WHITE PAPER

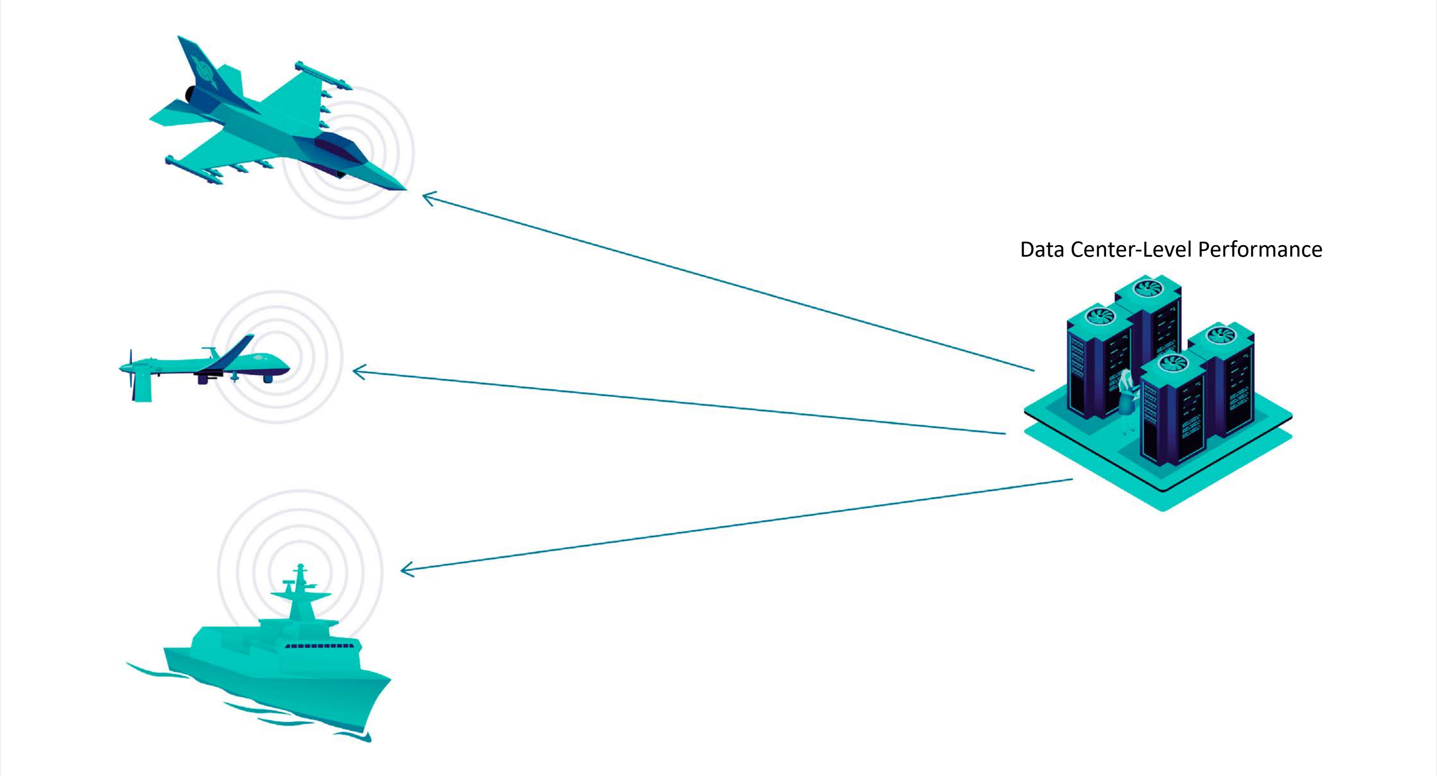
High-speed networking to unleash AI

ANTON CHUCHKOV
Edge Product Manager
Mercury Systems

Massive Disaggregated Processing for Sensors at the Edge

Edge applications need support for more powerful, deployable computing subsystems that can process extremely high bandwidth, ever-growing sensor data streams and exploit the rapidly emerging capabilities of Artificial Intelligence (AI).

This paper highlights a novel architectural approach that addresses this growing need by combining innovative data processor unit (DPU) technology with high-performance graphics processing units (GPUs) in a rugged, SWaP-optimized configuration—without the need for an x86 CPU host.



EDGE APPLICATIONS NEED FLEXIBLE, HIGH-PERFORMANCE PROCESSING

Must be able to rapidly allocate, and re-allocate, computing resources to process data steams for multiple applications

Advanced computing resources are moving from data centers to edge systems, adding efficiency and new capabilities to applications ranging from petroleum exploration to radar signal processing. These high-performance edge systems must be able to rapidly allocate, and re-allocate, parallel processing resources to handle data streams from multiple sensor sources through various types of algorithms, including deep learning/machine learning neural networks for AI.

The system is the network

Networking speeds are keeping up with the constantly expanding data streams, as communications standards like PCIe Gen 5 and 200/400+ GbE delivering huge leaps in data transfer bandwidth. Effective edge systems will exploit those leaps, recognizing that data movement and data stream processing are functions distributed throughout a network; essentially, the system is the network.

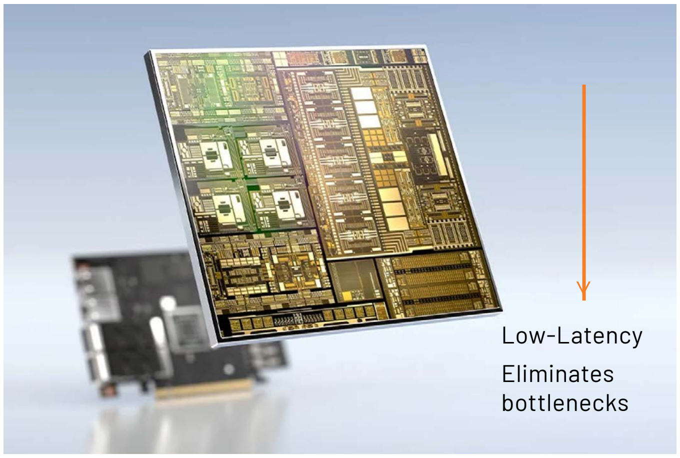


Figure 1: NVIDIA BlueField DPU

Low latency assurance is essential, at every node

As technologies evolve, data bandwidth is not the only area with significant increases; compute architecture is also benefiting from latency reductions offered by accelerators like DPUs that combine powerful CPUs, such as ARM, directly with high-speed networking and other accelerators for efficient, low latency I/O, which is critical for many edge applications, removing traditional CPU bottlenecks that may have introduced latency at each node.

New AI-based techniques, like Cognitive Radar, must operate in near-real time and will add further processing requirements

New, still-evolving application areas are adding further edge processing requirements. For example, cognitive radar applies AI techniques to extract information from a received return signal and then uses that information to improve transmit parameters such as frequency, waveform shape, and pulse repetition frequency. To be effective, the cognitive radar must execute those AI algorithms in near-real time, which, in turn, requires powerful graphics processing units (GPUs) in the processing chain..

THE DPU CONCEPT

Data centers have a similar challenge

Fortunately for edge applications, new technology has emerged to address a similar set of challenges across cloud, data center, and edge environments. Whether in the data center or at the edge, high-speed data movement is essential to application efficiency, and that movement demands a significant percentage of all computing cycles—if it is managed only by general-purpose CPUs. Sometimes the data in a stream must be processed directly by a CPU, while other streams are directed to storage. Many emerging AI applications operate using continual high-bandwidth data streams that are sent to GPUs, where immense numbers of math operations are executed in parallel. All the nodes and data streams need security. CPUs, designed for decision-making and general-purpose computing, can be programmed to manage any of those tasks, but they are not optimized for directing data streams, nor for storage and security management.

What is DPU and what it does

The data center solution for high-speed and low-latency networking is a data processing unit or DPU; it is a new class of programmable processors that is joining CPUs and GPUs as one of the three pillars of computing. Architecturally, a DPU is a system-on-a-chip (SoC) that combines three elements:

- A light-weight, multicore CPU
- A high-performance networking interface focused on parsing, processing, and moving data at line-rate speeds (i.e. 400G Ethernet)
- Programmable hardware acceleration engines for specific tasks, most especially controlling data storage, implementing security, and improving application performance for AI and machine learning

Working together, these elements allow a DPU to perform the multiple functions of offloading, accelerating, and isolating software-defined network connectivity. One function very important to edge applications is the ability to feed networked data directly to GPUs using the direct memory access (DMA), without any involvement by a system CPU. More than just a smart NIC (network interface card), DPUs can be used as standalone embedded processors that

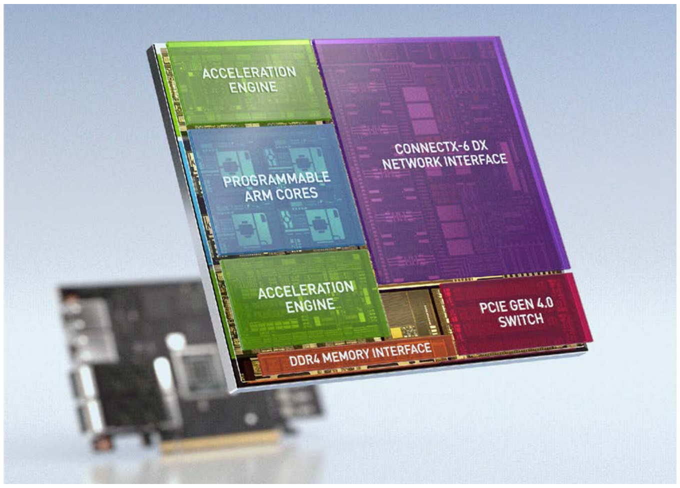


Figure 2: NVIDIA BlueField DPU Block Diagram

incorporate a PCIe switch architecture to operate as either root or endpoints for GPUs, NVMe storage, and other PCIe devices. Another critical edge function is crypto acceleration performed in line with IPSEC and TLS security protocols.

A DESIGN THAT ADAPTS DPUs AND GPUs FOR EDGE APPLICATIONS

A rugged, small form factor product combining a DPU and a GPU

Mercury's recently announced Rugged Distributed Processing (RDP) product family is a new class of processing systems for edge applications. Rugged and built to meet SWaP constraints, the RDP products feature powerful GPUs and DPUs from NVIDIA. The first member of the product family is the RDP 1U Rugged Distributed GPGPU Server, described here in some detail.

The Bluefield DPU

Central to the RDP 1U architecture is the NVIDIA® Bluefield® DPU. It is perfectly designed to support high-performance, distributed GPU edge processing, offering:

- 200 Gb/s of Ethernet connectivity to sensors, storage, and other systems
- A PCIe Gen4 switch + 16 lanes of connectivity for high-bandwidth data movement to and from GPU processing
- 8 ARM CPU cores to initiate stream processing applications and control the routing of data streams without adding latency
- Accelerated switching and packet processing (ASAP^2) engine for bolstering advanced networking
- Dedicated encryption acceleration engines, supporting security at line-rate speeds
- Storage control engines, with compression and decompression acceleration
- Path to add GPU processing to an existing high-speed compute rack without requiring the replacement of compute servers.

The A100 Tensor Core GPU

The RDP 1U edge server delivers cutting-edge computing power from NVIDIA's A100 Tensor Core GPU. Its highly parallel math operations are ideal for signal processing and AI algorithms. The A100's performance-enhancing features include:

- 6,912 processing cores that can be partitioned into seven isolated GPU instances to dynamically adjust to shifting workload demands

Figure 3: NVIDIA A100 80 GB PCIe GPU

- 80GB of GPU memory supporting 2 TB/s of memory bandwidth to enable extremely high-speed processing of huge data streams
- Native support for a range of math precisions, including double precision (FP64), single precision (FP32), half precision (FP16), and integer (INT8)

The 1U Short-depth Chasis

The physical form factor and high-speed Ethernet connectivity of the RDP chassis reflect its design goal—disaggregated parallel processing for edge applications. Some of its defining characteristics are:

- A 19" rack-mountable unit, suitable for shipboard, large aircraft, or remote ground installations
- Compact dimensions of 1U height (1.75"), 17" width, and just 20" depth
- Integrated air cooling
- Tested for 0°C to 35°C operation
- 16 lanes of internal PCIe Gen4 communications, linking the DPU and GPU
- Two 100 Gbps (or single 200G) fiber-optic Ethernet network ports delivering data streams to and from the DPU
- A 1 Gbps copper Ethernet network port for system control communications

UNIQUE CAPABILITIES THAT EMPOWER ADVANCED APPLICATIONS

It makes powerful GPU processing profoundly accessible for edge applications

The 1U RDP GPU server allows Mercury customers to deploy ruggedized, distributed GPU processing at the edge with data center-class server technology. NVIDIA DPU and GPU modules are packaged to reduce SWaP, cost, and complexity for networked GPU servers. Customers can now scale aggregate GPU resources by adding standalone GPU servers to an existing high-speed compute rack without requiring the replacement of compute servers.

Further, instead of specifying and restricting GPU resources within an x86 host system and then deploying a mix of compute and inferencing-optimized machines, RDP GPU node resources are available to functions across the network; only an Ethernet connection is required. Performance accelerations range up to 249x (compared to CPU only) for AI inferencing applications such as BERT-LARGE Inference workload, performed using an A100 80GB.

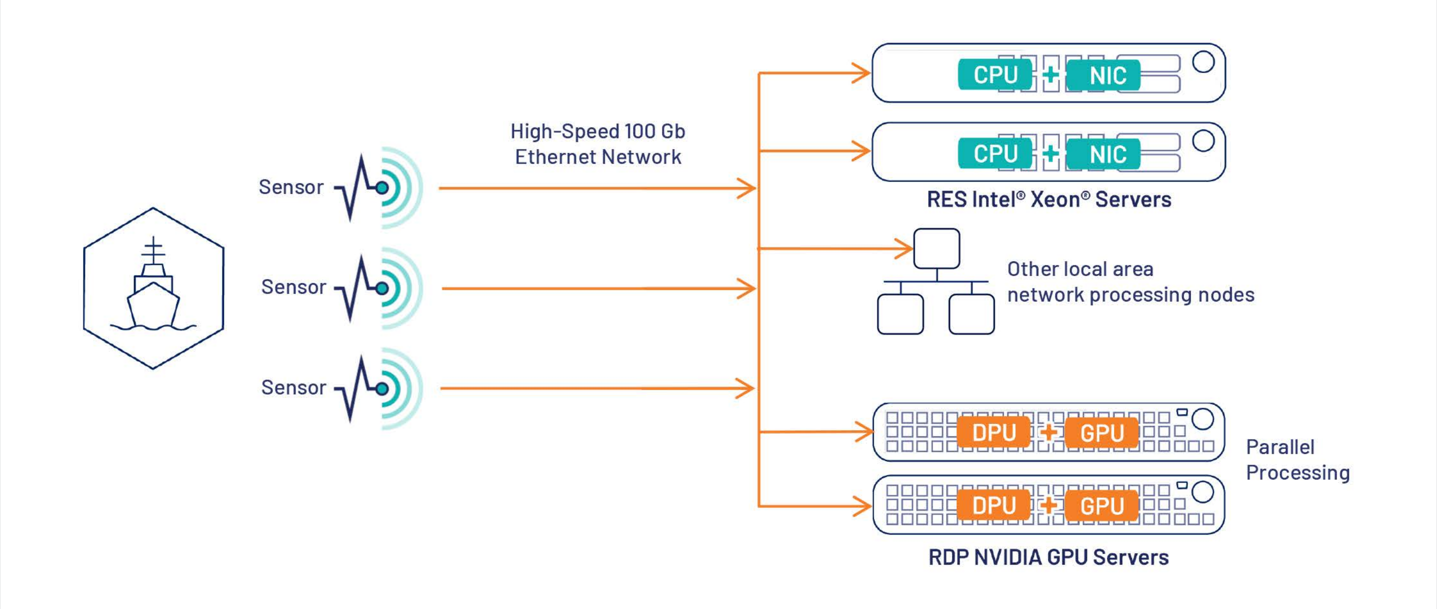


Figure 4: High-Performance Distributed GPU Edge Processing

The ARM processors in the DPU allow designs to avoid the cost and power requirements of an intel CPU

Avoid x86 Host CPU Cost and Power Requirements

The ARM cores integrated within the BlueField DPU mean system configurations do not need to include an x86 CPU, reducing costs, SWaP and power requirements. Running the Linux operating system, these ARM cores are easily capable of initiating decoupled data-streaming applications that execute on the GPU. Those applications could be signal processing of input from a radar antenna or deep learning algorithms used to extract insights from sensor stream input. An additional general-purpose CPU is not needed to manage the GPU applications or data flow. And since an x86 host server is not needed for management, x86 servers no longer need to be over-specified to handle such tasks, which translates to cost and power consumption savings within the system architecture.

The DPU's data management functions move data flexibly into, and out of, the GPU for efficient, low-latency processing

Data Management that Exploits Direct Memory Access

The DPU's data management functions, combined with 100/200 GbE and PCIe Gen4 interfaces, deliver highly efficient data-stream transfers. Using DMA, these data streams can move flexibly into and out of the GPU for uncompromised, low-latency processing.

The flexible assignment of data streams can be extended to other GPUs in other RDPs operating on the same rack, or to general-purpose servers that don't have on-board parallel processing, for efficient support of complex and demanding application environments. Networking speeds are maintained from end to end, without signal slowdown through the RDP nodes.

Accelerated encryption - in silicon - enables advanced security for data streaming applications

Low Latency Security for Data Streaming Applications

Security engines within the DPU execute IPSEC and TLS algorithms in silicon, supporting low-latency security for data-streaming applications; one key function is deep packet inspection at line speed. This capability also simplifies low-latency support for multi-level security implementations controlling access to data. At the network level, the advanced security inherent to every RDP server means each node is protected, a concept referred to within data centers as micro-segmentation.

Similar silicon-based support for data storage

High Speed, Secure Data Storage and Retrieval

While the DPU integrates on-board storage to host the operating system, additional storage functions can be managed by specialized hardware engines integrated inside a DPU, if connected to an NVMe fabric. These engines, operating in parallel with GPU processing, provide high-speed data encryption as well as compression and decompression acceleration.

Software frameworks for developers to accelerate applications

Software Frameworks to Accelerate Application Development

A key added value of NVIDIA technology is in the software frameworks that simplify and accelerate development of high-performance applications.

CUDA® is a parallel computing platform and programming model developed by NVIDIA to help developers speed up math-intensive application algorithms by harnessing the parallel processing power of GPUs. CUDA is supported by hundreds of libraries, software development kits (SDKs) and optimization tools, many focused on specific types of application areas, including deep learning, natural language processing and other branches of AI.

An important CUDA characteristic is consistent support for new generations of GPUs. Customers can migrate CUDA-based applications to new, more powerful GPUs in a very straightforward fashion, preserving their software investment and quickly accessing the performance advantages of faster hardware.

NVIDIA DOCA™ is an analogous software framework for DPUs. DOCA unlocks innovation in the data center, cloud and at the edge by enabling developers to rapidly create applications and services on top of BlueField DPUs. It consists of an SDK and a runtime environment. The DOCA SDK provides industry-standard open APIs and frameworks, including the Data Plane Development Kit (DPDK) for networking and security and the Storage Performance Development Kit (SPDK) for storage. The DOCA runtime environment includes tools for provisioning, deploying and orchestrating containerized services on hundreds or thousands of DPUs across the data center.

As with CUDA, DOCA provides consistent support for new generations of DPUs, simplifying and accelerating software migrations.

Aerial is an application framework for building high-performance, software-defined, cloud-native 5G applications to address increasing consumer demand. Customers leverage the NVIDIA Aerial™ SDK to build and deploy GPU-accelerated 5G virtual radio access networks (vRAN) that optimize results with parallel processing on the GPU for baseband signals and data flow.

BRINGING HIGH PERFORMANCE, HIGH BANDWIDTH COMPUTING TO THE EDGE

Mercury's RDP product family allows customers to access powerful, rugged, SWaP-optimized computing for edge applications, exploiting the parallelism of GPU technology without the overhead of a traditional x86 processor. The RDP's integrated DPU enables efficient, low latency processing of extremely high bandwidth data streams without the cost or power impact of a general-purpose CPU. A compact, rack-mount chassis design simplifies deployment across a range of edge environments.

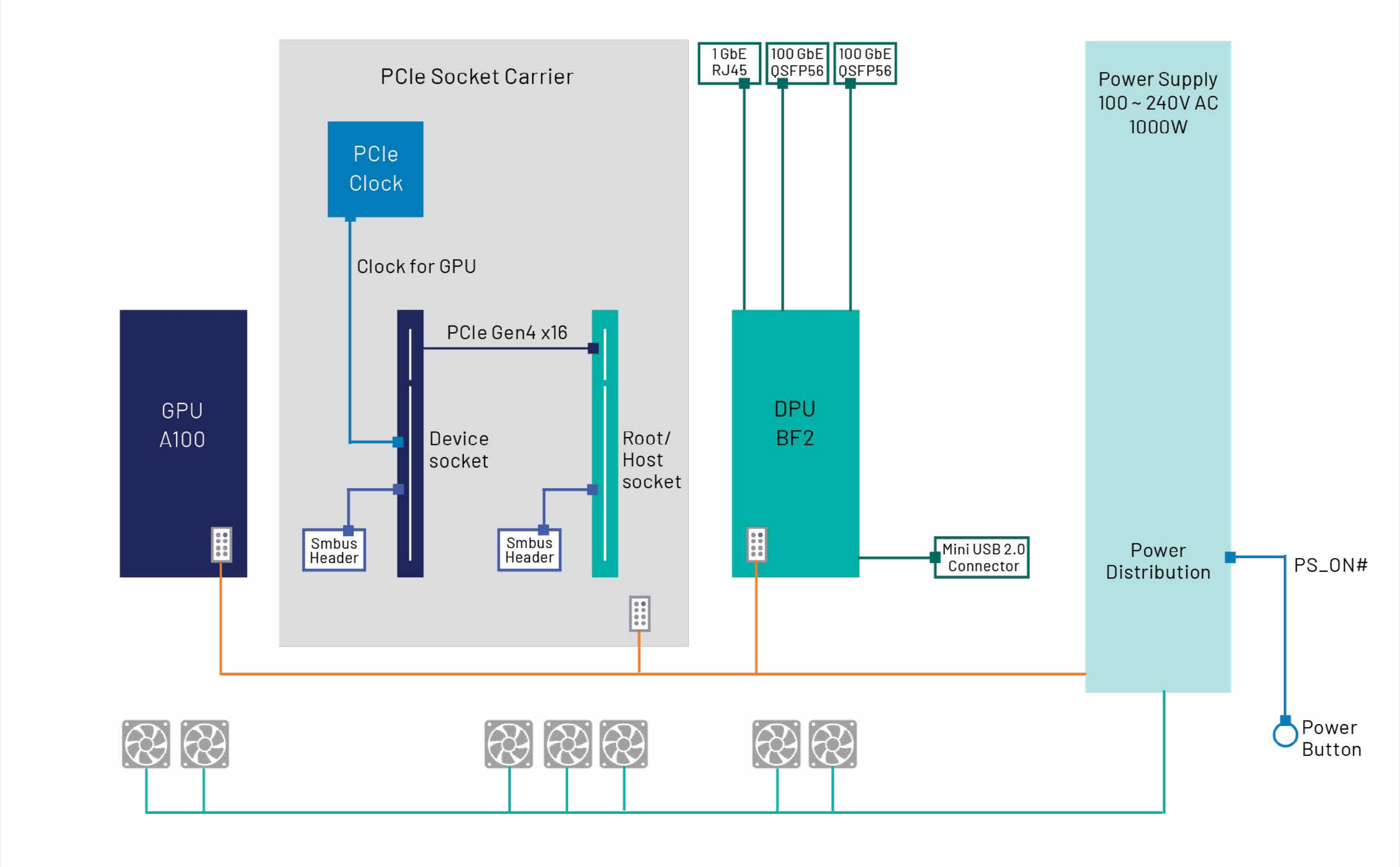


Figure 5: RDPB02-1U20RA Functional Block Diagram

ENGAGE WITH US

At Mercury, we know that all our technology innovations are complemented and enhanced by customer partnerships, collaborating to solve problems. We see close cooperation with technology visionaries, program managers, and engineering teams as key to getting the maximum value from our new RDP solutions.

Engage with our team to explore how they can meet your most demanding requirements for edge applications and help move your programs forward

FIRST IN A FAMILY OF RDP SOLUTIONS

Mercury's 1U RDP server is the first in a new rugged distributed processing rackmount server series. This product family will exploit the processing capabilities available when DPUs are combined with GPUs in a tightly integrated design. As an NVIDIA Preferred OEM partner, Mercury intends to continue leveraging roadmapped GPU and DPU innovations to solve challenging customer problems at the edge.

ADAPTING LEADING EDGE COMMERCIAL TECHNOLOGY

The commercial electronics market uses billions of dollars to drive technology

The enormous worldwide commercial electronics market continually drives technology forward at a rapid pace with hundreds of billions in R&D investments every year.

The much-rugged marketplace exploits that huge commercial investment

The much smaller rugged solution marketplace can exploit that huge commercial investment by adopting new technologies and then adapting them to the unique requirements posed by difficult environments on the edge

The challenge is to compress the adopt and adapt cycle as much as possible

The challenge is to compress the adopt-and-adapt cycle as much as possible. If edge applications can rapidly access new technologies, they can expand their functions and offer users new capabilities.

Mercury meets the need for accessible leading-edge technology

Mercury is the leader in adapting commercial technology for applications in difficult environments, making them more affordable, safe and secure. We accelerate the process of moving new technology beyond the data center, as we bridge the gap between commercial technology and edge applications

Through close collaboration with our technology partner NVIDIA, we have been able to adopt the DPU concept and rapidly adapt it for new types of edge applications. In close collaboration with the semiconductor industry, we will continue to integrate latest-generation technologies to enable server-class and HPC processing at the edge

Corporate Headquarters

50 Minuteman Road
Andover, MA 01810 USA
+1 978.967.1401 tel
+1 866.627.6951 tel
+1 978.256.3599 fax

International Headquarters
Mercury International

Avenue Eugène-Lance, 38
PO Box 584
CH-1212 Grand-Lancy 1
Geneva, Switzerland
+41 22 884 51 00 tel

Learn more

Visit: mrcy.com/servers