Solution Brief

Intel Agilex® 7 FPGAs
4th Gen Intel® Xeon® Scalable processors
Eideticom NoLoad Computational Storage Processor

Eideticom's NoLoad Unleashes the Power of 4th Gen Intel® Xeon® Scalable Processors for Storage Compute

Offloading compression from the host CPU to Intel® FPGA-based accelerators results in significant TCO savings for Financial Services Industry (FSI), HPC, and Enterprise Data Center use cases.

> " NoLoad computational storage combined with Intel Agilex® FPGAs and Intel® Xeon® processors provides best-in-class performance and TCO for demanding storage and compute workloads."
>
> **Sean Gibb**
> VP of Engineering
> Eideticom

Today's computational workloads are larger, more complex, and more diverse than ever. The explosion of applications such as high-performance computing (HPC), artificial intelligence (AI), machine learning (ML), data analytics, and other specialized tasks is driving the exponential growth of data. In turn, processing all this data requires vast amounts of computational power coupled with low latency and high bandwidth access to the data.

Computational Storage

As the storage market grows, new capabilities are needed to move, manage, and protect stored data. Storage processing features—such as virtualization, data protection, data security (encryption), and data compression—are essential to increase storage capacity. However, these capabilities involve many infrastructure services that consume many compute cycles.

The first step towards offloading the infrastructure functions from CPU cores was the introduction of SmartNICs, which augment the ethernet chipsets in regular network interface cards (NICs) with FPGAs. Unfortunately, "SmartNIC" has become a somewhat overloaded term, with different vendors offering wildly different implementations. However, a SmartNIC may be defined as a programmable NIC at its most fundamental level. Another way of looking at this is that a SmartNIC allows the data path portions of infrastructure functions to be offloaded from the CPU cores.

More recently, Intel has taken this offloading process to the next level with FPGA-based infrastructure processing units (IPUs). These cards feature a high-end FPGA coupled with a high-end processor like an Intel Xeon CPU. The combination of the FPGA, which can handle data path functionality, and the CPU, which can handle control path functionality, means that the IPU—which is an evolution of the SmartNIC and which can be thought of as a "Smarter SmartNIC"—can offload the host system to a much greater extent.

The combination of high-speed transceivers, dense logic, and deep memories offered by Intel Agilex® 7 FPGAs and the intellectual property (IP) solutions provided by Intel and its partners allow developers to easily create ideal solutions for online, nearline, and offline storage.

In the data center space, FPGAs offer the low latency offloading necessary to accelerate functions, such as data analytics, artificial intelligence, smart networking, hyper-converged storage, and other functions. FPGAs support in-line, look-aside, and multifunction processing modes to offload CPU workloads by reducing complex bottlenecks (Figure 1).

Authors

Thomas M. Schulte
Sr. Product Line Manager
Intel Corporation

Don Grabski
Strategic Business Development Manager
Intel Corporation

Sean Gibb
Vice President of Engineering
Eideticom

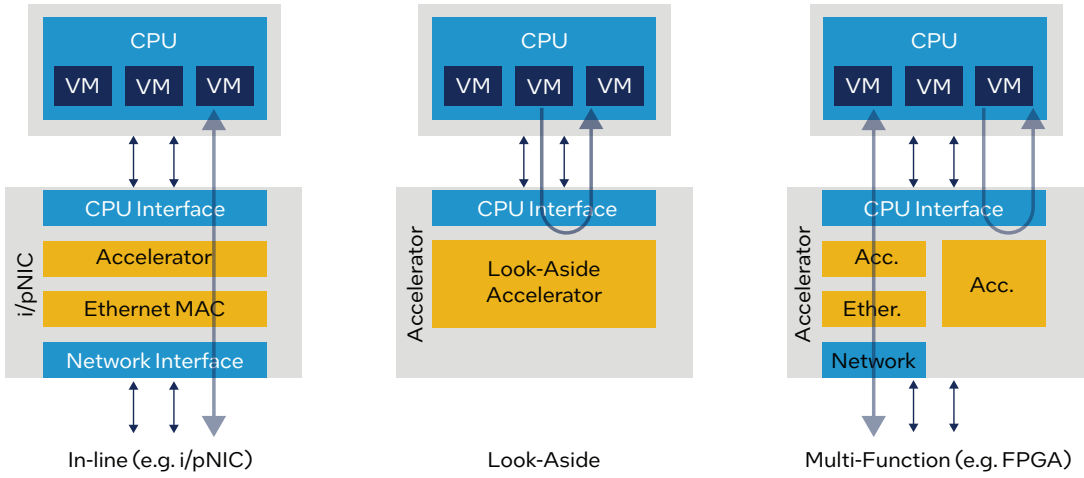

Figure 1. FPGAs support in-line, look-aside, and multi-function processing solutions.

In the case of a new storage paradigm known as computational storage (CS), the system is architected in such a way that computational storage functions (CSFs) are coupled to the storage devices themselves, thereby offloading host processors and reducing the movement of data. These architectures allow CSF compute resources to be deployed either in the SSD storage devices themselves, in which case these would be classed as computational storage devices (CSDs), or on devices located between the SSDs and the host, such as FPGA-based accelerator cards, SmartNICs, or IPUs.

Additional acceleration functions which FPGA-based accelerators could implement include compression and decompression, encryption and decryption, SQL query acceleration, and the acceleration of graph algorithms (Centrality, Pathfinding, Community Detection, and so on).

Another potential application is data transcoding. For example, many databases are currently adopting the open-source in-memory format Apache Arrow—a language-agnostic software framework for developing data analytics applications that process columnar data—because it facilitates efficient analytic operations on modern CPU and GPU hardware. There's also the open-source on-disk data storage format Apache Parquet, which provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk. Transcoding data between Apache Arrow and Apache Parquet, and vice versa, will become increasingly important.

Eideticom unleashes the performance of 4th Gen Intel® Xeon® Scalable processors

As discussed previously, computational storage benefits could improve application performance and/or reduce the host CPU cores utilized, freeing them up to perform other revenue-generating tasks. This increases infrastructure efficiency and improves the total cost of ownership (TCO).

Eideticom is a leader in developing computational storage solutions for data center storage or compute and is an Intel Partner company[1]. Eideticom's NoLoad solution is an NVM Express (NVMe) computational storage processor (CSP). NoLoad computational storage solutions solve the limitations of CPU-centric computing for storage and compute-intensive workloads.

NoLoad is in production and shipping on various form factors from its hardware partners, such as the IA-220-U2[2] U.2 module and the IA-420F[3] card from BittWare, both of which feature Intel Agilex 7 FPGAs which communicate with the host CPU using PCIe 4.0 (Figure 2).

NoLoad provides a suite of functions, including compression and decompression, encryption and decryption, deduplication, and data analytics.

One market that can benefit from NoLoad technology is FSI, a portmanteau of "financial technology." FSI refers to firms using new technology to compete with traditional financial methods in delivering financial services. Artificial intelligence, blockchain, cloud computing, and big data are regarded as the "ABCD" (four key areas) of FSI.

Figure 2. The IA-220-U2 U.2 module (left) and the IA-420F card (right) from BittWare

Benchmark Testing for FSI Analytics Use-case

Recently, a benchmark compared a typical high-end FSI task performed on two different computational environments. This real-world example featured a high-performance, software-defined packet capture and analytics engine.

The first benchmark scenario was run only in software on a pair of 4th Gen Intel® Xeon® Scalable processors, previously codenamed Sapphire Rapids (SPR). The second scenario augmented the CPUs with the FPGA-based accelerator cards, all tied together with Eideticom's NoLoad solution.

The FSI task involves performing compression and decompression on stock market data in 1GB data files.

The hardware for scenario 1 featured two 4th Gen Intel® Xeon® Scalable processors (two socket platforms) running with a 2 GHz clock (Figure 3). Each CPU contained 56 cores with two threads per core, resulting in a total of 224 available cores.

The hardware for the second scenario (Figure 4) used the NoLoad solution implemented on the CPUs and FPGA-based accelerator cards. Two IA-220-U2 cards were used for the compression/decompression.

In this benchmark, packets are compressed and written to an SSD array. The NoLoad software stack allows access to NoLoad compression and decompression services in the filesystem, kernel space, or user space.

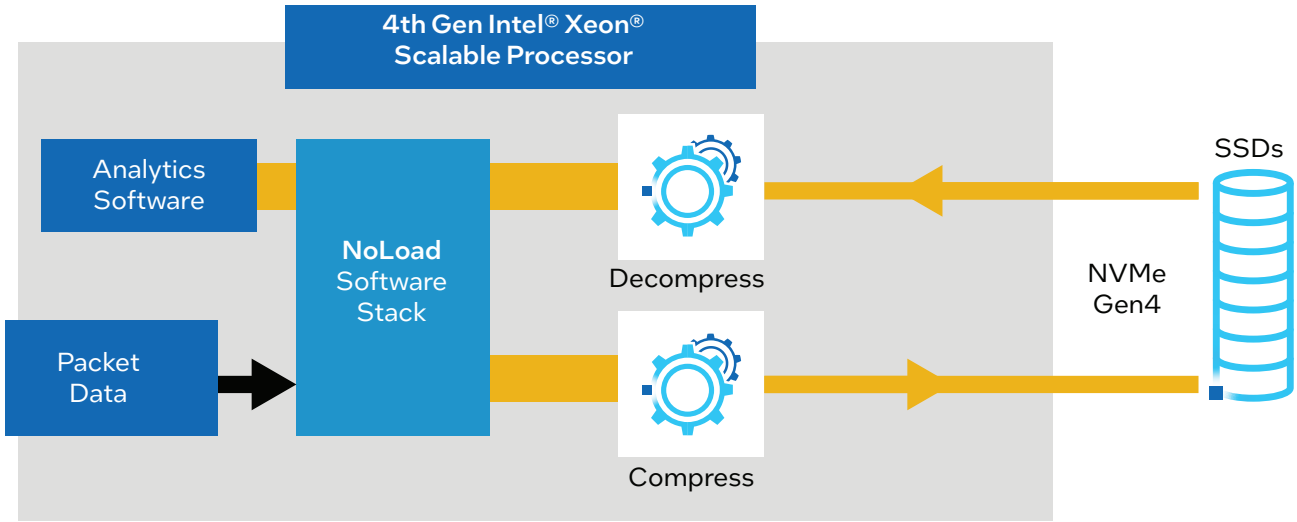

Figure 3. All processing on the data packets is performed on Host CPU processor cores.

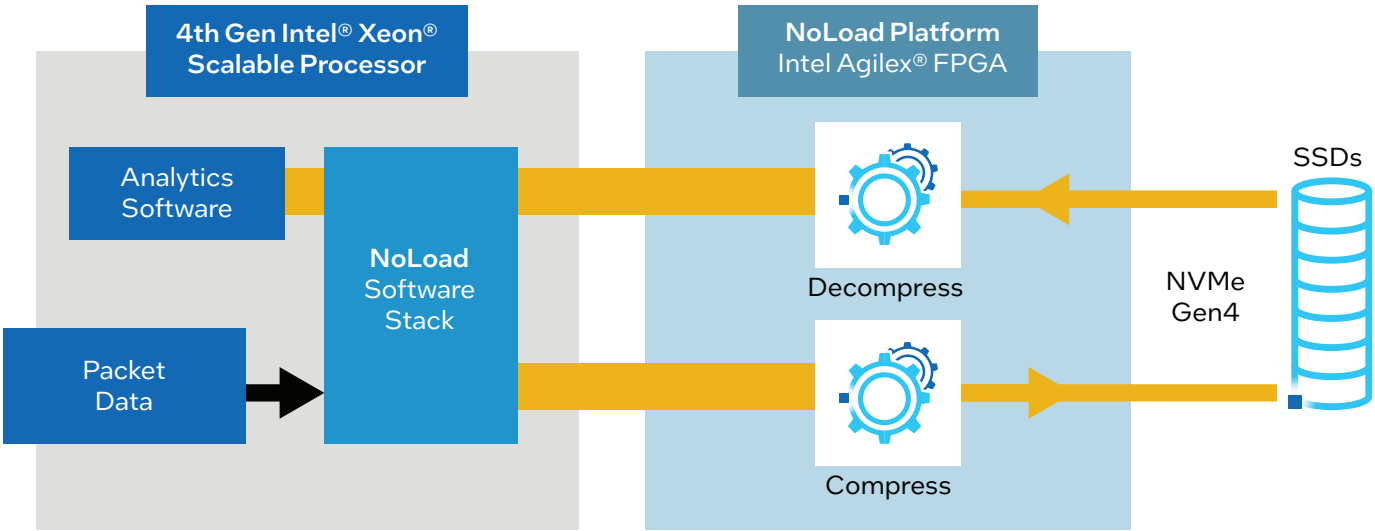

Figure 4. Packets from the host are moved over PCIe to the BittWare IA-220-U2 for FPGA storage services acceleration using the NoLoad framework and IP.

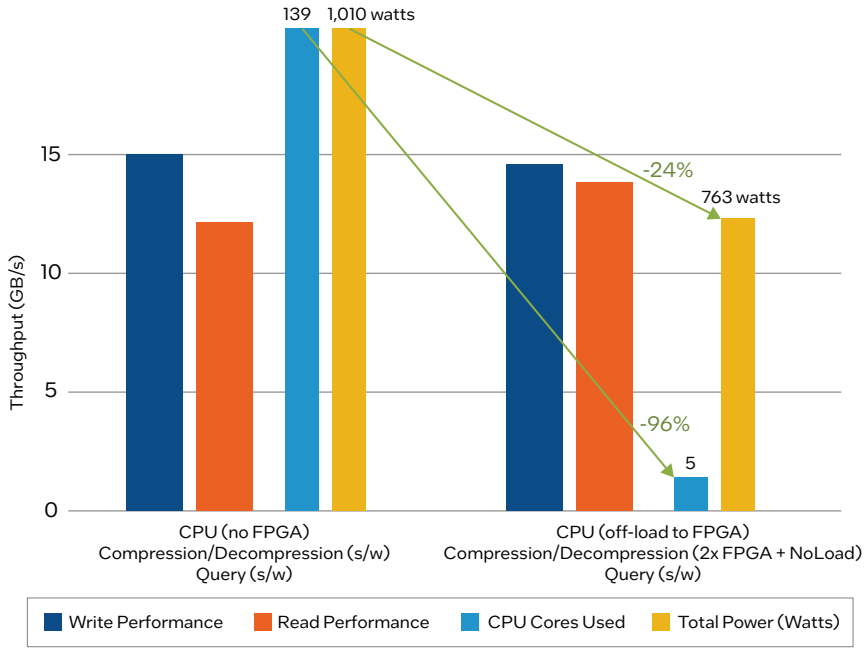

Figure 5. Benchmark results for Eideticom NoLoad solution on 4th Gen Intel Xeon Scalable processors and Intel Agilex 7 FPGA accelerator cards. Data sourced from Eideticom.

Benchmark Results

Analyzing the measurement results (Figure 5) for three key metrics (throughput performance, # of CPU cores used, and total power used) for scenario 1 compared to scenario 2, it is clear the FPGA-based accelerator scenario resulted in nearly identical performance but significantly used fewer number of CPU cores and lower power. The total power for the FPGA scenario is the system's total power, including CPU cores, NMVE storage, and the two FPGA cards. The CPU cores freed up by the FPGA off-load + NoLoad solution are now available for other tasks or workloads.

- **96% reduction** in CPU cores used (lower is better)
- **24% reduction** in power used (lower is better)

TCO Savings Estimates

The financial considerations for the FPGA-based accelerator scenario are shown in Table 1. As a proxy for the value of each CPU core, we use pricing from Amazon's Cloud services EC2 instances. Based on Amazon EC2 F1 instances (virtual cloud service providing FPGA-based accelerators + per CPU core services), we assume a $1.06 per hour cost.[4] This cost includes access to one FPGA and eight virtual CPU cores but to simplify the calculations, assume this price only applies to the CPU cores.

Please contact your local Intel salesperson to create a more detailed TCO estimate for your specific requirements, algorithms, and use case.

Item	Category	Scenario 1 Costs	Scenario 2 Costs	Savings	Comments
Platform + CPU's	CapEx	n/a	n/a	None. Used in both scenarios	–
FPGA cards	CapEx	–	Note 1	One-time, up-front cost	–
NoLoad solution	CapEx	–	Note 2	n/a	–
CPU cores used	OpEX	139	5	$12,784 every 30-days	Note 3
Energy cost/used	OpEX	1,010 watts	763 watts	$12,804 every 30-days	Note 4, 5

Table 1. TCO calculations and assumptions using benchmark results.

Notes:

1. Depends on the FPGA card (Design it yourself or buy off-the-shelf from 3rd party supplier)
2. Contact Eideticom for the NoLoad solution price quote.
3. Value of each CPU core = ($1.06 per hour / 8 cores) x 24-hrs. x 30 days =$95.40.
4. Assumes electricity cost = $0.04 / kW
5. Assume the operation completes in 2 seconds for both scenarios.

4th Gen Intel® Xeon® Scalable Processors

4th Gen Intel® Xeon® Scalable processors are designed to accelerate performance across the fastest-growing compute-intensive and memory-intensive workloads.

With built-in accelerators and software optimizations, previous generation Intel® Xeon® Scalable processors have been shown to deliver leading performance-per-watt on targeted real-world workloads.[5] This results in efficient CPU utilization, low power consumption, and higher return on investment (ROI) while helping businesses achieve their sustainability goals.

These 4th Gen Intel Xeon Scalable processors have the most built-in accelerators of any CPU on the market to deliver performance and power efficiency advantages across the fastest-growing workload types in AI, analytics, networking, storage, and HPC. To enable new built-in accelerator features, Intel supports the ecosystem with OS-level software, libraries, and APIs. Other key features of the new Intel Xeon Scalable processors include support for DDR5, PCI Express 5.0, and Compute Express Link (CXL) v1.1.

Intel Agilex® 7 FPGAs

FPGAs are increasingly used for important roles in modern applications, from the data center to the network to the edge. Their flexibility, power efficiency, massively parallel architecture, and huge input/output (I/O) bandwidth make FPGAs attractive for accelerating and/or offloading a wide range of tasks from AI to storage and networking. Many of these applications put enormous demands on memory, including capacity, bandwidth, latency, and power efficiency. To handle these high-demand applications, Intel has created Intel Agilex 7 FPGAs and SoC FPGAs (Figure 6).

Intel Agilex 7 FPGAs I-Series[6], which are built using Intel 10 nm SuperFin process technology, are targeted at bandwidth-intensive applications. These FPGAs and SoC FPGAs include hardened controllers supporting external DDR4 memory. They also support the FPGA industry's first CXL hard IP, which enables developers to offload latency-sensitive functions to accelerators over CXL interconnect.

Intel Agilex 7 FPGAs M-Series[7] are the first Intel Agilex FPGAs implemented on the Intel 7 process technology and feature in-package HBM2e memory. Intel 7 brings higher programmable fabric capacity and performance while consuming less power. Hardened controllers provide support for state-of-the-art memory technologies such as DDR5 and LPDDR5.

Intel Agilex 7 FPGAs and SoC FPGAs bring high-performance I/O bandwidth, which is critical for systems processing today's enormous data loads, with transceiver rates up to 116 Gbps and support for PCIe 5.0 and CXL 1.1/2.0.

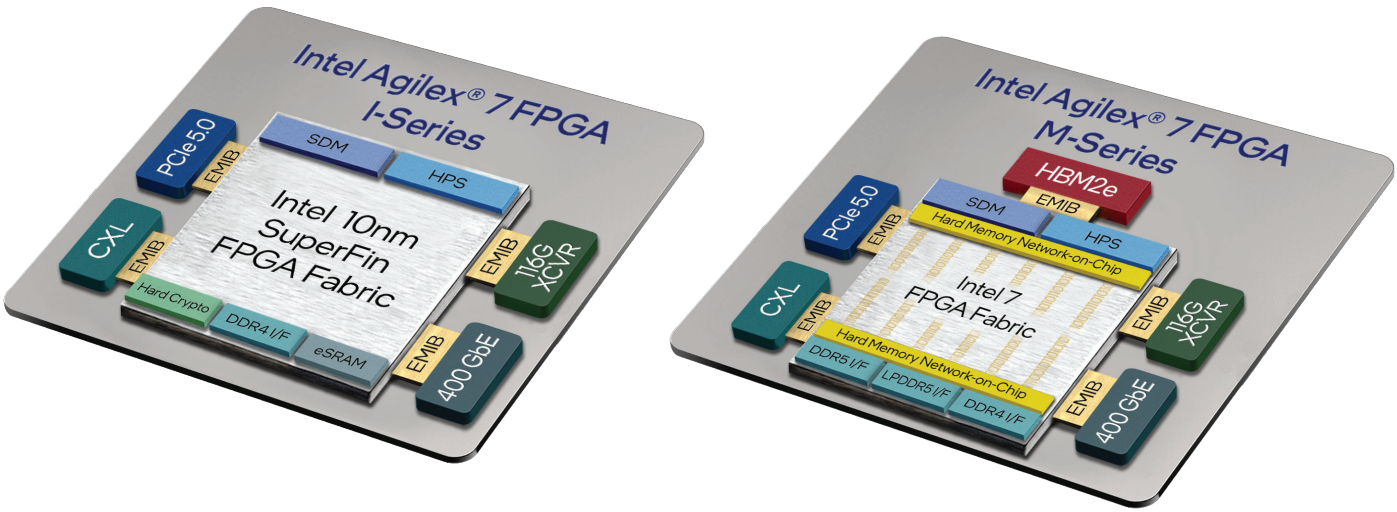

Figure 6. Example members of Agilex 7 SoC FPGAs: I-Series (left) and M-Series (right).

Summary

Today's computational workloads are larger, more complex, and more diverse than ever. By combining the newest Intel products with innovative solutions from partners, such as Eideticom and Bittware, customers could achieve significant TCO savings for targeted use cases or workloads.

FSI is one of many cases where it is advantageous to offload algorithmically intensive and latency-sensitive functions to Intel Agilex FPGA-based accelerator cards, thereby unleashing the power of 4th Gen Intel Xeon Scalable processors by freeing up host CPU cores to perform other money-earning tasks.

Contact Eideticom as they investigate porting the Eideticom NoLoad solution to a newer BittWare card (IA-440i[8]), which might reduce the FPGA cards used from 2 to 1 for this FSI use case.

To learn more, contact your local Intel representative today.

References

1 Intel Partner = Eideticom
https://www.intel.com/content/www/us/en/partner/showcase/storefront/a5S3b000000BBQ8EAO/eideticom.html

2 https://www.bittware.com/fpga/ia-220-u2/

3 https://www.bittware.com/fpga/ia-420f/

4 https://aws.amazon.com/ec2/instance-types/f1/ , price per hour for f1.2xlarge 1-yr. reserved instance effective for Linux/Unix in the US East (Northern Virginia), as of April 1, 2023.

5 4th Gen Intel Xeon Scalable Processor performance benchmarks at
https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/

6 https://www.intel.com/content/www/us/en/products/details/fpga/agilex/7/i-series.html

7 https://www.intel.com/content/www/us/en/products/details/fpga/agilex/7/m-series.html

8 https://www.bittware.com/fpga/ia-440i/

Additional Resources

Overview of Intel Agilex FPGA Portfolio:
https://www.intel.com/content/www/us/en/products/details/fpga/agilex.html

Learn more about Intel Agilex FPGAs with R-Tile:
https://www.intel.com/content/www/us/en/products/docs/programmable/fpga-leadership.html

SS-1161-1.0