FPGA-AS-A-SERVICE: A GUIDE FOR CSPs
The data center is coming under immense pressure. Connected cars, smart factories and smart cities are among the technologies driving a boom in connected devices, which Gartner predicts will see 20.4 billion connected things worldwide by 2020¹. The volumes of data flowing from these devices are also increasing.

At the same time, cloud service providers (CSPs) must work within tight constraints in the physical environment of their data centers, with restrictions on the space, power and cooling available. CSPs need to process the growing data volumes in a way that is power efficient, and has sufficiently low latency to support real-time applications, such as artificial intelligence.

Some Tier 1 CSPs are using Field Programmable Gate Arrays (FPGAs) to accelerate customer workloads at a lower level of power consumption. FPGAs are semiconductor devices which can be reconfigured after shipping to provide hardware circuitry that is tailored for a particular process, such as encryption or data analytics. Because the processing is carried out in hardware, it’s significantly faster than it would be if carried out in software on a general purpose processor. FPGAs can be used for a wider range of workloads than GPUs, including networking, storage and encryption, and are more power efficient than a CPU. By increasing the throughput of servers, FPGAs help CSPs to meet demanding data processing requirements within the constraints of their data centers.

In the past, using FPGAs has required specialist programming skills, including a good understanding of how the hardware works, to configure each accelerator. Thanks to the latest developments in software and the FPGA ecosystem, it’s much easier to use FPGAs today. In this guide, we’ll introduce some of the technologies and strategies that can help next-wave CSPs to deliver FPGAs-as-a-Service, providing acceleration to their customers on demand.
To help simplify the deployment of FPGAs in the data center, Intel has created the Intel® Programmable Acceleration Card (Intel® PAC). The first board in the family is based on the Intel® Arria® 10 GX FPGA (Intel® PAC with Intel® Arria® 10 GX FPGA). The second generation board will be based on the Intel® Stratix® 10 SX FPGA and is sampling now. It will be in production in 2019. Devices in the Intel PAC family will be able to share the same accelerator code.

The Intel PAC seamlessly pairs with an Intel® Xeon® processor over the PCIe bus.

The Intel PAC with Intel Arria 10 GX FPGA is being validated with server OEMs now. It is already available in servers based on the Intel® Xeon® Scalable processor from Dell® and Hewlett Packard Enterprise* (HPE*).
The Intel® Acceleration Stack for FPGAs makes it easier for next-wave CSPs to integrate FPGAs in their data centers. It provides:

- An acceleration environment that makes it easier to deploy workloads to the FPGA and to interact with the FPGA from within user applications;
- A suite of developer tools to support developers as they create accelerator configurations;
- Pre-built accelerator libraries for popular applications; and
- Industry-standard software frameworks.

Using the acceleration stack, the FPGA can be accessed directly from within virtual machines or containers, or indirectly through standard libraries that take advantage of FPGAs. Using an encryption library that has been extended for FPGAs, for example, enables encryption functions to be accelerated in hardware without the host application needing to work with the FPGA directly.

The acceleration stack introduces a degree of standardization that will enable accelerator functions to be reused more easily across different FPGA models and generations.

**Figure 2:** The Intel® Acceleration Stack for FPGAs
ADDING OFF-THE-SHELF ACCELERATOR FUNCTIONS

The easiest way to configure the FPGA is to use preprogrammed accelerators or frameworks that support FPGAs. Here’s a selection of acceleration solutions that work with the Intel Acceleration Stack.

ACCELERATING IMAGE PROCESSING APPLICATIONS

CSPs that provide image processing functions as a service, or whose customers carry out image recognition and manipulation, can benefit from using the following software tools with pre-built accelerator functions.

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

OpenVINO™ is short for Open Visual Inference and Neural network Optimization, and is designed to accelerate real-time vision applications on Intel® architecture. Such applications might include facial recognition or image recognition.

The Intel® Distribution of OpenVINO™ toolkit provides a single API that works across CPUs and different computer vision accelerators, including FPGAs, GPUs, and the Intel® Movidius™ Neural Compute Stick (NCS). It also accelerates the OpenCV®, OpenCL™ and OpenVX* computer vision libraries.

The Deep Learning Deployment Toolkit is part of the OpenVINO toolkit and it enables you to use a high level API to deploy a trained neural network for Caffe® and TensorFlow® (among others) on the FPGA.

Find out more about the Intel® Distribution of OpenVINO™ toolkit

CTACCEL

CTAccel Image Processing (CIP) is an FPGA-based image processing accelerator, which can be used for a wide range of image-based applications including thumbnail generation and transcoding, and image sharpening or color filtering. It is compatible with the popular open source image processing application ImageMagick®, and the OpenCV library, so cloud customers can use familiar software while benefiting from hardware acceleration. The FPGA can be used to accelerate JPEG decoding, image resizing and cropping, and webp encoding operations.

Find out more about CTAccel Image Processing
ACCELERATING DATABASE AS A SERVICE

Some CSPs specialize in helping their customers to handle the vast volumes of data generated by business today. For those using Apache Cassandra*, rENIAC’s Data Engine can accelerate performance on servers with FPGAs.

APACHE CASSANDRA*

Apache Cassandra is an open source database, which can distribute data across clusters within or spread across data centers. rENIAC’s Data Engine enables FPGAs to be used with standard servers to accelerate data centric workloads. The Data Proxy for Cassandra, built on this Data Engine, enables Cassandra’s read performance to be sped up without any changes to the application code. The solution is hosted on an additional server equipped with an FPGA. This server sits between the database clients and the database server.

Find out more about rENIAC’s Data Proxy for Cassandra

ACCELERATING ANALYTICS AS A SERVICE

For CSPs providing analytics services to their customers, there are off-the-shelf options for accelerating Apache Spark* and common SQL databases.

APACHE SPARK*

The Bigstream Hyper-acceleration Layer* (HaL*) is designed to accelerate database and analytics solutions, with Apache Spark being the first supported platform. The Apache Spark representation of the computation is translated into a platform-independent dataflow, which the Bigstream solution can then translate into FPGA-optimized code.

Find out more about Bigstream’s Hyper-acceleration Layer

SQL DATABASES

Relational databases and SQL underpin many big data and analytics applications today. S64DA* from Swarm64* extends common databases including PostgreSQL*, MariaDB* and MySQL*, using Intel FPGAs to accelerate analytics performance. The solution works by optimizing the data structures of the database and the dataflow, so that each query needs to touch less data, and data that does need to be processed can move through the system more quickly.

Use cases for S64DA include applications that insert data at high velocity (up to millions of rows per second), searching or aggregating ranges in large data sets, and near real-time requirements (sub-second timing between information being received by the database and it being made available to queries).

Find out more about S64DA
FPGAs have in the past had a reputation for being difficult to program. In the early days, hardware description languages (HDL) such as Verilog* were required. Rather than describing the expected program behavior, developers had to describe the hardware configuration, which required highly specialized and technical knowledge.

Today, OpenCL™ provides a standard language for creating programs that can be implemented across CPUs, GPUs, digital signal processors (DSPs), and FPGAs. OpenCL makes coding for FPGAs much more approachable, and the C-like programming language offers an easy migration path for those experienced in coding for GPUs.

The Intel® FPGA SDK for OpenCL™ abstracts away the traditional hardware-focused FPGA development flow, with the OpenCL accelerator code translated into Verilog, and then implemented on the FPGA by the Intel Quartus Prime software suite (see Figure 3).

The SDK enables users to quickly emulate their OpenCL kernels to validate their functionality, and receive feedback on any bottlenecks. The profiler can be used to examine system performance and get direct insight into the architectural bottlenecks of the design. The SDK helps with future-proofing, because code can be reused across different FPGA families or generations without code modification.

CSPs can take advantage of OpenCL by coding their own accelerator functions and making them available through an API. Customers could, for example, use an API to send and receive files for encryption or media transcoding, with the performance enhanced by FPGA acceleration. CSPs can also consider allowing customers to deploy their own OpenCL accelerators in their cloud infrastructures, although they may wish to add consultancy and testing services to mitigate the risks associated with using untested hardware configurations.

Many of the functions that benefit most from acceleration, such as media transcoding and encryption, are common to a wide range of applications. CSPs might be able to avoid the cost of recreating software that others have already written by licensing existing accelerator functions from other companies.

Find out more about the Intel® FPGA SDK for OpenCL™
Discover the Intel® Quartus® Prime software suite
**USING FPGAS WITH ORCHESTRATION SOFTWARE**

Intel’s Acceleration Stack for FPGAs enables orchestrators to discover and manage FPGAs like any other compute resource, and hot swap accelerator cores into them in seconds. Supported orchestrators include OpenStack*, Kubernetes*, VMware vSphere*, and KVM in Linux*. The integration with standard orchestrators simplifies the deployment and use of FPGAs at cloud scale. Using VMware vSphere 6.7 Update 1, VMware DirectPath I/O technology is used to enable virtual machines to access the FPGA directly, enabling near-bare-metal performance.

The FPGA can only be used by one virtual machine at a time, but use of the FPGA can be allocated to different virtual machines according to customer and business requirements. Multiple FPGAs can be installed in the same server enclosure.

**CONCLUSION**

Offering FPGAs-as-a-Service gives CSPs an opportunity to differentiate by enabling customers to get increased performance for some of their most intensive processing requirements.

The Intel PAC can be used with validated servers to more easily add FPGAs to the data center, and the Intel Acceleration Stack streamlines the discovery, provisioning and management of FPGAs at data center scale. Applications can be accelerated by using existing software frameworks, libraries and acceleration layers, and OpenCL can be used to streamline the development of new accelerator functions. CSPs can also seek to license ready-made accelerator functions that are compatible with the Intel Acceleration Stack.

Popular orchestration software, including OpenStack, Kubernetes, VMware vSphere and KVM can be used to manage FPGA compute resources in the data center.
FURTHER READING

- Intel® FPGA Acceleration Hub
- Intel Resources for Cloud Service Providers


All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer, or learn more at Intel.com/csp

Intel, the Intel logo, Xeon, Arria, OpenVINO, Quartus, and Movidius are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos.

*Other names and brands may be claimed as the property of others.

© Intel Corporation 0219/CAT/LH/PDF 338652-001EN