserve power and provide the highest possible performance-to-power ratio.

Next-level Computing

So, with GPGPU-based processing, we are meeting the call for better intelligence, more intuitive computing capabilities and increased system performance. GPU accelerated computing has helped to elevate AI into new depths of learned intelligence, a world that can optimize complex, highly sophisticated computing across many industries by reliably managing higher data throughput and balancing system processing for more efficient computing operations. ■

Aitech

Embedded Computing *without* Compromise

19756 Prairie Street
Chatsworth, CA 91311
Toll Free: (888) Aitech-8 (248-3248)
Ph: (818) 700-2000
Fax: (818) 407-1502
sales@aitechsystems.com

No. 91 Prestige South End Road
560004 Basanvagudi
Bengaluru, Karnataka, India
Ph: +91-80-4866-8105
Fax: +91-80-4866-8106
Indiasales@aitechsystems.com

4 Maskit Street
PO Box 4128
Herzliya, 4673304 Israel
Ph: +972 (9) 960-0600
Fax: +972 (9) 954-4315
sales@aitechsystems.com

AitechSystems.com

The information in this document is subject to change without notice and should not be construed as a commitment by Aitech. While reasonable precautions have been taken, Aitech assumes no responsibility for any errors that may appear in this document. All products mentioned are trademarks or registered trademarks of their respective owners.