White Paper

GPGPU Computing for Next-Generation Mission-Critical Applications

www.adlinktech.com
Introduction

Recent developments in artificial intelligence (AI) and growing demands imposed by both the enterprise and military sectors call for heavy-lift tasks such as: rendering images for video; analysis of high-volumes of data in detecting risks and generating real-time countermeasures demanded by cybersecurity; as well as the software-based abstraction of complex hardware systems. While IC-level advancements in compute acceleration are not sufficiently meeting these demands, this discussion proposes heterogeneous computing as an alternative.

Among the many benefits that heterogeneous computing offers is the appropriate assignment of workloads to traditional CPUs, task-specific chips such as graphics processing units (GPUs), and the relatively new general purpose GPU (GPGPU). These chipsets, combined with application-specific hardware platforms developed by ADLINK in conjunction with NVIDIA provide solutions offering low power consumption, optimized performance, rugged construction, and small form factors (SFFs) for every industry across the globe.
In 2015 former Intel CEO Brian Krzanich announced that the doubling time for transistor density had slowed from the two years predicted by Moore’s law to two and a half years, and that their 10nm architecture had fallen still further behind that schedule (https://blogs.wsj.com/digits/2015/07/16/intel-rechisels-the-tablet-on-moores-law/). This is evidence that every shrinking of transistor size is becoming more difficult and costly to achieve. Similarly, thermal issues also continue to plague CPU evolution. To achieve traditional performance gains from each successive silicon generation, chips must make up in speed (frequency) for what they lack in increased transistor count. Unfortunately higher frequencies mean more heat generation.

While higher transistor counts and frequency increases are brute force approaches to CPU scaling, the limits of those methods are proving to be increasingly cumbersome. Efficiency optimizations are merely a stopgap solution to enhanced performance. We can trace the evolution of task-specific accelerators all the way from the Intel® MMX™ instruction set in 1997 to modern-day security enhancements like Intel® QuickAssist Technology (Intel® QAT). While MMX was the first of several accelerators designed specifically to improve graphics performance, accelerators can only make generic CPUs better at general purpose tasks. CPUs cannot beat the performance of mature task-specific chips such as graphics processing units (GPUs).

GPUs do the heavy lifting of rendering images and video for playback. These tasks involve complex mathematical and geometric calculations – many of them done simultaneously and in parallel. The need for highly parallel computing led to GPUs evolving along different architectural lines than CPUs. In time, software developers realized that non-graphical workload types (financial analysis, weather modeling, fluid dynamics, signal processing, and so on) could also be processed efficiently on GPUs when dedicated to highly parallelized task computation. Thus, general purpose GPU (GPGPU) computing was born.

GPGPU architecture involves many parallel computing pipelines, each of which can run a small program called a shader. As a customizable program, a shader can perform a wide variety of tasks. NVIDIA has capitalized deeply on this ability for more than a decade with its Compute Unified Device Architecture (CUDA) software platform. NVIDIA® CUDA® provides an application programming interface (API) for software developers to let programs written in a variety of languages access GPU functions (e.g., C, C++, Fortran, and many more via third-party wrappers).

CUDA is not the only available API for GPGPU software. For example Open Computing Language (OpenCL) is a Khronos Group open-source cross-platform alternative to CUDA. OpenCL helps programs execute on heterogeneous platforms – systems that compute across CPUs, GPUs, and other types of processors. OpenGL, also managed by Khronos, offers a cross-platform API specifically made for vector graphics and leverages the GPU to accelerate rendering. Ultimately CUDA remains popular for its easy programmability. Field-programmable gate array (FPGA) chips can be customized for accelerating certain tasks, but CUDA remains the more user-friendly and efficient approach for many programmers.
GPGPUs Today and Tomorrow

While CPU performance advances have recently rolled off, the development of GPUs continue to track more closely to Moore’s Law. As such, GPUs stand more ready than ever to help process the very large data sets commonly seen in data centers and IoT hubs. However, GPUs alone are not the answer for next-gen computation. As an alternative, heterogeneous computing looks to be the most effective approach for tackling tomorrow’s large computing needs.

With heterogeneous computing, different chip types perform complementary tasks. For example, a heterogeneous platform might have FPGAs manage information as it flows into the system. Output from the FPGAs could then stream into GPUs for processing, and CPUs would handle data management and routing. APIs help programmers devise software that lets each chip type work in concert with each other while leveraging each processor’s strengths.

General-purpose graphics processing (GPGPU) is becoming the cornerstone of digital signal processing in aerospace and defense applications like radar and sonar signal processing, image processing, hyperspectral sensor imaging, signals intelligence, electronic warfare, and persistent surveillance. This versatile GPGPU approach to maximizing data processing and throughput is applicable to a wide range of markets:

- AI and machine learning, including autonomous vehicles
- Analytics
- Big data and data mining
- Computational fluid dynamics
- Computer vision
- Database processing
- Design and visualization
- Digital warfare
- Molecular dynamics
- Preventative maintenance
- Radar processing
- Robotics
- Software-defined radio
- Video processing and analysis
- Virtual and augmented reality

To illustrate, let’s spotlight three of the above markets.

In digital warfare, attacks can take many forms: from disinformation campaigns, to espionage, to the widespread exploitation of network infrastructure, which can then be used to attack other infrastructure such as energy grids. Defending against such attacks requires a formidable amount of real-time network monitoring, event analysis, and the ability to inject countermeasures instantaneously. Some countermeasures may involve massive amounts of processing for traffic handling. Heterogeneous computing can address these various stages of digital warfare and process them with maximum efficacy.

Armed forces can reap benefits from heterogeneous computing in numerous ways. For instance, even though radar processing systems often deploy on large cruisers, submarines, and similar platforms, such vehicles must still deal with the same size, weight and power (SwaP) constraints as the rest of the military. A legacy radar processing systems might require four cubic feet to house an 18-blade server weighing over 50 kg and consuming 2000W to achieve a peak processing speed of 576 GFLOPS. Compare that with a modern VITA-75 system, such as one of ADLINK’s HPERC family of platforms. To reach a nearly identical 574 GFLOPS, ADLINK’s fanless HPERC measures only 0.8 cubic feet while weighing less than 5 kg and consuming just 200W. This is made possible in part by the onboard GPU taking over a large portion of the radar signal processing workload.
One near-future application of GPGPU technology in the military/aerospace sector will be the forthcoming F-35 Block 4 updates, expected in 2023, which must be applied simultaneously across the platform’s Autonomic Logistics Information System (ALIS) fleet management backbone. Upgrades are expected to include 11 radar and electro-optical system enhancements as well as the addition of a wide-area high-resolution synthetic aperture radar (SAR) mode to the Northrop Grumman APG-81’s active electronically scanned array (AESA) radar. Taken together, the Block 4 updates represent a significant jump in signal processing demand that will require a commensurate upgrade to the processing infrastructure. The acceleration inherent in GPU-driven parallel processing will likely be essential.

Finally, consider software-defined radio (SDR). Military threats are constantly evolving, and our ability to intercept, monitor, and decode communications must keep pace. With traditional hardware components abstracted as software functions (amplifiers, mixers, filters, etc.), SDR can address a wide variety of radio protocols and perform operations on radio signals in real-time. Naturally, such operations rely heavily on signal analysis and filtering. SDR products are widely and increasingly adopted in the military, where the ability to upgrade and update existing software to meet and beat new and emerging threats can give warfighters the combat edge. SDR is a baseline technology that enables dynamic spectrum access systems with cognitive or “smart radio” functionality. Predictably, GPUs are adept at performing this kind of software-based computation.

NVIDIA: Making a Market

How big is the future of GPUs? In looking at NVIDIA’s self-reported market for Tesla compute, the market may be as large as $30 billion by 2020 — with $4 billion in HPC, $11 billion in deep learning training, and $15 billion in deep learning inference.

To nurture this market, NVIDIA aggressively developed and promoted GPU frameworks and libraries. Such tools provide simple interfaces for accessing GPU-specific code to perform certain tasks. For example, NVIDIA frameworks for deep learning include Python-based PyTorch, MXNEt, TensorFlow, MATLAB, and NVIDIA Caffe.

One can get a sense market growth in this area from the amount of framework downloads. The following approximate download numbers lends a sense of the recent growth GPGPU development.

- 2015: 400,000 downloads
- 2016: 2 million downloads
- 2017: 8 million downloads
- 2018: 13 million downloads

To provide stability in this rapidly shifting market, NVIDIA offers five year lifecycle support for its NVIDIA® Quadro® graphics processors intended for solution developers and partners such as ADLINK. This means organizations that typically plan for solution deployments to last for many years in the field (such as the military and aerospace sectors) can look forward to five years of unchanging silicon availability and complete software support. This allows sufficient time for full R&D cycles and realization of ROI on designs.
ADLINK and NVIDIA

With many years of leadership across embedded computing markets, ADLINK deepened its long-term commitment to GPGPU platforms in 2018 by becoming an NVIDIA Quadro Embedded Partner for embedded graphics, HPC, and deep learning solutions. While ADLINK is one of 23 NVIDIA OEM Preferred Partners worldwide, only six of these (including ADLINK) have obtained NVIDIA’s permission to create chip-down Quadro GPU solutions for ruggedized embedded markets. Going further still, ADLINK is one of only ten NVIDIA Jetson Preferred Partners. This rare degree of collaboration and commitment provides ADLINK with the highest levels of technical support from NVIDIA to help meet the demanding design, production, and product lifecycle requirements of the embedded market.

ADLINK’s GPGPU products for 2019 are based on the following NVIDIA chips, all of which qualify for five year lifecycle support:

- **NVIDIA® Quadro® P1000 (Pascal™ architecture):** 1.8 TFLOPS SP Peak, 512 CUDA cores, 4GB DDR5, 96 GB/s, up to 48W GPU
- **NVIDIA® Quadro® P2000 (Pascal GP107):** 2.3 TFLOPS SP Peak, 768 CUDA cores, 4GB DDR5, 96 GB/s, up to 58W GPU only
- **NVIDIA® Quadro® P2200 (Pascal):** 3.8 TFLOPS peak FP32 performance, 1280 CUDA cores, 5GB DDR5, 200 GB/s, up to 75W GPU
- **NVIDIA® Quadro® P3000 (Pascal):** 3.9 TFLOPS peak FP32 performance, 1280 CUDA cores, 6GB DDR5, 168.2 GB/s, up to 75W GPU
- **NVIDIA® Quadro® P5000 (Pascal):** 6.4 TFLOPS peak FP32 performance, 2048 CUDA cores, 16GB DDR5, 192.2 GB/s, up to 100W GPU only
- **NVIDIA® Quadro® T1000 (Turing™ architecture):** 2.6 TFLOPS FP32 performance, 768 CUDA cores, 4GB GDDR6, 192 GB/s, up to 50W GPU
- **NVIDIA® Quadro® RTX3000 (Turing):** 6.4 TFLOPS FP32 performance, 2304 CUDA cores, 6GB GDDR6, 3336 GB/s, up to 80W GPU

ADLINK is now bringing a range of embedded products to market that will blend the advantages of NVIDIA’s graphics speed, power efficiency, and software stack with ADLINK’s proven dedication to best-of-breed embedded computing solutions for the world’s most demanding industries and applications.

AI on Modules (AloM)

In order to facilitate the development of a stronger GPGPU portfolio enabling customers to develop solutions with AI and machine learning technologies, ADLINK is expanding on the computer-on-module concept to develop a line of AI-on-Module (AloM) products to systematically address customers’ application requirements for a variety of industries. AloM products integrate GPUs, VPUs and NPUs in a variety of standard form factors such as MXM, PEG, VPX, PC/104, and CompactPCI. Being introduced later, the CMS-P1000, VPX3-P5000, VPX6H10, MXM and PEG cards are all under the AloM development initiative. Leveraging its expertise in developing highly rugged and reliable embedded modules, and modern graphics processing technologies, ADLINK’s AloM products will help customers across industries rapidly bring AI and machine learning to both existing and new system architectures, enabling them to gain competitive advantage with the capability to drive the performance required by next-generation applications in a cost-effective manner. ADLINK also plays an active role in collaboration with ecosystem partners and leading consortia to fully realize the benefits of AloM, namely its superior flexibility, modularity, scalability and affordability.

Mobile PCI Express Modules (MXM) and PCI Express Graphics (PEG) Cards

ADLINK offers a series of MXM and PEG cards as embedded graphics building blocks for integration with carrier boards and platforms featuring expansion slots. The MXM standard provides an industry-adopted socket for mobile graphics processors and is well-suited for SWaP-constrained applications such as portable ultrasound, airborne radar, and aerial infrared imaging. PEG cards connect via a commonly used interface and are easy to integrate and use in many
embedded market segments including healthcare, industrial automation, networking and telecommunication. Closely coordinated with NVIDIA’s embedded roadmap, ADLINK’s MXM and PEG products cover a wide range of GPUs spanning the Pascal through Turing architectures, providing customers with great flexibility to choose the right solution that meets the specific power, thermal and cost requirements for their application.

PC/104 Single Board Computers

In 1992, the PC/104 Consortium formed to help adapt proven desktop technologies into embedded applications. One of the group’s founders, Ampro, brought with it 16 years of PC/104 leadership and innovation when ADLINK Technology acquired the firm in 2008. Ever since, ADLINK has spearheaded efforts to keep PC/104 amply supported, relevant, and able to meet next-generation embedded needs across military and aerospace sectors.

Today, there are five major PC/104 standards, starting with the original ISA-based PCI/104 and extending to the latest PCI Express-based PCIe/104, available with various lane capabilities and configurations. The stackable, scalable nature of PC/104 allows these small boards to serve in a broad range of capacities, from cube satellite control to in-jet camera processing. ADLINK PC/104 options span from low-power, rugged designs made for legacy platform maintenance to the CMx-SLx, based on a 6th Gen Intel® Core™ Processor i3 with fully modern display support. For more forward-looking applications that can benefit from GPGPU acceleration, consider ADLINK’s PCIe/104-based CM5-P1000. The board features an NVIDIA Quadro P1000 GPU, equipped with 640 CUDA cores and backed by 4GB of GDDR5 memory. This provides for 1.8 TFLOPS performance with output across four HDMI connections.

VPX Blades

As detailed in our “Proven Precision” white paper, VPX continues to bring cutting-edge connectivity and bandwidth to small form factor solutions tailored to demanding, often rugged environments. With VPX blades, organizations can easily realize computation in a much smaller footprint than previously required by closet-sized rackmount solutions.

Bringing GPGPU performance to VPX, ADLINK’s flagship VPX3-P5000 is a 3U VPX graphics blade that utilizes a carrier board architecture to host a range of NVIDIA Quadro GPU MXM modules, including the Pascal line’s P1000, P2000, P3000, and P5000, as well as the Turing line’s T1000, RTX3000, and RTX5000. This allows ADLINK’s customers to have greater choice and lower design and implementation costs when designing powerful graphics platforms. The VPX6H10 6U VPX processor blade integrates a 12-core Intel® Xeon® Processor D-1559 and onboard NVIDIA Quadro P1000 GPU. Such features support the processing power required by graphics-intensive and mission-critical applications. ADLINK also offers a selection of other GPU-equipped VPX blades for a wide range of application demands and project criteria.

Small Form Factor (SFF) Tactical Edge Server Solution Group

The military and aerospace sectors have decades of experience and development invested in SFF computing applications. Over time, these efforts have led organizations to standardize on a few key form factors and standards that serve to bring COTS advantages to applications without sacrificing performance or reliability. ADLINK introduced the GPGPU-based High Performance Extreme Rugged Computer (HPERC) to its SFF Tactical Edge Server (SFF-TES) solution group. The goal is to better optimize its SFF portfolio to meet one of today’s most in-demand and mission-critical computing requirements: small systems that can handle both the data loads and physical pounding common in edge and fog deployments. With dimensions of only 304.8 x 150 x 130 mm, the new HPERC-KBL-MH ruggedized platform features a quad-core Intel® Xeon® processor with optional NVIDIA Quadro P1000 embedded MXM graphics via PCIe x16 Gen3 connectivity. Equipped with a fanless passive thermal solution for silent operation, ingress protection and operational reliability, the HPERC-KBL-MH complies with a range of temperature, shock, and vibration MIL-STD standards, making this compact solution suitable for deployment in the most extreme environments.

NVIDIA Jetson ARM-based Platforms

NVIDIA® Jetson™ family is a line of Tegra SoC-equipped embedded computing modules targeting deep learning applications that require low power envelopes. Such constraints often exist in network edge and fog computing environments. Applications include autonomous vehicles and drones, and robotic control.
ADLINK’s ROScube-X is a heterogeneous computing system measuring just 190 x 210 x 80 mm in its core configuration. The system serves as a robotic controller powered by NVIDIA Jetson AGX Xavier™ module. Rich in I/O ports and expansion options, the passively cooled ROScube-X features 16GB of system memory, 32GB of eMMC storage, HDMI output, and twin Gigabit Ethernet ports, making it amply capable of tackling advanced tasks with a range of Linux distributions. The system’s remarkably compact form factor makes it highly convenient for taking into the field to harness in AI inference and/or training roles, particularly when enabling automated mobile robots (AMRs).

The ADLINK M100-Nano-AINVr is another compact AI-optimized system, built around NVIDIA’s Jetson Nano processing/inference engine. The Jetson Nano contains a quad-core ARM Cortex-A57 MPCore processor and 128 NVIDIA CUDA cores. Measuring only 210 x 170 x 55 mm, the M100-Nano-AINVr hosts eight PoE-enabled Gigabit Ethernet ports and a tray-mounted 2.5” SATA drive bay, making this a highly optimized solution for multi-stream IP camera input and AI-powered NVR. Opportunities for such systems abound in applications such as defense security, persistent surveillance, and access control. This is exactly the sort of functionality organizations need when deploying camera vision at the network edge for faster time-to-decision.

For more ADLINK options based on Jetson, look for solutions such as the NEON-i1000, a rugged industrial camera for AI machine vision, or the DLAP-201-JT2, a compact inference platform for edge AI applications. ADLINK and NVIDIA will support the Jetson Nano and AGX Xavier platforms until at least January 2025.

Software Support for Optimization

ADLINK and NVIDIA collaborate to ensure that solutions such as HPERC systems are as efficient and optimized as possible. This includes assisting customers in enabling innovations such as GPUDirect RDMA. GPUDirect is a suite of evolved data transfer methods that allows Tesla and Quadro GPUs as well as third-party system devices read and write directly to CUDA host and device memory. This streamlines memory use and significantly lowers CPU overhead. GPUDirect RDMA uses Kepler-class and later GPUs and CUDA 5.0 to establish a direct data path between the GPU and third-party peer devices using standard PCI Express features. Examples of third-party devices include network interfaces, video acquisition devices, and storage adapters. ADLINK continuously develops deep expertise in both the hardware surrounding GPUDirect RDMA and the applications best poised to leverage it.

Also of note is how ADLINK’s Vortex middleware can play an instrumental role in edge computing deployments, especially when combined with GPU-equipped heterogeneous computing solutions. The Vortex data distribution service (DDS) is middleware that enables a secure publish/subscribe pattern for sending and receiving data, events, and commands among network nodes. DDS addresses the needs of applications in aerospace and defense, air-traffic control, autonomous vehicles, medical devices, robotics, power generation, simulation and testing, smart grid management, transportation systems, and other applications that require real-time data exchange.

In a distributed system, middleware is the software layer that lies between the operating system and applications. It enables the various components of a system to more easily communicate and share data. Middleware simplifies the development of distributed systems by letting software developers focus on the specific purpose of their applications rather than the mechanics of passing information between applications and systems. Such functionality becomes particularly important at the edge, where multiple AI and conventional applications may need to collaborate as part of a larger edge strategy. ADLINK has been diligent in making sure its Vortex DDS, edge systems, and NVIDIA frameworks all comprise a compatible, optimized ecosystem.

Conclusion

From cybersecurity to transportation logistics to target identification, AI is now permeating an increasing array of military, aerospace, industrial, and similar market applications. According to Analytics Insight, 51% of IT professionals report having to work with data for days to multiple weeks before it becomes actionable. A similar number said that they only “sometimes” have the resources needed to act on data. Extrapolate these limitations to military scenarios and the potential for AI to assist in decision making becomes a mission-critical advantage.

GPU-enhanced heterogeneous computing has the ability to introduce critical high-load applications such as AI to many arenas. Data centers, mobile troop encampments, robotics and autonomous vehicles can all take advantage for accelerated, low-power computing resulting in faster time-to-decision. As the NVIDIA Quadro Embedded Partner with the most market-proven experience and broadest embedded product portfolio, ADLINK continues to excel in making some of the world’s most rugged, dependable, high-performance embedded computing solutions. With NVIDIA embedded technology, ADLINK solutions are now even more capable with integrated GPGPU acceleration.

To find out more about ADLINK heterogeneous computing solutions and how they can meet your organization’s present and future needs, contact us at adlinktech.com.