Software-Defined Hardware Provides the Key to High-Performance Data Acceleration

WP019

2023.05.22

Copyright and Trademarks

Notice of Disclaimer

Achronix®

Data Acceleration

Achronix Semiconductor Corporation

2903 Bunker Hill Lane
Santa Clara, CA 95054
USA

Website: www.achronix.com
E-mail : info@achronix.com

Executive Summary

Across a wide range of industries, data acceleration is the key to building efficient, smart systems. Traditional general-purpose processors are falling short in their ability to support the performance and latency constraints that users have. A number of accelerator technologies have appeared to fill the gap that are based on custom silicon, graphics processors or dynamically reconfigurable hardware, but the key to their success is their ability to integrate into an environment where high throughput, low latency and ease of development are paramount requirements. A board-level platform developed jointly by Achronix and BittWare has been optimized for these applications, providing developers with a rapid path to deployment for high-throughput data acceleration.

A Growing Demand for Distributed Acceleration

There is a massive thirst for performance to power a diverse range of applications in both cloud and edge computing. To satisfy this demand, operators of data centers, network hubs and edge-computing sites are turning to the technology of customized accelerators.

Accelerators are a practical response to the challenges faced by users with a need for high-performance computing platforms who can no longer count on traditional general-purpose CPUs, such as those in the Intel Xeon family, to support the growth in demand for data throughput. The core of the problem with the general-purpose CPU is that Moore's Law continues to double the number of available transistors per square millimeter approximately every two years but no longer allows for growth in clock speeds. In addition, parallelism within CPUs quickly reached a practical limit. As a result, other technologies are better placed to support new types of workloads such as machine learning, genomics research, mathematical and statistical analysis, speech and image recognition, and data mining and search.

In contrast to traditional database-driven applications, these new workloads often do not map well onto conventional CPU pipelines. Some, such as the training of neural networks have been shown to work well on GPUs. These algorithms can take advantage of the hundreds of parallel floating-point shader cores to iterate through the trillions of steps needed to update a large network. Genomics and data searches, on the other hand, make use of large numbers of comparison steps and work with lower-resolution integer data. Though these workloads could take advantage of either CPU or GPU processing, the computational and energy efficiency of these tasks is comparatively low when run on such platforms. Custom ASIC- or FPGA-based accelerators can offer much greater throughput at lower power because they enable designers to construct dedicated circuits optimized for these operations and data types.

Hyperscale data-center operators in sectors such as Internet search and social media have seized on the accelerator concept to support the workloads that underpin their services. Voice-response systems are now in everyday use and supported by artificial-intelligence algorithms running on combinations of conventional server blades with custom accelerators. As the need for applications based on technologies such as machine learning and data mining expand, a wide variety of enterprise users are turning to accelerators to let them keep pace with demand.

In addition to this growth, the use of accelerators is set to expand beyond the data center. Sectors such as virtual reality, autonomous driving, robotics, and Industry 4.0 cannot tolerate the telecommunication delays of relaying information from distant data centers. Increasingly, compute horsepower will need to be deployed in edge-computing racks mounted in roadside cabinets, next to mobile base stations or within on-campus closets.

Across the data-center and edge-computing use cases, there are a number of common drivers. Energy efficiency is a core requirement to reduce the cost and complexity of cooling as well as to minimize electricity bills. Low-power

operation is critical in edge-computing installations where there is less control over ambient temperatures and where maintenance needs to be kept to a minimum.

In many sectors, rapid change is inevitable, which leads to the need to be able to tune and rework applications to suit requirements as they arise. It is not just updates to existing applications. Often, new use cases arise that can challenge the ability of users to react in a timely fashion. These use cases may call for the development of hybrid applications that bring together different technologies and concepts, such as adding AI functionality to mathematical modeling and data-mining systems. To cope with these changes, users need to be able to call on accelerator technologies that can work well together, with the individual components communicating at high speed across network connections.

Scalability is similarly important. As the customer base for a particular service grows, operators need to know they can add capacity easily. Again, highly programmable solutions with efficient communications to support scaling through increased parallelism are vital. Support for protocols such as 100 Gbps Ethernet and faster links ensures that the growth can be accommodated using distributed processing. For example, an edge application may call on cloud support until the local cabinet is upgraded with additional processing power.

Hardware Platforms for Acceleration

Accelerator hardware can take many forms. The ideal configuration is one that offers the combination of PCI Express (PCIe) and high-speed Ethernet connectivity, with the option to add custom connections to support a variety of topologies such as ring, mesh, and daisy chain structures to match the application's data-throughput needs. PCIe support allows tight integration of the acceleration engines with host processors and other accelerators through memory-mapped interfaces. The ability to exchange data through shared-memory structures over an interface such as PCIe can greatly ease the development of distributed applications.

Ethernet connections running at, and above, 100 Gbps provide further scope for expansion. With their own Ethernet ports, instead of having packets routed through the host's main network interface, accelerators can co-ordinate with each other efficiently. For example, in a distributed storage configuration, accelerator cards can be connected directly to embedded non-volatile memory (NVMe) modules where the independent search engines in each can easily coordinate, using messages sent over their Ethernet connections to identify data scattered across many nodes.

Whether used as the primary acceleration technology or in concert with GPUs and other technologies, FPGAs are highly suited to the needs of data-center and edge-computing applications. A key advantage of FPGAs is that they can be programmed in-system to create a huge variety of digital hardware circuits. Software can select the programming bitstream for the target application and send it to configure the FPGA. By loading new patterns into the logic array on the device, FPGAs can be updated dynamically to take on new tasks as required. The programmability creates hardware defined by software, fully supporting the ability of users to dynamically change not just applications, but the hardware on which they run. Hardware programmability, combined with the ability to link multiple accelerators, delivers an enormous degree of flexibility to the user.

A number of computing users have realized the power of FPGAs in acceleration applications. For example, Microsoft's Catapult project uses FPGAs to build accelerators for its search services and in its Project BrainWave for high-speed AI inferencing. Amazon offers the ability to use FPGAs in the cloud through its F1 services, making it easy to deploy the technology to remote users.

Other sectors have chosen to use FPGA acceleration for some time. For example, arrays of FPGAs have been employed for many years to handle the processing of radar in military and aerospace applications and real-time

imaging in medicine. As the industrial sector adopts concepts such as real-time machine health monitoring as part of the move towards Industry 4.0, users can turn to FPGAs to improve the quality and responsiveness of their algorithms.

Although it is possible to use GPUs for data acceleration, FPGA implementations often benefit from lower latency and greater energy efficiency. A key issue with GPUs is that their compute efficiency is often a fraction of their theoretical throughput. Because they are optimized for 3D graphics rendering pipelines, the shader cores tend to operate out of relatively small local memories because they benefit from execution pipelines that have a high degree of data reuse. Streaming workloads exhibit fewer opportunities for data reuse, which means the memories need to be filled with new data more frequently, which impacts processing time. The cache-oriented subsystems of CPUs suffer from similar issues. Because they can implement complete pipelines through which data flows freely, FPGAs offer much greater levels of compute efficiency. For example, benchmarks for genomics applications demonstrate a speedup of 80-fold on FPGA-based hardware compared to CPU-based implementations.

Within high-performance compute and cloud-computing environments, architects are turning to FPGA acceleration to avoid bottlenecks appearing in other parts of the system. By devolving more work to the storage subsystems themselves, data-center users can achieve large improvements in efficiency. Database acceleration, data analytics and other forms of processing suitable for computational storage can be deployed on accelerators alongside functions for low-level services such as encryption, deduplication and secure erasure coding.

As concepts such as software-defined networking (SDN) and network function virtualization (NFV) take hold, server blades are being used to take on a more critical role in the management of communications within and between data centers. However, the processing burden on Xeon-class server processors as line speeds increase to, and beyond, 100 Gbps is immense, and data-center operators are keen to offload the handling of many of the SDN functions to nearby accelerator cards. In emerging architectures, the general-purpose server CPUs are used to handle exceptional events while the accelerators take care of the bulk of networking traffic. FPGAs provide the ability to update algorithms and protocol handlers as new requirements, applications and security threats arise, making them excellent substrates for networking acceleration.

Implementing Effective Acceleration

The first wave of accelerators employed by hyperscale users such as Amazon, Facebook, and Microsoft were largely custom designs. These companies were able to justify the economies of scale necessary to build their own custom boards, whether based on ASICs of their own design or off-the-shelf FPGAs and GPUs. From a cost and time perspective, such custom chip-level designs are difficult for enterprise data-center and edge-computing users to justify. However, designing custom ASICs and boards is not essential. The need for standard interfaces such as Ethernet and PCIe make it not only possible to use standard board-level products, but also desirable.

As a long-standing supplier of acceleration hardware, BittWare has gained deep experience in designing FPGA-based cards around the PCIe form factor for customers across a broad range of sectors from high-performance computing through cloud acceleration to instrumentation. Now part of Molex family of companies, BittWare is able to call on global supply networks and deep relationships with server suppliers such as Dell and HP Enterprise. BittWare is the only FPGA-vendor agnostic supplier of critical mass able to address the qualification, validation, lifecycle and support requirements of enterprise customers who want to deploy FPGA accelerators in high volumes for mission-critical applications.

A key differentiator for BittWare in these applications is the extensive software support the company provides for its FPGA-based accelerators. Each card is supplied with software drivers for Linux and Windows that enable rapid

integration into a wide range of systems with PCIe and Ethernet connectivity. In addition to supporting communications between a host CPU and the card, the drivers provide access to embedded firmware on the card. This firmware handles a variety of management and self-test functions. They enable the FPGA circuitry to be reconfigured with new images as needed in addition to a number of monitoring programs for power, voltage and temperature. If the cooling in the host system fails, the firmware can perform a managed shutdown of the card to avoid thermal overloads. In addition, the software bundle includes reference designs so that developers can quickly build configurations that let them test card functionality and start work on their own applications.

For the latest generation of accelerator cards, BittWare has worked closely with Achronix, the only FPGA vendor able to provide both standalone FPGAs and embedded FPGA (eFPGA) IP. The VectorPath® S7t-VG6 card (**Figure 1**) uses Achronix's 7nm Speedster®7t FPGA, which combines a number of features that provide not just high-throughput data acceleration internally but also supports the highly distributed, networked architectures that are now required from machine learning to advanced instrumentation.

Figure 1 – ***VectorPath S7t-VG6 Accelerator Card***

Software-Friendly Hardware Provides Maximum Flexibility

Providing direct support for distributed architectures, the Speedster7t FPGAs used in the VectorPath S7t-VG6 mark a significant shift from traditional FPGA architectures by making it easier for software-oriented developers to build custom processing units. This innovative new architecture is in contrast to conventional FPGAs made by suppliers such as Intel and Xilinx, which were not designed with a focus on data acceleration.

In designing the Speedster7t architecture, Achronix created an FPGA that maximizes system throughput while improving ease of use for computer architects and developers. In a key departure from traditional FPGA architec-

tures, Speedster7t FPGAs include an innovative, 2D network-on-chip (NoC) that streams data between processing elements within the logic array and the various on-chip high-speed interfaces and memory ports.

Conventional FPGAs require the user to design the circuitry to connect their accelerators to high-speed Ethernet or PCIe data ports and/or memory ports. Typically, a single system consists of multiple accelerators connected to multiple high-speed ports. For example, (**Figure 2**) illustrates a scenario where there are two accelerators connected to two memory ports for a shared memory space. In this scenario, the clock-domain-crossing (CDC) between the memory and FPGA clocks is managed using FIFOs. Additionally, a switch function is needed in the FPGA fabric to manage the addressing, arbitration and back pressure. In conventional FPGAs, this functionality consumes a significant amount of the FPGA resources and are complex enough to degrade the system performance as well as complicate timing closure.

Achronix took the approach of enabling software-designed hardware in which Ethernet and other high-speed I/O ports can be easily connected to custom accelerator functions using the 2D NoC (**Figure 3**). The Speedster7t NoC eliminates the need for designing the CDC and switch functions that connect the accelerators to the high-speed data or memory ports. By simply connecting functions to the NoC, the connectivity challenges are eliminated,

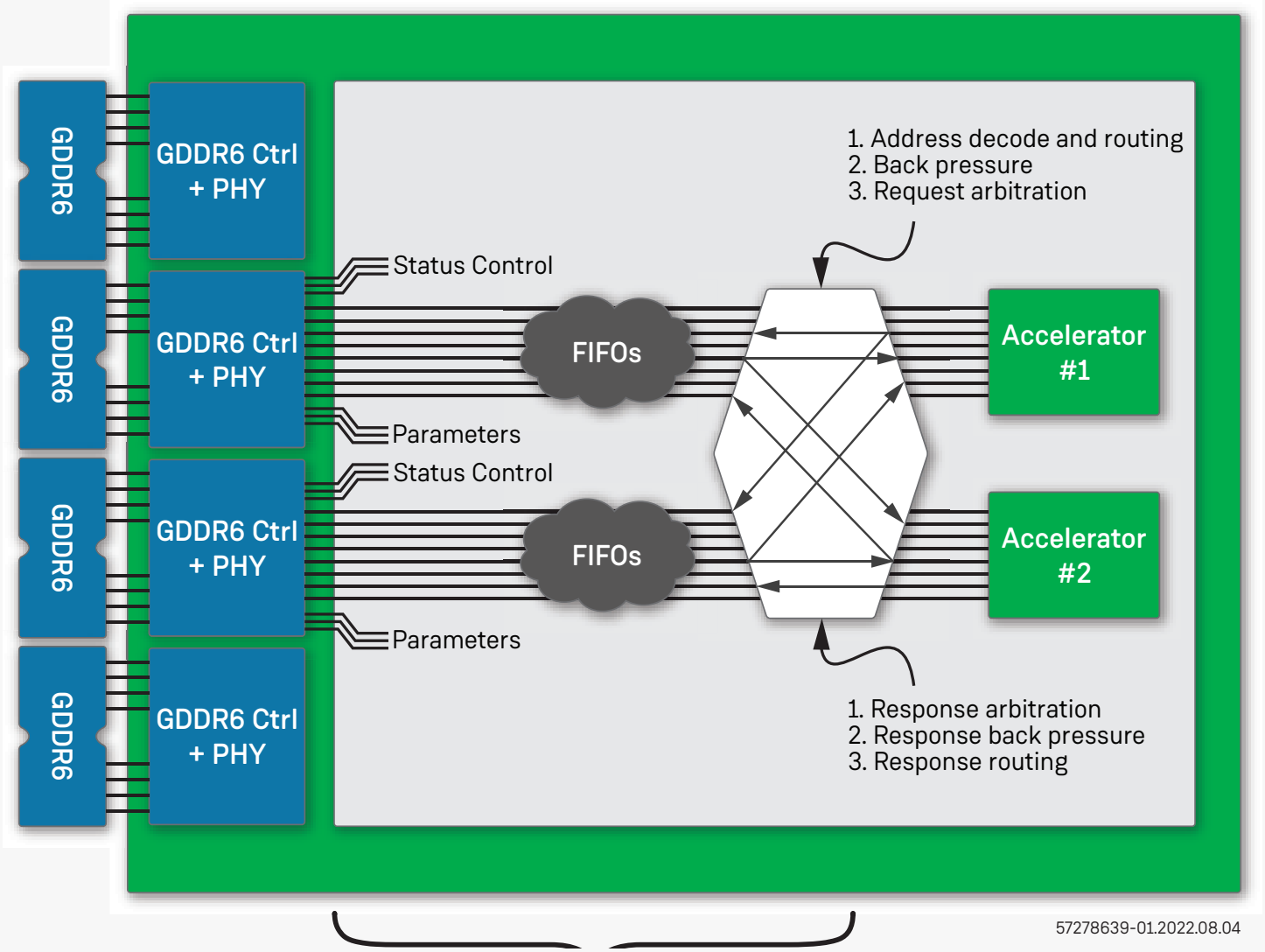

Figure 2 – *Challenges with Traditional FPGA Design*

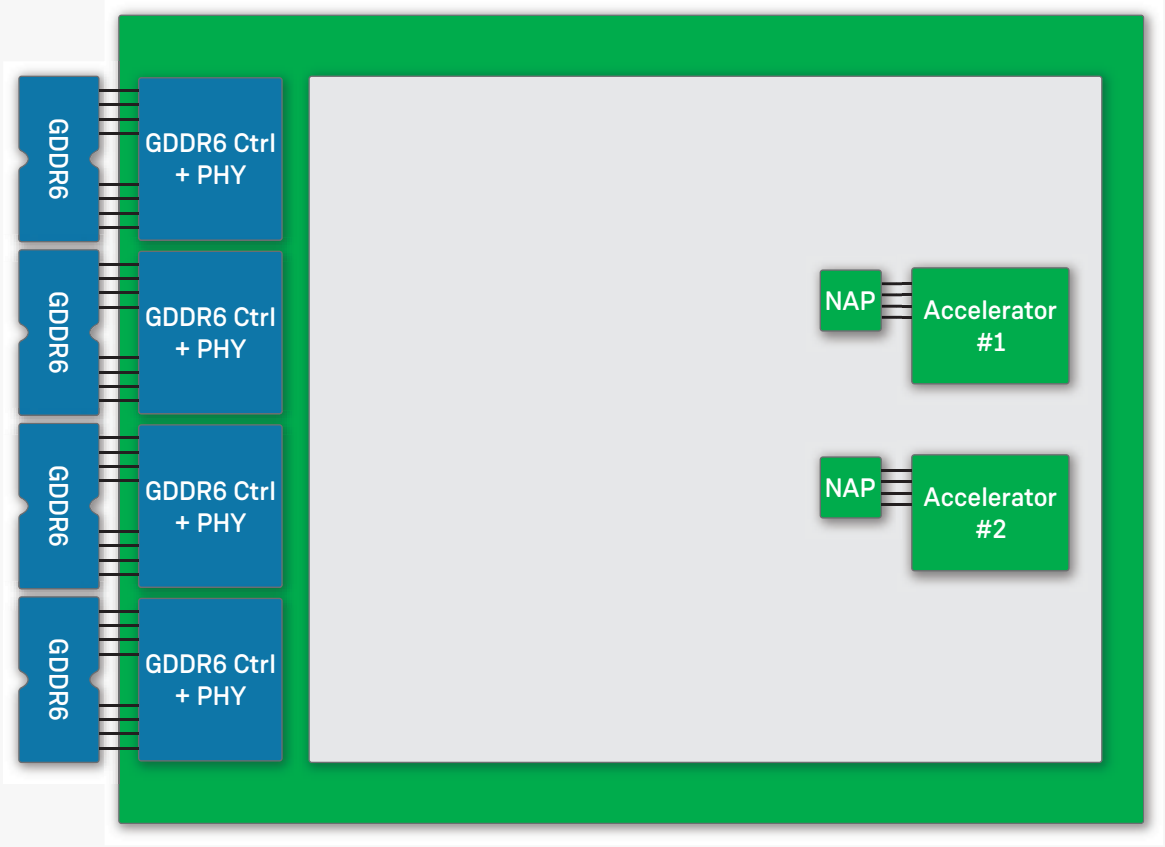

Figure 3 – ***The Speedster7t 2D NoC Enables Software Friendly Hardware***

which simplifies the design, reduces FPGA resources consumption, improves performance and simplifies timing closure.

To enable high-performance arithmetic operations, each Speedster7t device features a large array of programmable compute elements, organized in machine learning processor (MLP) blocks. The MLP is a highly configurable, compute-intensive block that can support up to 32 multiplier/accumulate (MAC) operations in each cycle. The presence of the MLP allows an efficient sharing of resources between fully programmable logic and hardwired arithmetic units in accelerator-focused designs.

Whereas some FPGAs favor the use of the HBM2 memory, in which FPGA and memory are combined into an expensive 2.5D package, the Speedster7t family instead employs the GDDR6 memory standard. This interface delivers the highest performance available for off-chip memories today and at significantly lower cost, making it easier for teams to implement accelerators with high-bandwidth memory arrays. A GDDR6 memory controller can sustain 512 Gbps of bandwidth. With eight banks available on the VectorPath S7t-VG6 (**Figure 4**), aggregate memory bandwidth can reach 4 Tbps. Additionally, there is a DDR4 interface that can be used for data that is accessed less frequently or does not require the throughput of GDDR6.

The VectorPath S7t-VG6 provides a number of high-performance interfaces to support distributed architectures and high-speed host communication. Today, the card is PCI-SIG certified for PCIe Gen 5.0 ×16 at 32 GT/s. In terms of Ethernet connectivity, the card handles ultra-high line rates up to 400 Gbps using widely supported optical interface modules built around the QSFP-DD and QSFP56 standards.

A further MCIO expansion port is available on the opposite end of the card to support a wide array of additional

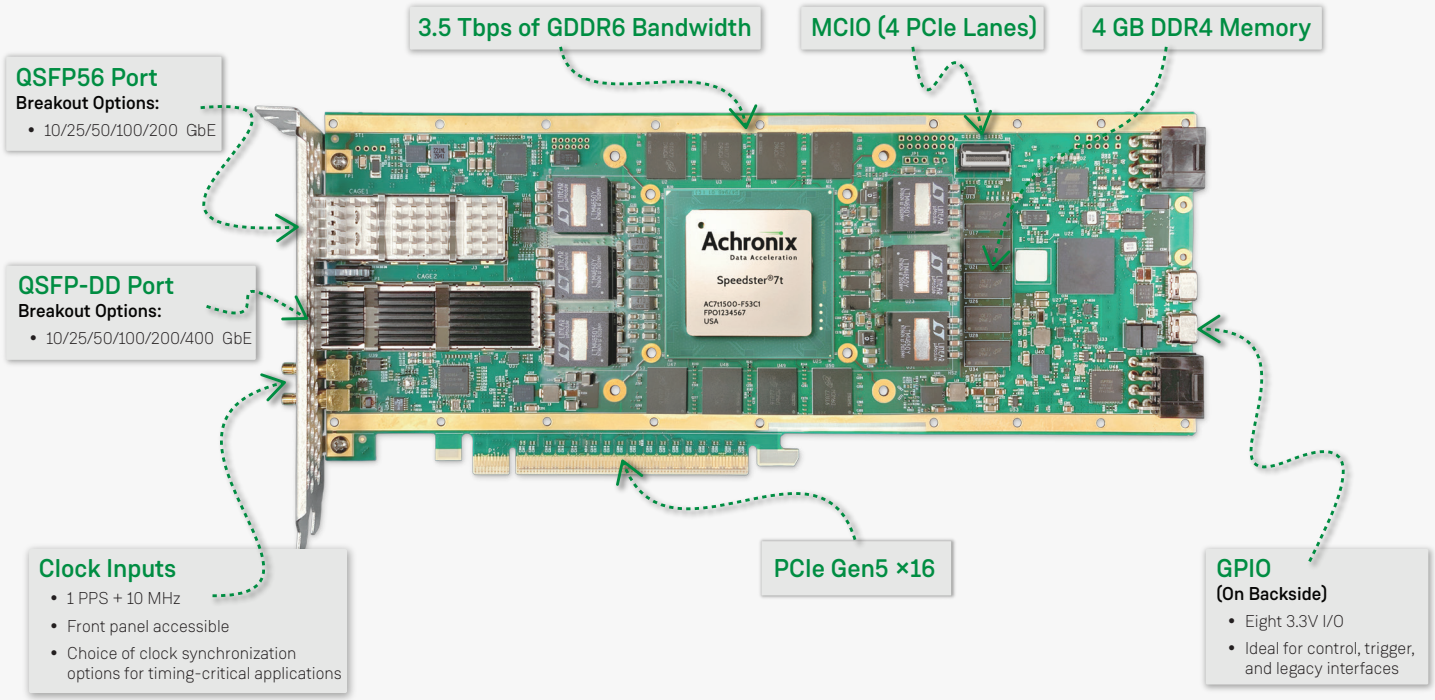

Figure 4 – ***VectorPath Network and Memory Interfaces***

latency critical application use cases. For example, the MCIO port can be used to connect the card to peripherals such as an NVMe storage array for computational storage or database acceleration applications. The MCIO port can also bring in additional network connections beyond the QSPF-DD or QSFP56, allowing for port expansion.

The VectorPath S7t-VG6 also includes multiple clock inputs on the front panel, which is often needed to synchronize multiple cards together. The two SMB clock input connectors support 1 PPS and 10 MHz clock inputs which are connected to jitter cleaners before they enter the FPGA. Once in the FPGA, these clocks can be multiplied or divided to the frequencies needed for the specific application.

Further expansion is available through a general-purpose digital I/O header. This I/O port supports single-ended 3.3V connections and LVDS signals, which allows custom signals such as external clocks, triggers and application specific I/O to be directly connected to the Speedster7t FPGA. This expansion port can also be used to retrofit the VectorPath accelerator card into legacy hardware.

Good for Low and High-Volume Requirements

Every detail has been considered for the VectorPath S7t-VG6 accelerator card. For example, there is support for passive, active air-cooling and liquid cooling. Additionally, long-term supply and support are guaranteed by BittWare and Achronix for sectors such as medical and defense, which require long lifecycle support. In these markets, the short production lifetimes of GPU-based PCIe cards are incompatible with the need for systems that can demonstrate service lifetimes in excess of ten years.

For higher volume requirements, particularly in cases such as edge computing, customers can use BittWare's cost-reduction program to streamline the hardware to support only the I/O options they need. Additionally, BittWare will license board design files and the software and drivers included with the VectorPath S7t-VG6 card. A path to custom system-on-chip devices is also possible with Achronix's Speedcore eFPGA IP. Customers can build their

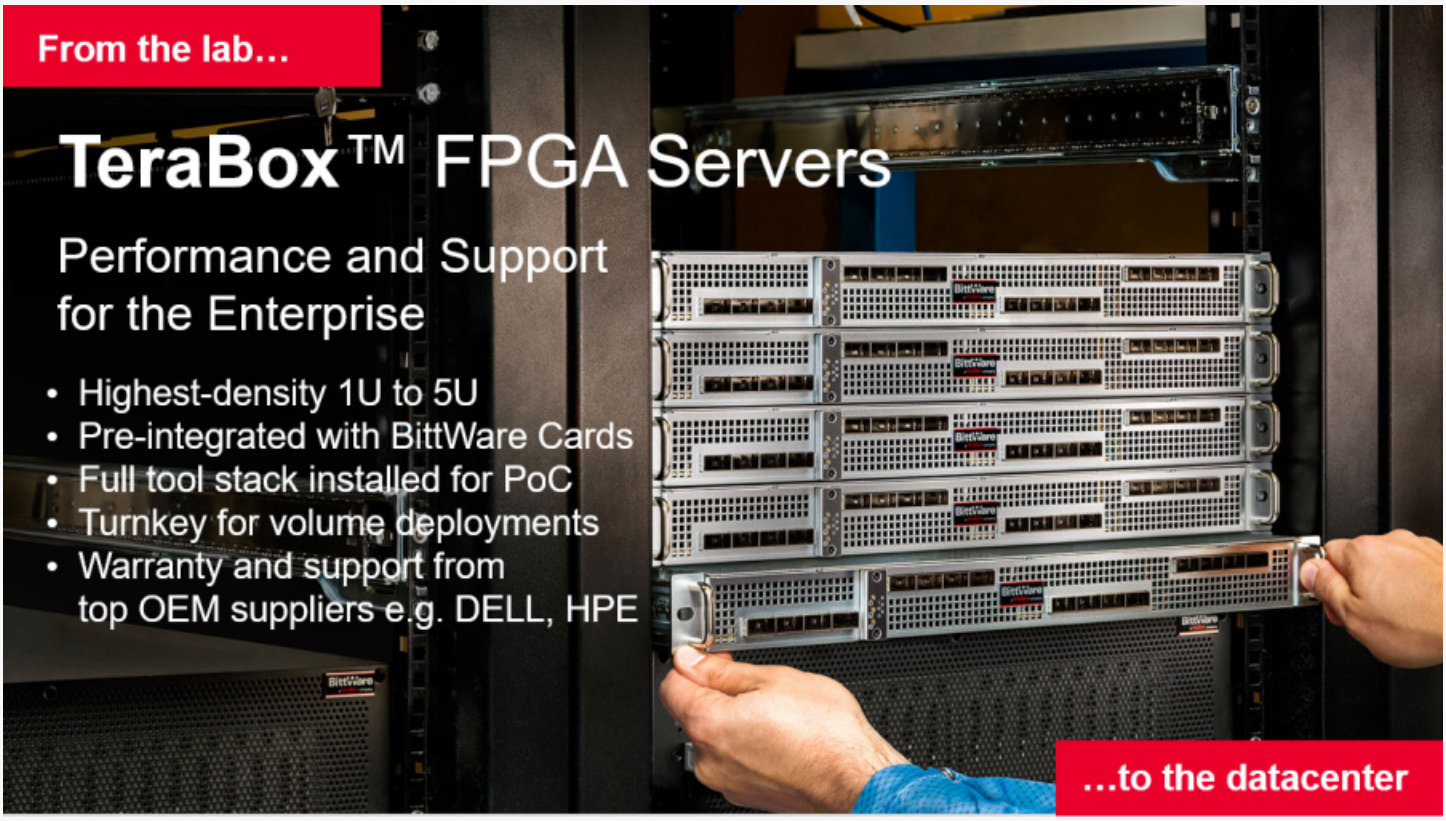

Figure 5 – ***TeraBox Deployment***

own SoC that includes the Speedster7t programmability, but has ASIC cost structure.

For greater development and deployment convenience, the VectorPath S7t-VG6 can be supplied by BittWare in pre-integrated, multicore servers in the form of the TeraBox platform. Ranging from 2U to 5U, TeraBox rack-mount chassis can hold up to 16 BittWare PCIe cards, managed by dual Intel Xeon processors. As a fully packaged solution, TeraBox offers customers the fastest mechanism to get up and running with FPGA development. Supported by the Bittworks II and FPGA Devkit software, users can plug in the TeraBox (**Figure 5**) and start development work straightaway. Alternatively, customers can buy preconfigured servers from Dell and HP Enterprise that contain BittWare cards.

Conclusion

By taking into account the needs of users looking for data acceleration across a wide range of applications, BittWare and Achronix have created a highly flexible engine that can be easily deployed, whether they are to be employed individually or as part of larger heterogeneous processing arrays. The Speedster7t FPGA at the heart of the card provides developers with the ability to build high-throughput applications that can take full advantage of programmable logic and PCIe and up to 400 Gbps Ethernet connectivity. Software and support from BittWare ensures those developers can begin work as soon as the card is plugged in. The flexible nature of the FPGA and the Speedster7t NoC means that the cards can maximize their useful life as applications change and evolve.