High-Performance Edge Computing for Real Time Decision Making

ONE STOP SYSTEMS

(877) 438-2724

onestopsystems.com

Increasingly businesses across multiple industries will embrace high performance edge computing to fulfill the demand for processing locally sourced digital data to drive precise real time decision making.

Edge computing enhanced with HPC capabilities, such as deploying rugged *AI on the Fly*® systems at the edge from One Stop Systems (onestopsystems.com) powered by NVIDIA V100S GPUs, will let many companies overcome the performance, latency, and security challenges that occur when remote locations rely on traditional data center or cloud computing models. In this article PureB2B explores the underlying market trends and unique requirements driven by this imperative to create actionable intelligence at the edge.

Tim Miller, VP of Product Management at One Stop Systems
TMiller@onestopsystems.com

High Performance Edge Computing for Real-Time Decision Making

By John Palmer, PureB2B

The time is now to embrace high-performance computing technologies and advanced artificial intelligence capabilities for your upcoming edge computing deployments.

In today's digital economy, data is exploding. Digital sensors of every variety generate vast amounts of raw data. Data scientists and corporate managers across different industries must now harness these raw data streams and convert them to actionable intelligence for real-time decision making. Until recently, the ability to deal with all but the simplest data sets involved culling the data and shipping it to centralized data centers for offline processing. Complex data sets were abandoned as simply too hard to exploit.

Today, the emergence of high performance computing (HPC) and artificial intelligence acceleration hardware deployment directly at the edge, close to the point of data generation, opens up AI advancements across industries previously unable to benefit from centralized datacenters. Edge computing enhanced with HPC capabilities will let many companies overcome the performance, latency, and security challenges that occur when remote locations rely on traditional data center or cloud computing models. Increasingly, businesses across multiple industries embrace high performance edge computing to fulfill the demand for processing digital data in disparate remote locations.

As a result, high-performance Internet of Things (IoT), AI, machine learning and deep learning—and other emerging technologies—require the powerful computing typically housed in data centers or the cloud to be brought to the edge.

Why High Performance Edge Computing Now

Exploiting AI training and inference capabilities demands HPC, which aggregates processing power, storage, and more to complete complex computational tasks involving volumes of data with intense speed, security, and accuracy. The demand for this capability is growing. Hyperion Research expects the market for HPC server-based AI, including machine learning and deep learning, to grow at a rate of 26.3% CAGR to top $1.6 billion in 2022. This market will be driven largely by investments in verticals such as manufacturing, energy and utilities, government and defense, and other industries with a need to resolve complex problems and calculations quickly.

"HPC has already moved to the forefront of R&D for AI and deep learning, including leading-edge developments for autonomous vehicles, precision medicine, smart cities, and the Internet of Things," according to the Hyperion Research paper The Business Value of Leading-Edge High Performance Computing: 2019 Update.

"These economically important new use cases are natural extensions of activities HPC has supported for some time—vehicle design, health care, urban transportation, regional/national power grids, and large distributed networks."

HPC server-based AI is expected to grow by
26%

Industry watchers only expect this trend to grow. A November 2019 report from Gartner says that by 2022, more than half of enterprise-generated data will be created and processed by technology outside the data center or cloud. And according to a Forrester Research blog, "the biggest benefits organizations seek from edge computing include flexibility to handle present and future artificial intelligence demands and the fact that computing at the edge avoids network latency and allows faster responses."

Industries Tackling HPC at the Edge

Some industries take the straightforward approach to edge computing and install computer equipment in climate-controlled warehouses as close as possible to the point of capture, to avoid latency issues and security vulnerabilities that can occur when data is streamed to faraway data centers. But for other industries with unique use cases, it is critical to couple the needed computing processing power in the appropriate package to handle the business and technology challenges that remote locations present. As more businesses are expected to speed the delivery of more intelligent applications, the need for robust and custom-sized computer systems that can handle those needs will also grow. A 2018 report from McKinsey and Company identified more than 100 edge computing use cases across 11 different industries that would deploy by 2025, representing more than $200 billion of hardware in use.

The use cases are diverse and many. AI systems deployed at the edge can remotely monitor mine and oil rig operations, control air quality and congestion lane monitoring in smart cities, and track unmanned vehicles operating autonomously on land, sea, and air. The automotive industry, as another example, requires computing systems that can reside in the trunk of an autonomous vehicle. In this scenario, split-second local processing enables self-driving cars to take action to prevent an accident.

The military uses geospatial visualization to create real-time maps of the modern battlefield that can keep track of every aircraft, troop movement, vehicle, and enemy combatant within a given area. Powerful AI systems reside on ships, airplanes, and tanks with the all-important task of sifting through massive amounts of sensor data on possible targets to decide who is friend or foe. These systems must also withstand extremes of weather, altitude, temperature changes, shock, and vibration.

Entertainment audiences constantly demand more content and better graphics. High-performance systems allow artists to interact with their audience or power immersive experiences in location-based entertainment venues. Heavy-duty video servers help provide high-performance hardware and software on the entertainment stage, helping performing artists bring striking videos and graphics to their live shows with massive amounts of video effects. Upon the completion of each show, the entire set of servers must be packed and moved to the next venue, so the development of smaller server packages that can deliver the needed HPC requirements can eliminate bulky equipment.

Industries such as healthcare, where HPC operates in the cramped spaces of diagnostic machines while maintaining regulatory compliance, must consider compact sizes with secure features. For instance, HPC powers CT scan machines, MRIs, and genome mapping applications that help speed the development of medicine. These machines grow smaller with each iteration, but they must still adhere to many compliance requirements when dealing with patient information. Strict requirements for noise and vibration make it difficult to build fans into the design of these machines, and therefore the build requires customization and miniaturization as the machines get smaller. The industry use cases for HPC in edge environments will only continue to multiply.

Ruggedized, Ready for Anything: Requirements at the Edge

When HPC resides in less forgiving locations—depending on the industry—lacking pristine computing environments new hardware design is required that can take a beating while continuing to deliver fast, reliable performance in any condition. Many specialized use cases require HPC edge capabilities that must be coupled with specialized, ruggedized form factors that can withstand several unique physical conditions such as extreme heat or cold while also fitting into small spaces. For companies looking to implement HPC at the edge, there are many considerations, especially since they will entrust their AI capability, information, and ultimately, business success to specialized machines meeting edge requirements.

In the HPC tech field, the acronym SWaP (Size, Weight, and Power), refers to the concept of optimizing the overall dimensions, power, and weight of a device while increasing its efficiency and adhering to the overall footprint available.

The concept is applied to the design of military applications, but it can also be used when considering HPC for applications as well.

Edge systems need to be designed with the future in mind, as edge technology constantly changes and becomes more advanced. When looking for an HPC system specialized for the edge, consider the

- **Cooling** — Rugged computing systems often require designs without fans to cool interior components, help maintain reliability in extreme temperatures, or where factors such as vibration, dust or debris, or varying power voltage could cause failure in typical systems. Conduction cooling, cold plates, and liquid cooling provide viable alternatives to HPC systems at the edge.
- **Connectivity** — Look for systems that feature a variety of input and output ports to allow expandability and connection with existing legacy systems as well as paths to advanced networks such as 5G wireless and 100Gb or more HPC interconnects.
- **Smaller scale deployment** — Systems designed to operate in the field, in vehicles, or in medical equipment must fit in a small enough package that does not interfere with the design, but must also deliver the powerful processing needed. A reliable and scalable system can accept performance accelerators in the form of more powerful semiconductors (CPU, GPU, VPU).
- **Expandability** — Edge devices must be ready to work with the upcoming release of the fourth-generation expansion of the PCI Express (PCIe) standard, a bus standard that computer hardware uses to transfer data internally. Essentially, PCIe provides the “data highway” inside systems that connect elements such as data acquisition, storage, and compute engines. The new generation, PCIe Gen 4, will double the standard transfer rate from 8.0 GT/s to 16.0 GT/s.
- **Unique form factors and size constraints** — AI devices in many industries have become smaller. Unique designs require that the computing package must be customizable and be able to fit into any space, whether it’s the trunk of a car, in a medical device, or onboard a drone aircraft.
- **Environmental tolerance** — HPC edge systems must operate reliably in any type of weather or environmental condition including extreme changes in temperatures. Look for a build that can withstand the shock and vibration of extreme operations. If dust or dirt are a common element (such as in mining operations), the design must ensure the components are protected from those conditions.
- **Rugged, one-piece design** — A reliable HPC edge system is built to last, with as few parts as possible to eliminate moving parts and decrease the chance of failure. In many cases, data capture, compute accelerators, and flash storage capacity can be integrated in one package.
- **Power considerations** — HPC edge systems should operate independently, in environments where power sources may be sporadic or subject to surges of varying voltages. The system’s electrical components should maintain reliability and potentially switch to battery power if needed to handle changing conditions.

Partnering for Computing on the Edge

HPC at the edge brings more powerful and robust computing to remote locations, enabling more powerful and reliable AI applications to operate using small, rugged packages that pack a punch. It is critical to work with the right provider for applications that need to pack that power and dependability into specialized packages that can survive and perform anywhere on and off the planet. Consider the following factors when looking for a high-performance edge computing provider:

- A reliable provider thinks of the future, thereby designing systems that have forward-thinking features. For instance, features such as self-encrypting NVMe solid state hard drives native to PCIe, which makes data security and scale-out capacity easier and leads to higher bandwidth with lower latency.

- An experienced provider has a proven track record of working with OEMs and end users to develop customized 'perfect fit' solutions to meet specialized requirements cost effectively.

- A preferred provider should have strong partnerships with OEMs and tech vendors that develop the fundamental elements of high-performance edge systems, including GPUs, FPGAs, CPUs, network interfaces, NVMe storage, etc. Partnerships are crucial, as these relationships are required to remain on the competitive leading edge.

- Providers should have a long history of designing hardware systems that cater to the needs of applications that require robust and rugged builds, while providing custom connectivity and scaling, such as adding more graphics processing units accelerators in conjunction with more storage.

- Computing needs vary by industry, and the right provider must design a system that takes into consideration the special regulations and certifications that must be followed. For instance, the military has strict Federal Information Processing Standards (FIPS) security and reliability standards, while HIPAA regulations in the medical field are in place to protect sensitive patient information.

- Providers will need to bring to the table ruggedization and specialized form factors expertise to design for conditions such as extreme shock, temperature, and vibration. The provider must have the experience to anticipate everything that can happen in the field, and as a result will work with companies to develop the appropriate product that can handle anything.

- To stay competitive in a fast-changing tech world, companies will need providers that have deep experience in latest generations of PCIe and high-speed signaling design.

HPC deployed at the edge enables businesses across industries to more quickly process vast volumes of data, power more intelligent applications and support real-time decision making. Selecting the right specialized HPC edge provider will enable companies to drive innovation at the accelerated pace today's digital economy demands.