Solution Brief

System Integrators, Aggregators, ISV
Smart Factory, Medical

DFI Server-Grade Industrial Board for AI Visual Inspection

The DFI ICX610-C621A features a 3rd generation Intel® Xeon® Scalable processor for compute, with SATA and PCIe 4.0 expansion slots for storage and connectivity, providing edge computing for AI and video processing in smart factories.

Industrial edge computing brings multiple advantages to industrial operations. Increasingly, it provides a very fast, more responsive ability to process, analyze and act on data gathered by Industrial Internet of Things (IIoT) sensors compared to the common cloud-based processing models.

Sensors with high-speed communications capabilities are inexpensive now and deployed throughout manufacturing plants and other industrial settings. Moreover, new factory equipment is usually computer-equipped to monitor and relay performance data.

When those sensors or components have to relay data to the cloud for processing, the process introduces latency, which can make the data less useful especially in applications that need real-time response. Factory robotics, for example, is a widely deployed technology today, and generates large amounts of data through daily operations. If this data is sent to the cloud, bandwidth eventually becomes overwhelmed as the volume of data grows.

The data that these systems relay can include information on their speed, efficiency, cycle time, operating temperature, usage patterns, maintenance needs or other data pertinent to the system's purpose in the facility. It also includes environmental information such as ambient temperature, airborne particles, or potentially, anything else that sensors can measure. This data is critical for decision making and, when augmented with artificial intelligence, can enable systems to take actions without human intervention.

Processing at the edge

Industrial edge computing moves the IoT and industrial data workload processing to the outer reaches of the network. This enables very fast response and the use of AI for process improvement. To be effective, the edge compute node must provide ample bandwidth and processing power.

Although edge computing has been deployed in industrial automation for years, IoT, video and industrial process computing are now demanding more processing power. The number of IoT sensors in use is increasing, and switching to high-definition (HD – 1080 pixels) security video or even ultra-high definition (UHD – 4k pixels) can dramatically increase data processing needs. These new applications are outpacing these legacy compute systems which have traditionally been built by a system integrator (SI) using compute, storage and networking functionality from a variety of manufacturers. This approach can lead to higher overall operational costs due to longer time to integrate and debug. Moreover, this approach usually leads to increased hardware costs as lower performance systems are outmatched by high bandwidth workloads.

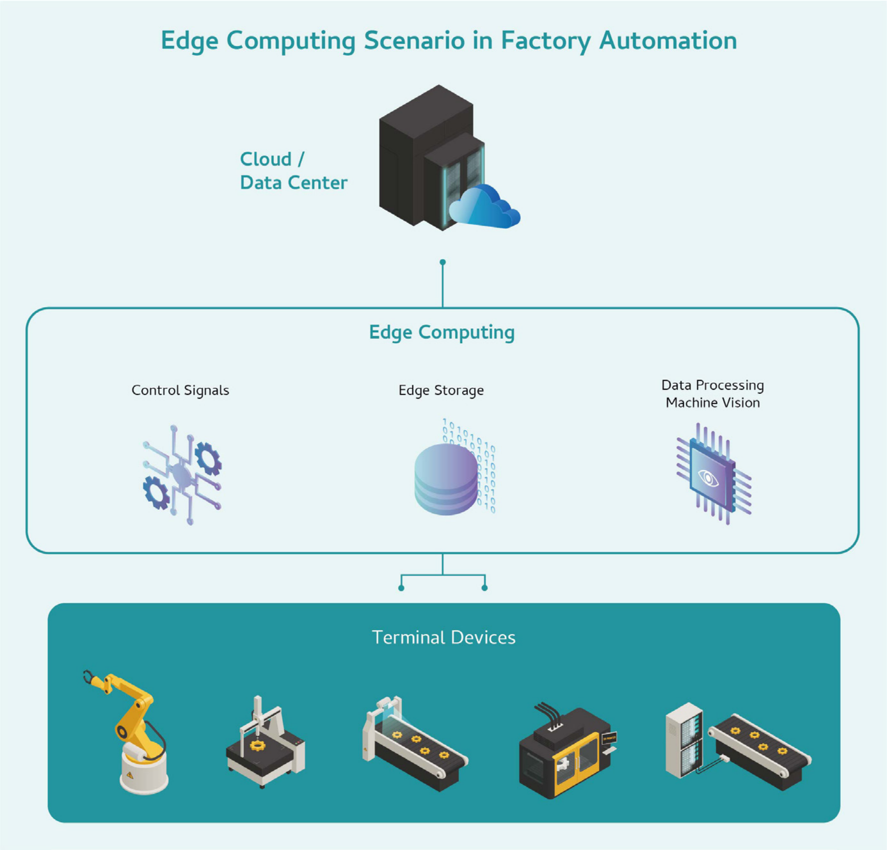

Figure 1. An example architecture for industrial edge computing in factory automation applications. The terminal device sends data to the edge computing server. The server analyzes the data, sends back the corresponding control commands, and stores the data.

DFI Inc., a global provider of embedded computing solutions and an Intel® Network Builders ecosystem member, developed the ICX610-C621A server board, enabled by 3rd Gen Intel® Xeon® Scalable processors to provide a complete solution for industrial edge networking applications. This flexible platform allows system integrators to design custom, powerful and more-secure AI-enabled solutions for industrial manufacturing. It serves as a motherboard for an edge server that offers compute power to match the data flows created by smart factory systems especially those using video or AI.

Processing Provided by 3rd Generation Intel® Xeon® Scalable CPUs

In a traditional industrial edge computing architecture, cameras and other sensors transmit their data to dedicated small form-factor computers. Most of these computers have the compute performance for a single HD camera each, or possibly up to three if the frames-per-second rate is low. In a smart factory, with fast-moving manufacturing processes and many cameras all providing HD video, one camera to one computer is the norm. The computers, in turn, relay their analytics to the cloud, and onto a data server for storage and retrieval.

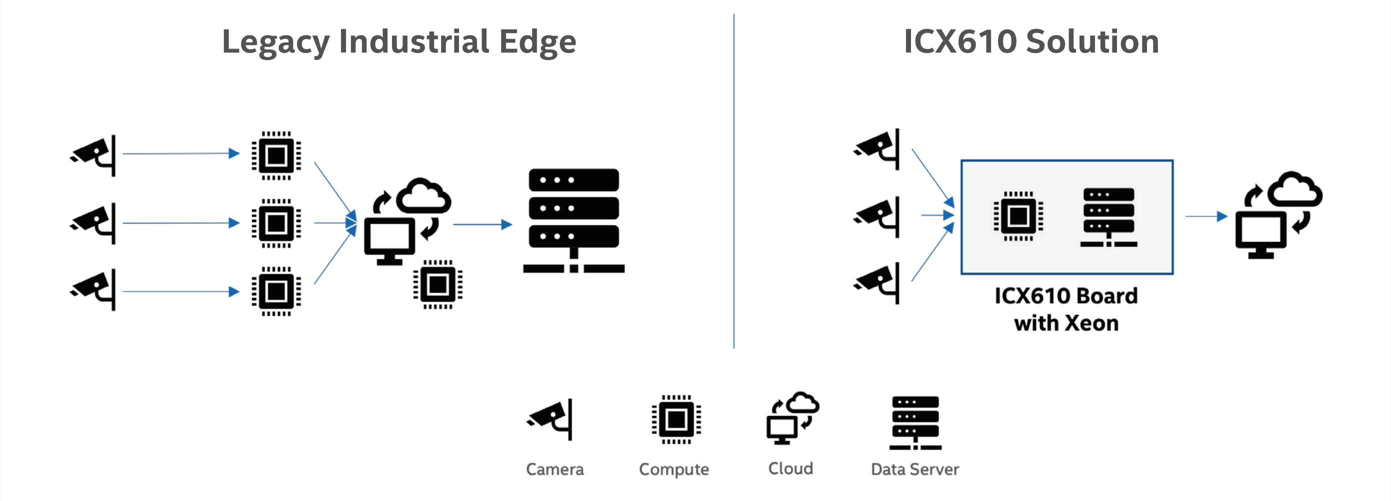

Figure 2. The typical industrial edge structure requires transmitting information to the cloud for storage and retrieval, introducing latency. DFI's ICX610 solution keeps the data on the motherboard.

The ICX610-621A provides an alternative, a high-performance compute with flexible expansion slots for a complete system including I/O for storage and networking all on a single board, streamlining industrial edge computing dramatically. Built on Intel's 3rd generation Xeon Scalable processor, it brings compute power equivalent to up to 40 small form-factor computers. This means buying fewer servers which can reduce capital costs and fewer servers also reduces electricity usage compared to buying and operating dozens of small form-factor computers.

Figure 3 is the layout of the ICX610-621A. The illustration shows a pad for a number of 3rd generation Intel Xeon Scalable CPUs that utilizes the LGA 4189 (Socket P+) socket. It also shows the three PCIe x16 slots and two PCIe x8 for high-performance input/output (I/O) connectivity along with six SATA 3.0 ports for connecting to storage drives.

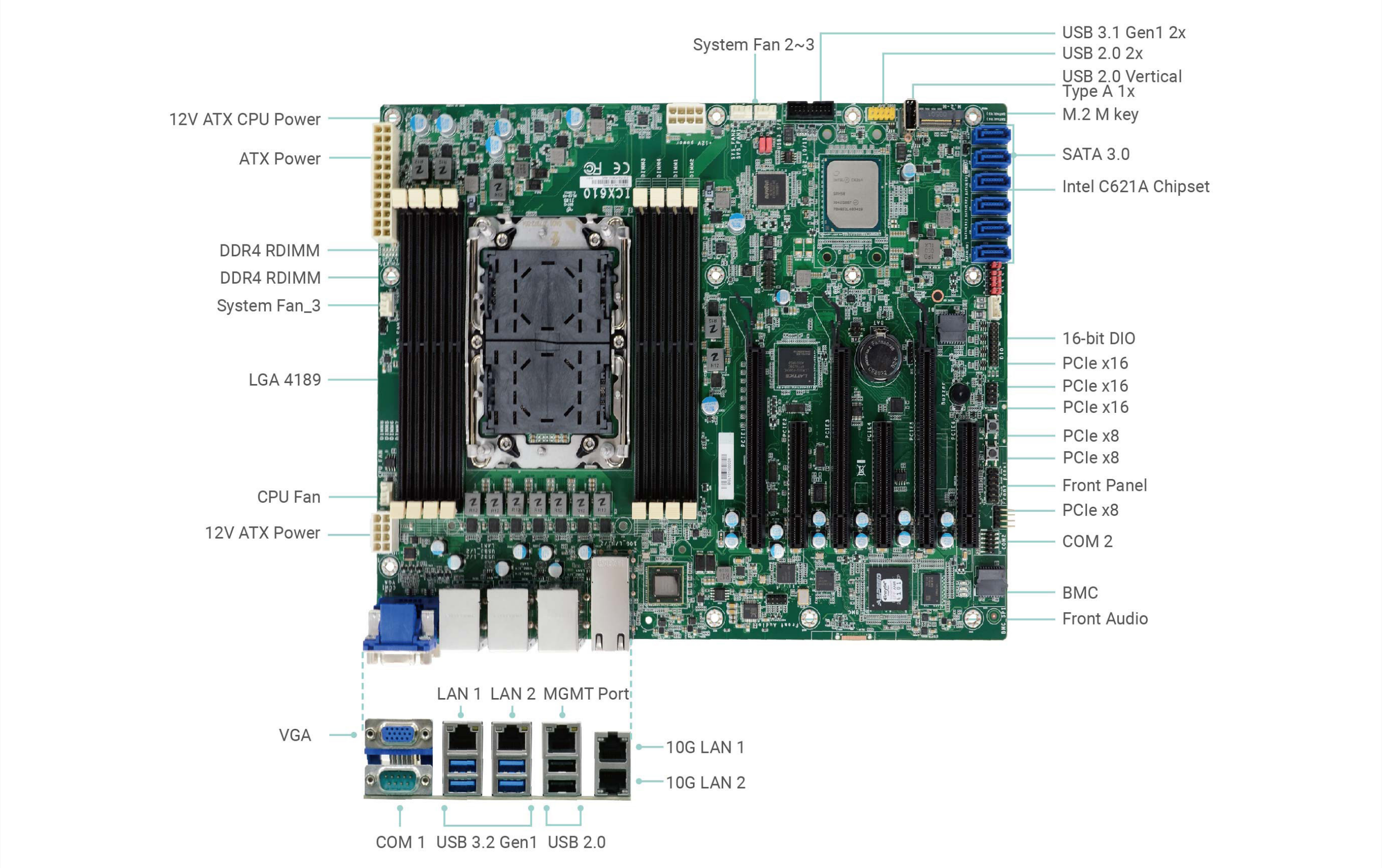

Figure 3. Layout of DFI ICX610-621A.

Powerful artificial intelligence inferencing

The compute power built into the ICX610-621A can support artificial intelligence inferencing, an advanced stage of machine learning in which AI software is able to make predictions and decisions based on new data.

The 3rd Gen Intel Xeon processors deliver up to 1.56x improvement in AI inference applications for image classification compared to the previous generation.[1] Embedded technologies, in particular Intel® Advanced Vector Extensions 512 (Intel® AVX-512), help increase the performance of inference workloads on Intel architecture.

The 3rd generation Intel Xeon Scalable processor features Intel® Deep Learning Boost, an instruction set for Intel Xeon Scalable processors that improves performance on deep learning tasks such as training and inferencing, sharpening the system's inferencing capability. The ICX610-C621A board features Intel's Deep Learning Boost on the CPU, strengthening its ability to act on data without human intervention.

This offers a significant speed increase for targeted workloads, including image recognition, image segmentation, and object detection that are common to AI-assisted medical image analysis. It can also apply to any industrial use case that involves analyzing visual data.

When creating an industrial AI solution using the ICX610-621A, SIs deploy the technology first in a laboratory setting for training. This enables the development of a certain degree of acumen in the system before it can affect anything in production. When the training is complete and the technology is implemented in production, it continues to learn desired behaviors and improve its ability to make decisions.

The training teaches the AI to infer information from the available data and act on its inference. AI Inferencing requires processing large volumes of data to provide the AI software enough information for it to infer conclusions and make decisions. High throughput and low latency are important to moving the necessary amount of data quickly, which the ICX610-621A line accomplishes by keeping all of the components on a single board.

ICX610-621A Inferencing Performance

DFI conducted testing on the ICX610-C621A in AI inferencing applications to demonstrate its performance. An ICX610-C621A was used in a high-precision bottle cap inspection application. AI inferencing requires rapid processing of information, which is partly influenced by the format of the data it is using. Single-point precision, which uses 32-bit numbers with a floating decimal point (also called FP32), has long been a common configuration. It provides precision, but throughput is comparatively slow in modern applications. Int8, which uses 8-bit numbers and a fixed decimal is more than twice as fast.

To show this, the company set up a test using Intel® OpenVINO™ Deep Learning (DL) Workbench for the FP32 results, with the DL Workbench set to automatically use the Vector Neural Network Instructions (VNNI) that are a part of Intel® Deep Learning Boost for the lower resolution Int8 testing. Int8 requires less overall storage capacity and less access bandwidth due to its smaller data size, which naturally reduces processing latency and increases throughput. When Int8 is selected from the optimize tab, DL Workbench automatically quantizes the selected model to Int8 by running a calibration procedure and then generating a quantized version of the model.

Figure 4 shows the speed and latency difference between FP32 and Int8, the latter data format demonstrated significantly higher throughput and much lower latency. With no other changes in the testing environment, Int8 demonstrated average throughput of 1399 TFLOPS compared to 253 TFLOPS for F32. Latency for Int8 was 0.69 ms, compared to 3.91 ms for FP32.

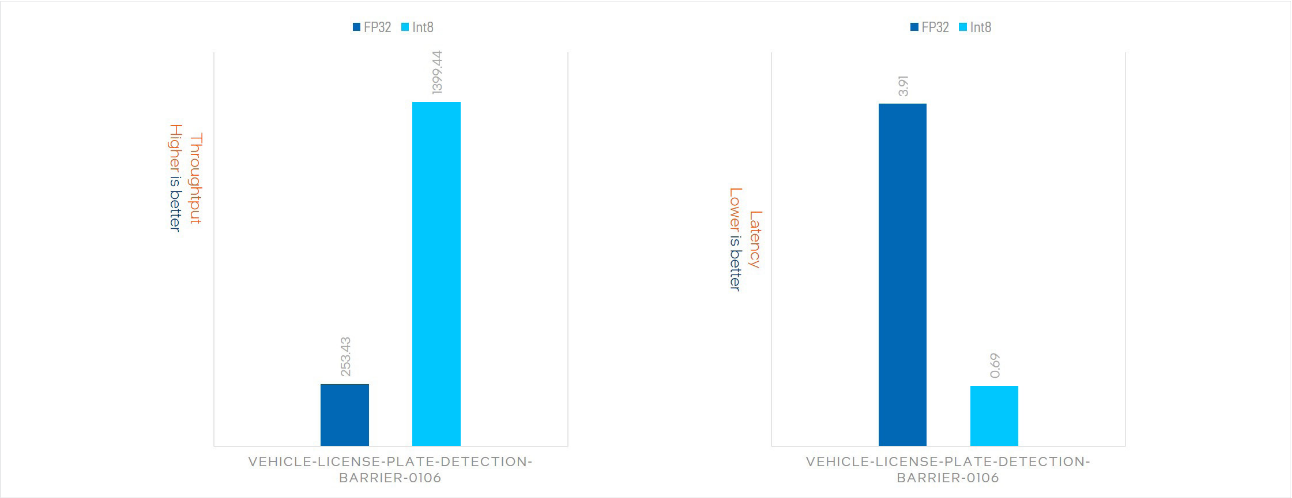

Figure 4. Performance of the ICX610-C621A single board computer in measuring FP32 and Int8 data formats in a license plate recognition application.

Conclusion

Industrial edge computing is a necessary part of the transformation of industries, including manufacturing, health care and many others. DFI's ICX610-C621A boards bundle all of the key components of edge computing on a single card, which vastly simplifies deployment and improves edge computing performance.

Organizations that choose this self-contained architecture can achieve performance gains and lower total cost of ownership, while enhancing their ability to apply automation to the work they do.

The edge is not the future of industrial computing; the edge is already here. DFI and Intel are making it easier for computing resources to operate at remote locations within the network.

Learn More

DFI.com

Intel® Network Builders

3rd generation Intel® Xeon® Scalable Processors

Intel® Deep Learning Boost

Notices & Disclaimers

[1] https://www.intel.com/content/www/us/en/products/docs/processors/embedded/3rd-gen-xeon-scalable-iot-product-brief.html

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

0322/TM/H09/PDF ♻ Please Recycle 349508-001US