Case Studies: HPC Acceleration with Intel FPGAs

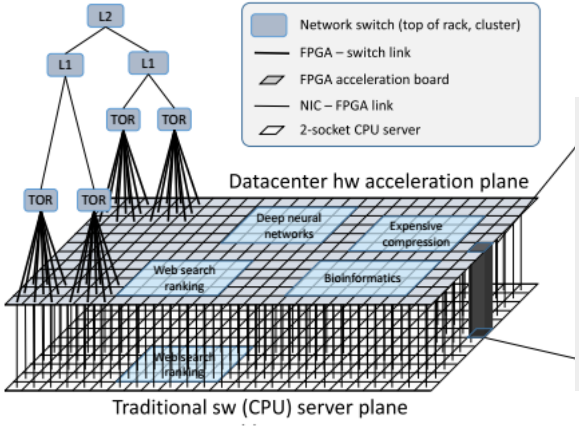
For International Workshop on FPGA for HPC
March 12th, 2018

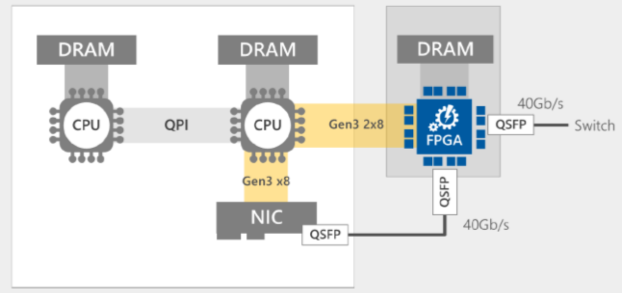
Brad Kadet
Intel® Programmable Solutions Group
FAE Director, Japan

Agenda

- Introduction to FPGA Acceleration (Microsoft Example)
- Machine Learning Acceleration
- Data Analytics Acceleration
- HPC Acceleration Examples
- FPGA Acceleration Platform and Card
- Summary

Microsoft FPGA Acceleration





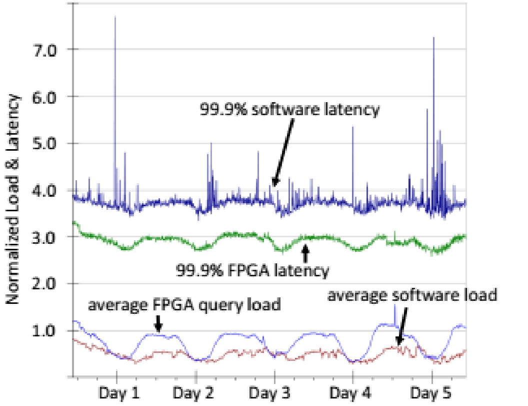


Fig. 7. Five day query throughput and latency of ranking service queries running in production, with and without FPGAs enabled.

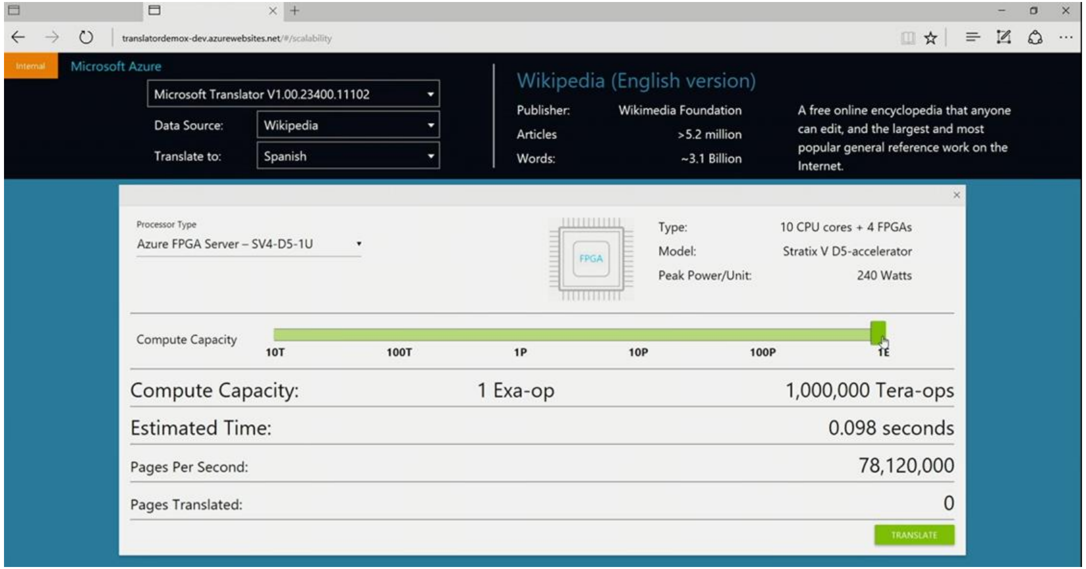
Microsoft Scale-Out FPGA Multi-Function Accelerator

- “Diversity of cloud workloads and … rapid … change” (weekly or monthly)
  - Search, SmartNIC, machine learning, encrypt, compress, and big data analytics
- Bing Search: 2X server level perf, 29% latency reduction, 10% increase in power
- Networking Virtualization: 10X latency improvement, 2X perf many db and OLTP workloads
- Machine Learning: Stratix 10 capable of 90 TFLOPs 8 bit floating point

Source: Microsoft

Microsoft Exa-op with FPGAs

(Ignite User Conference Sept. 2016)


translatordemox-dev.azurewebsites.net/#/scalability

Microsoft Translator V1.00.23400.11102

Data Source: Wikipedia

Translate to: Spanish

Wikipedia (English version)

Publisher: Wikimedia Foundation

Articles >5.2 million

Words: ~3.1 Billion

A free online encyclopedia that anyone can edit, and the largest and most popular general reference work on the Internet.

Processor Type

Azure FPGA Server – SV4-D5-1U

Type: 10 CPU cores + 4 FPGAs

Model: Stratix V D5-accelerator

Peak Power/Unit: 240 Watts

Compute Capacity

Compute Capacity: 1 Exa-op 1,000,000 Tera-ops

Estimated Time: 0.098 seconds

Pages Per Second: 78,120,000

Pages Translated: 0

"Translate every Wikipedia English page to another language in the blink of an eye"

Applications Acceleration:

Framework or APIs with OpenCL Underneath

INTEL® DEEP LEARNING DEPLOYMENT TOOLKIT

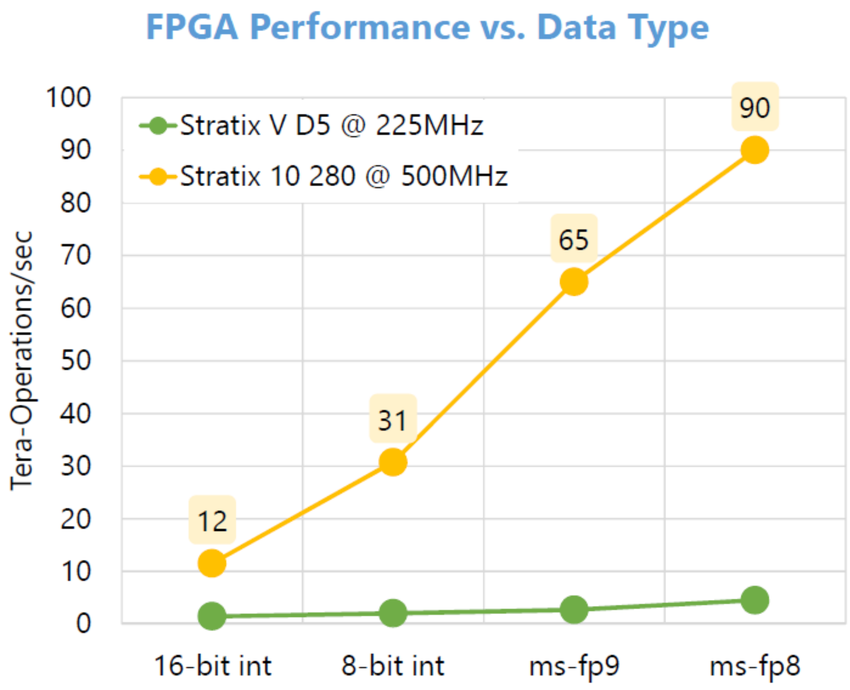
- With FPGA acceleration option
- Great FP8 performance

INTEL® SPARK FRAMEWORK WITH BIGDL

- FPGA acceleration POC

HPC: PROGRAMMER API

- Broad Institute GATK (PairHMM)
- Financial Library
- Government pattern matching
- Video transcode
- Emerging: oil & gas



Source: Microsoft

Infrastructure Acceleration

NETWORKING & DATA ACCESS

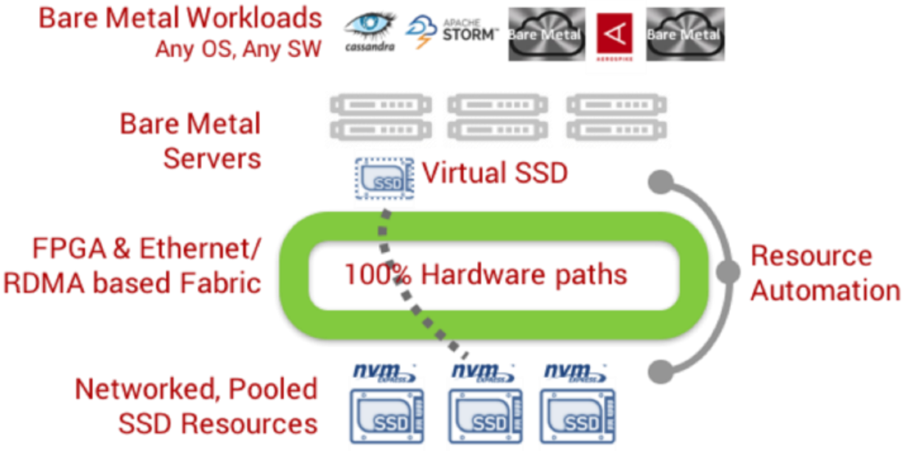
- In-line advantage over look-aside
- Compression, Encryption, Dedupe
- Virtualization or complete network stack

DATA ANALYTICS

- PostgreSQL, MariaDB, MySQL (Swarm64)
  - Data Warehouse
  - Real Time Analytics, 15M+ inserts/s
- Cassandra NoSQL (rENIAC)
- Hadoop/Spark (A3Cube)

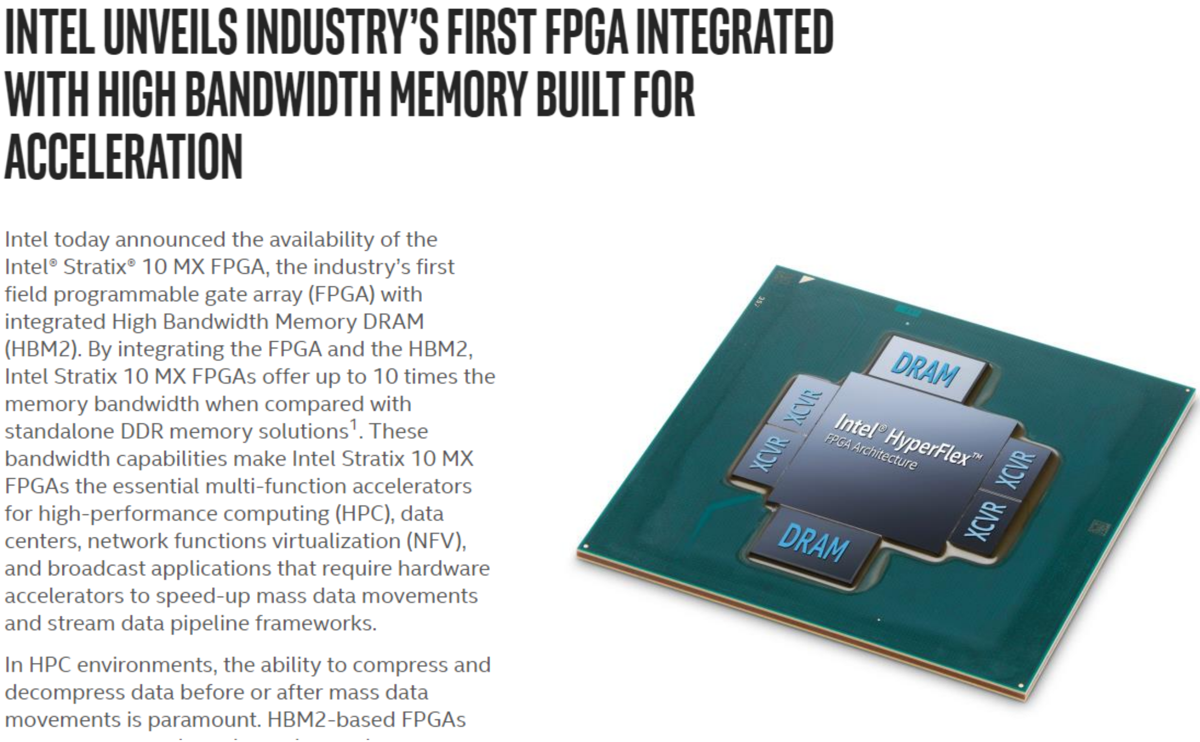
NVME OVER ROCE WITH ACCELERATORS

- Attala CPU offload & in-line acceleration



ATTALA SYSTEMS

Scale Up: Stratix 10 MX



INTEL UNVEILS INDUSTRY'S FIRST FPGA INTEGRATED WITH HIGH BANDWIDTH MEMORY BUILT FOR ACCELERATION

Intel today announced the availability of the Intel® Stratix® 10 MX FPGA, the industry's first field programmable gate array (FPGA) with integrated High Bandwidth Memory DRAM (HBM2). By integrating the FPGA and the HBM2, Intel Stratix 10 MX FPGAs offer up to 10 times the memory bandwidth when compared with standalone DDR memory solutions[1]. These bandwidth capabilities make Intel Stratix 10 MX FPGAs the essential multi-function accelerators for high-performance computing (HPC), data centers, network functions virtualization (NFV), and broadcast applications that require hardware accelerators to speed-up mass data movements and stream data pipeline frameworks.

In HPC environments, the ability to compress and decompress data before or after mass data movements is paramount. HBM2-based FPGAs

Scale Up: Falcon Mesa
Next Generation 10 NM FPGAs

CONTINUING PRODUCT LEADERSHIP

- Built on Intel Custom Foundry 10 nm platform
- 2nd Generation Intel® HyperFlex™ Architecture
- 2nd Generation EMIB-based heterogenous SiP
- Next Generation HBM Support
- Up to 112 Gbps Transceiver Rates
- PCI-Express Gen4 x16 Support

10nm FPGAs Built on
World's Most Advanced FinFET Process

Delivering Industry Leading Performance and Power

Scale Out: 3D-Torus w/OpenCL Option (Univ. of Florida)

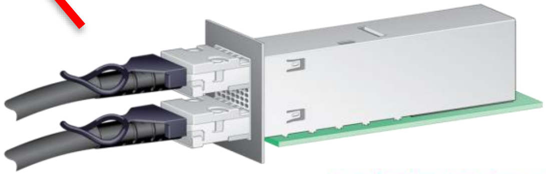
520N

High performance compute node with optical IO for creation of directly-coupled, dense FPGA clusters

- GPU/Phi form factor (3/4 length) - dual slot
- 16-lane PCIe Gen 3.0
- Intel Stratix 10 FPGA (GX2800 F1760 NF43)
- 4 QSFP28 Cages supporting 1G → 100G line rates
  - Upgrade to 6 100G network ports using QSFP-DD
  - Enables 3D-Torus network connectivity
- 4 banks of 8GB DDR4 SDRAM @ 2400MTPS
- On-board USB hub for system control and monitoring
- Board Support Packages (BSP) for Intel OpenCL SDK

OpenCL

intel



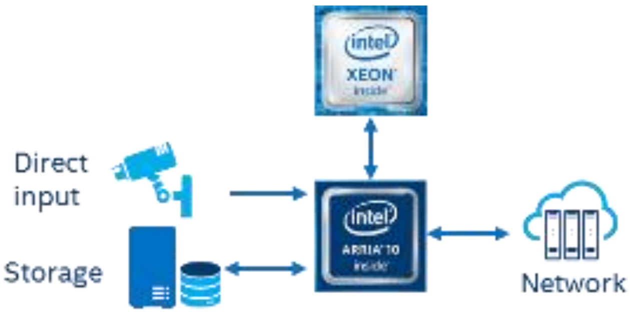
molex

QSFP-DD

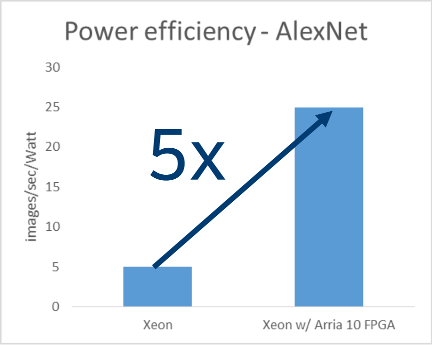
MACHINE LEARNING ACCELERATION

Intel FPGAs Offer Unique Value

High Throughput Deterministic Low Latency



Excellent Power Efficiency



Future Proof

- ✓ Current and future neural network topology
- ✓ Arbitrary precision data types (FloatP32 => FixedP2, sparsity, weight sharing)
- ✓ Inline and offload parallel processing; IO expansion
- ✓ More than 25 year silicon lifespans

Deep Learning Accelerator (DLA) Library

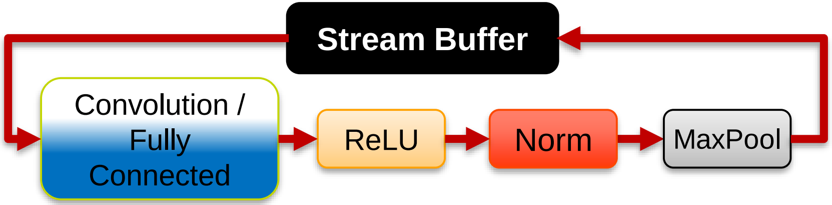
Implements common topologies in a graph loop architecture written in OpenCL™

Support for streaming or memory mapped data input

Support for various image sizes

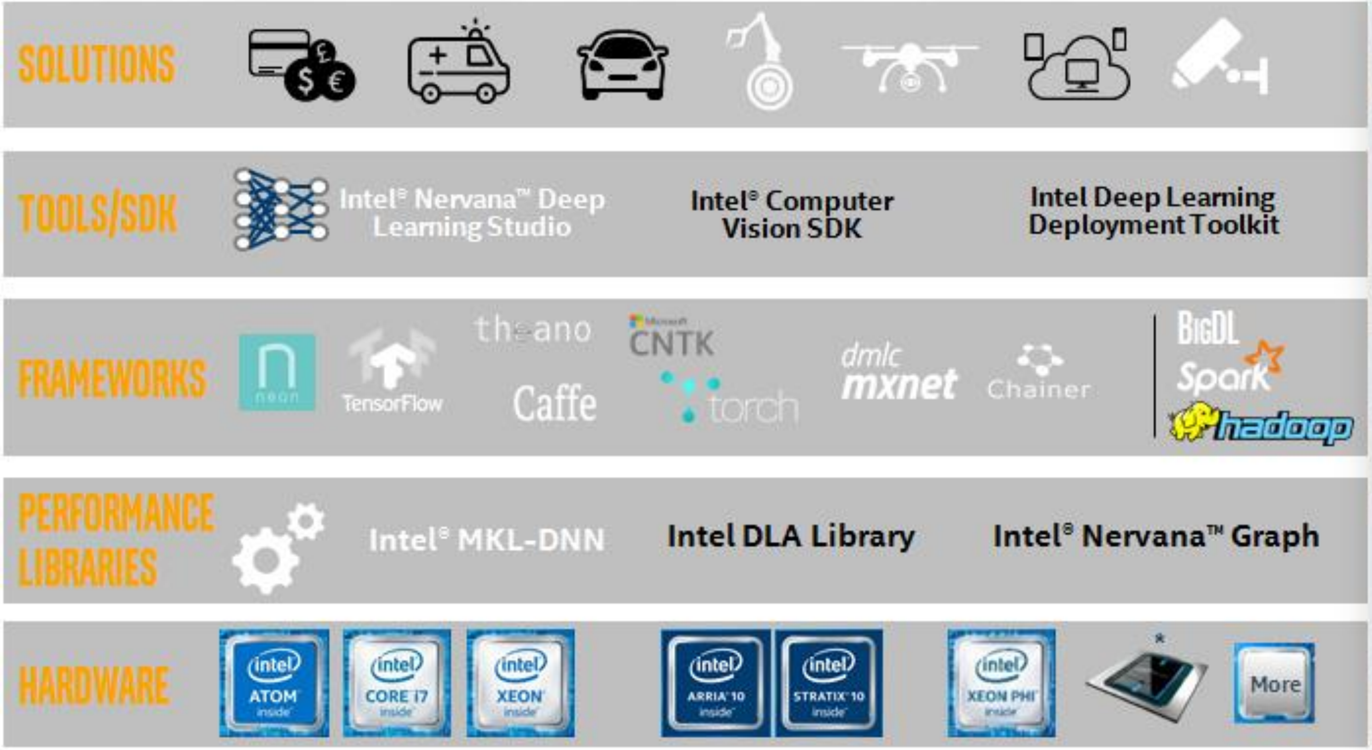
Static or dynamic architectures

- Static: Fixed architecture for maximum performance when reconfiguration not needed
- Dynamic: Run-time reconfigurable to different topologies
  - No FPGA compile required
- AlexNet, GoogleNet, LeNet, SqueezeNet, VGG16, ResNet, LSTM, SSD…



For more details, see our paper:
An OpenCL™ Deep Learning Accelerator on Arria 10 FPGA 2017

Intel® AI Ecosystem Now Enabled for FPGA



SOLUTIONS

TOOLS/SDK

Intel® Nervana™ Deep Learning Studio

Intel® Computer Vision SDK

Intel Deep Learning Deployment Toolkit

FRAMEWORKS

neon

TensorFlow

theano

Caffe

Microsoft CNTK

torch

dmlc mxnet

Chainer

BigDL

Spark

hadoop

PERFORMANCE LIBRARIES

Intel® MKL-DNN

Intel DLA Library

Intel® Nervana™ Graph

HARDWARE

intel ATOM inside

intel CORE i7 inside

intel XEON inside

intel ARRIA 10 inside

intel STRATIX 10 inside

intel XEON PHI inside

*

More

FPGA-ACCELERATED DATA ANALYTICS

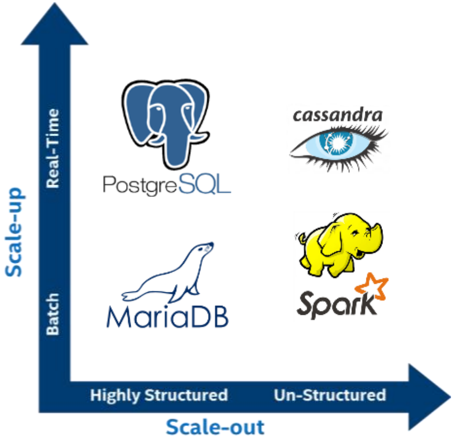
Accelerate Big Data Analytics with Existing Interfaces and FPGAs

1. **Intel Big Data Analytics Frameworks**
   Accelerate innovation in Big Data Analytics with frameworks built on Software Defined Infrastructure with open standard building blocks.

2. **Intel Frameworks & Libraries integrated with FPGAs**
   Run unmodified customer applications, use runtime orchestration with both Xeon ® and FPGA support, and leverage end to end virtualization & security.

3. **Accelerate Relational, NoSQL, and Un-Structured**

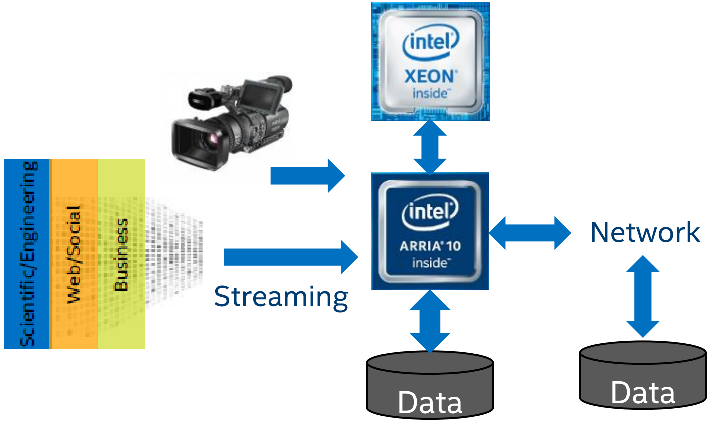
   FPGA data access, networking, and algorithm acceleration options with a single FPGA for highly structured, semi-structured, and un-structured data for better TCO, flexibility, and future proofing.



Analytics Landscape and Scaling

FPGAs Offer Unique Value for Analytics/Streaming

Single Multi-function Accelerator



Offloads algorithm, networking, and data access processing

Integrate to Intel Frameworks & APIs

- Run unmodified customer applications
- Orchestration run-time advantage: Xeon® or FPGA
- End-to-End Security & Virtualization framework

Moderate Acceleration is common

- PCIe lookaside acceleration (two data copies)

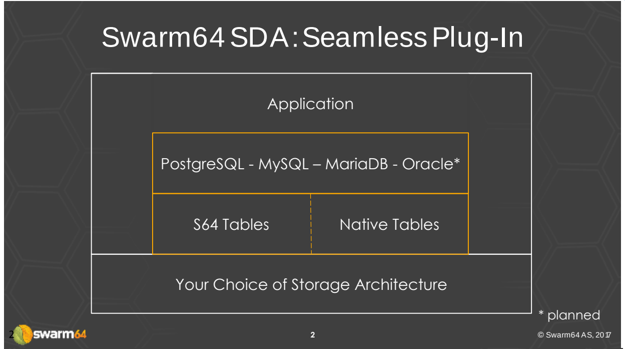
Significant Acceleration requires FPGA

- Multifunction and inline w/single FPGA
- Relational: 2X+ TPC-DS or TPC-H w/Swarm64
- NoSQL: 4X Cassandra w/rENIAC (80/20 R/W)
- Hadoop/Spark: 3X Terasort w/A3Cube (HDD)

Swarm64 Relational Database Acceleration

Two Workloads: Traditional Data Warehousing, Real Time Data Analytics

Database acceleration with a plugin



Swarm64 SDA: Seamless Plug-In

Application

PostgreSQL - MySQL – MariaDB - Oracle*

S64 Tables

Native Tables

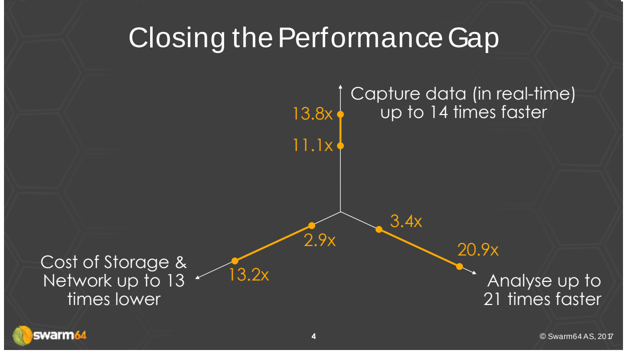
Your Choice of Storage Architecture

* planned

swarm64

2

© Swarm64 AS, 2017



Closing the Performance Gap

Capture data (in real-time) up to 14 times faster

13.8x

11.1x

Cost of Storage & Network up to 13 times lower

2.9x

13.2x

Analyse up to 21 times faster

3.4x

20.9x

swarm64

4

© Swarm64 AS, 2017

Acceleration Overview

- 10X+ single table inserts/s for real time data analytics
  - With modest tuning, 15M PostgreSQL INSERT/s*
  - 40%+ TCO savings over three years**
- 2X+ optimized queries for data warehousing
  - Using industry standard TPC-DS benchmark
- 3X+ storage compression
  - Data & tables managed by Swarm64

Note: this is SQL to relational d/b, not SQL to semi/unstructured data.

Note *: Dual Intel® Xeon® E5-2695 v4 processors, (8) 32GB DDR4-2400, (8) 512GB NVMe SSD.

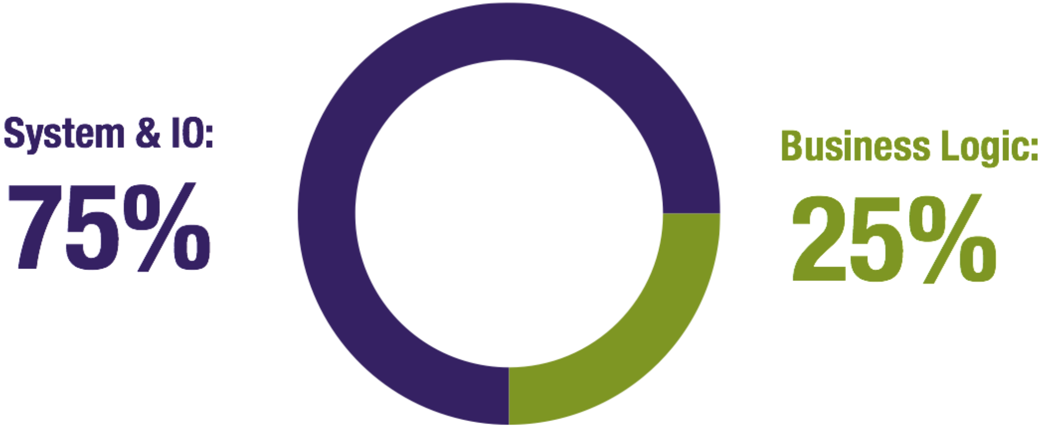
Note **: https://itpeernetwork.intel.com/data-center-application-acceleration-fpgas

Source: Swarm64

NoSQL: System & IO Acceleration Opportunity

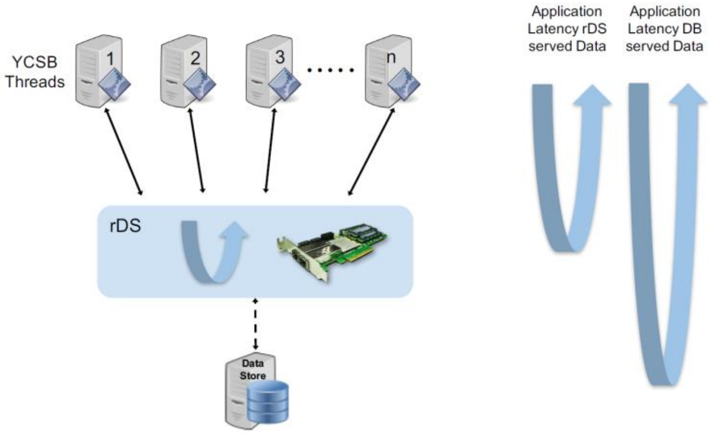
Source: rENIAC CEO

- Connection Management
- Compression/Encryption
- Book Keeping
- Data Encode/Decode



rENIAC Distributed Data Engine/Switch (rDS)

4X+ Cassandra acceleration (Source: rENIAC)


YCSB Threads
1
2
3
n
rDS
Data Store
Application Latency rDS served Data
Application Latency DB served Data

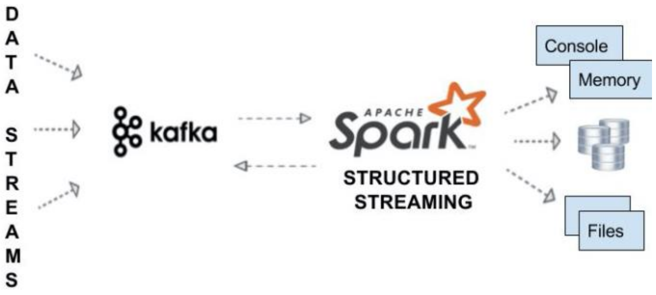
Overview

- No customer application change
- Plug-in card with 10GbE (Proxy tier or on database server)
- Distributed cache, proxy for reads and writes
- Predictable latency for SLAs
- Roadmap for storage compaction

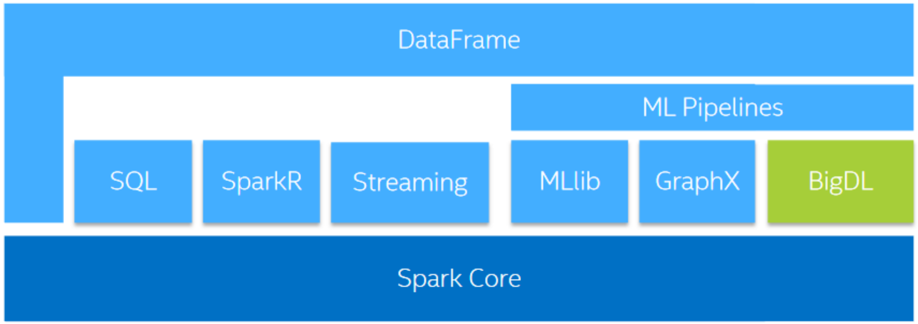
Significant Acceleration

- √ Networking/CQL acceleration
- √ Data access acceleration
- √ Compression
- √ Hashing

Spark: Five Acceleration Areas



BigDL: implemented as a standalone library on Spark (Spark package)

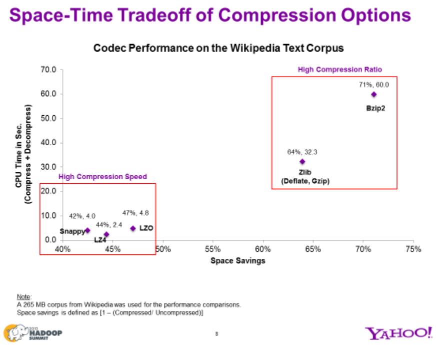


- Hadoop/Spark: Shuffle phase (by A3Cube)
- BigDL: Deep learning acceleration (by Intel - POC)
- Ingest/Kafka: Extract, transform, load and filtering (by BigStream)
- Machine learning MLlib: e.g. ALS (by Falcon Computing)
- SQL over SPARK POC (by BigStream)

Hadoop/Spark Shuffle Acceleration

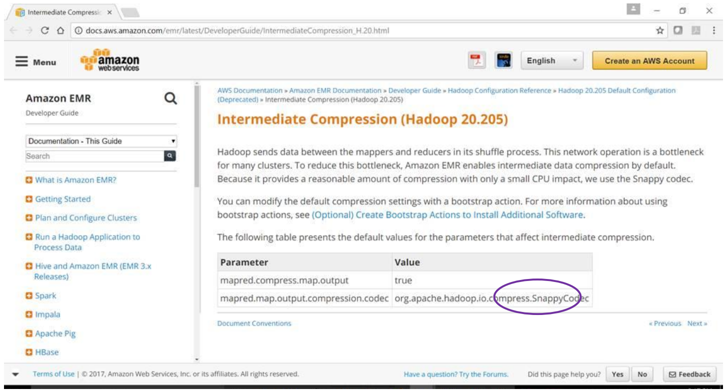
1.5X-3X TeraSort acceleration (Source: A3Cube)

- Baseline: snappy software compression
  - Zlib offers better compression
  - But too much cpu time & cycles
- Using hardware Zlib compression:
  - TeraSort speedup: up to ~1.3X[1] for disks
  - Against LZO software compression
- A3Cube acceleration versus snappy in s/w:
  - TeraSort speedup: ~1.5X SSD, ~3X disks
  - Networking acceleration
  - FPGA compression & file index lookup
  - No change to Hadoop/Spark application



Space-Time Tradeoff of Compression Options

Note:
A 265 MB corpus from Wikipedia was used for the performance comparisons.
Space savings is defined as [1 – (Compressed/ Uncompressed)]



Intermediate Compressi ×
docs.aws.amazon.com/emr/latest/DeveloperGuide/IntermediateCompression_H.20.html
Menu
amazon web services
English
Create an AWS Account
Amazon EMR
Developer Guide
Documentation - This Guide
Search
What is Amazon EMR?
Getting Started
Plan and Configure Clusters
Run a Hadoop Application to Process Data
Hive and Amazon EMR (EMR 3.x Releases)
Spark
Impala
Apache Pig
HBase

AWS Documentation » Amazon EMR Documentation » Developer Guide » Hadoop Configuration Reference » Hadoop 20.205 Default Configuration (Deprecated) » Intermediate Compression (Hadoop 20.205)

Intermediate Compression (Hadoop 20.205)

Hadoop sends data between the mappers and reducers in its shuffle process. This network operation is a bottleneck for many clusters. To reduce this bottleneck, Amazon EMR enables intermediate data compression by default. Because it provides a reasonable amount of compression with only a small CPU impact, we use the Snappy codec.

You can modify the default compression settings with a bootstrap action. For more information about using bootstrap actions, see (Optional) Create Bootstrap Actions to Install Additional Software.

The following table presents the default values for the parameters that affect intermediate compression.

Parameter	Value
mapred.compress.map.output	true
mapred.map.output.compression.codec	org.apache.hadoop.io.compress.SnappyCodec

[1] Note 1: https://www.exar.com/uploads/mwp-0002_a01_maximizinghadoopperfandstoragecapacitywithaltrahd.pdf

HPC ACCELERATION EXAMPLES

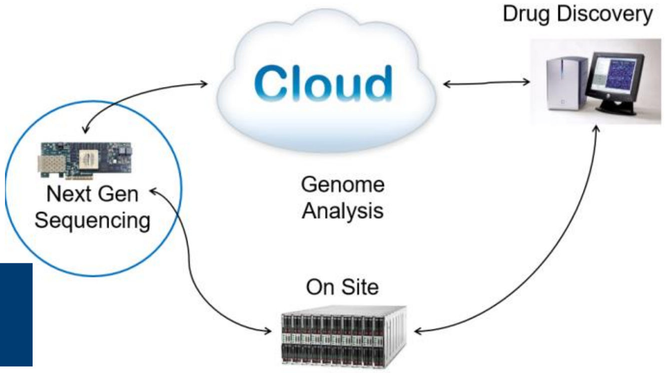
Genomics: GATK Acceleration

FPGA Acceleration in GATK

- Targets PairHMM full integration

Latest Intel Benchmark

Configuration	PairHMM	CPU Cores Used	Peak Perf (GCUPS)	Average Perf (GCUPS)
2 Socket Intel® Xeon® Processor E5 v4	AVX	1	0.699	0.676
2 Socket Intel® Xeon® Processor E5 v4	AVX	44	22.0	21.2
2 Socket Intel® Xeon® Processor E5 v4 + Intel® Arria® 10 FPGA	OpenCL	1	44.1	32.4

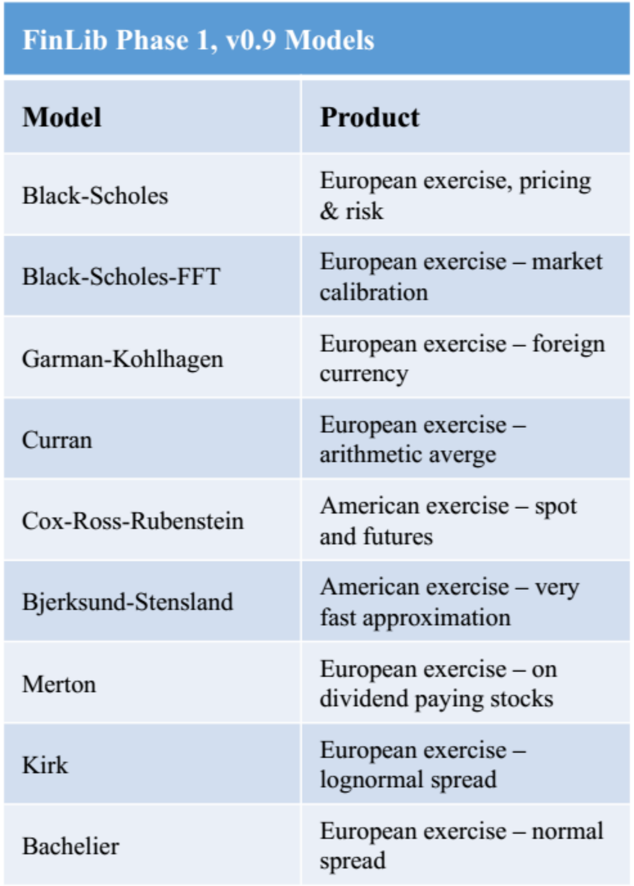


Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Financial Library (Intel)

Reference Design (Phase 1)

- Demo using the latest library is up and running on a dedicated server in Intel lab in Swindon UK, directly accessible to customers.
- Coverage is ~95% of exchange-traded options
- C++ and Python MKL api calls to an OpenCL kernel running on the FPGA
- FinLib can execute 3.2 billion option calculations/second using ~40% of an Arria10 1150 GX at 300MHz



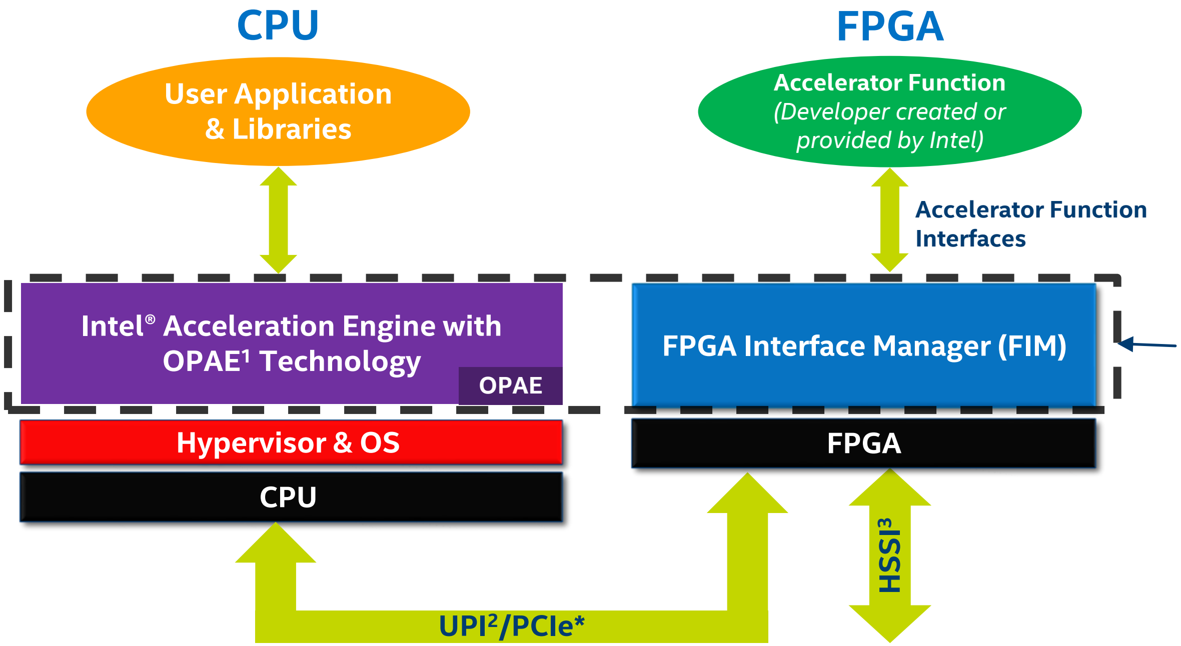
FinLib Phase 1, v0.9 Models

Model	Product
Black-Scholes	European exercise, pricing & risk
Black-Scholes-FFT	European exercise – market calibration
Garman-Kohlhagen	European exercise – foreign currency
Curran	European exercise – arithmetic averge
Cox-Ross-Rubenstein	American exercise – spot and futures
Bjerksund-Stensland	American exercise – very fast approximation
Merton	European exercise – on dividend paying stocks
Kirk	European exercise – lognormal spread
Bachelier	European exercise – normal spread

FPGA ACCELERATION PLATFORM AND CARD

ACCELERATION ENVIRONMENT

Common Developer Interface for Intel® FPGA Data Center Products



CPU
FPGA
User Application & Libraries
Accelerator Function
(Developer created or provided by Intel)
Accelerator Function Interfaces
Intel® Acceleration Engine with OPAE¹ Technology
OPAE
FPGA Interface Manager (FIM)
Hypervisor & OS
FPGA
CPU
HSSI³
UPI²/PCIe*

Intel® Acceleration Engine with OPAE[1] Technology

FPGA Interface Manager (FIM)

Optimized and simplified hardware and software APIs provided by Intel

[1] OPAE = Open Programmable Acceleration Engine

[2] UPI = Intel® Ultra Path Interconnect

[3] HSSI = High Speed Serial Interface

Some names pending final approval and may change in the future.

Supports: Red Hat Enterprise Linux* 7.3 w/ kernel 4.7, Intel® Xeon® Processors v4 or newer

OPEN PROGRAMMABLE ACCELERATION ENGINE (OPAE) TECHNOLOGY

Simplified FPGA Programming Model for Application Developers

Consistent API across product generations and platforms

- Abstraction for hardware-specific FPGA resource details

Designed for minimal software overhead and latency

- Lightweight user-space library *(libfpga)*

Open ecosystem for industry and developer community

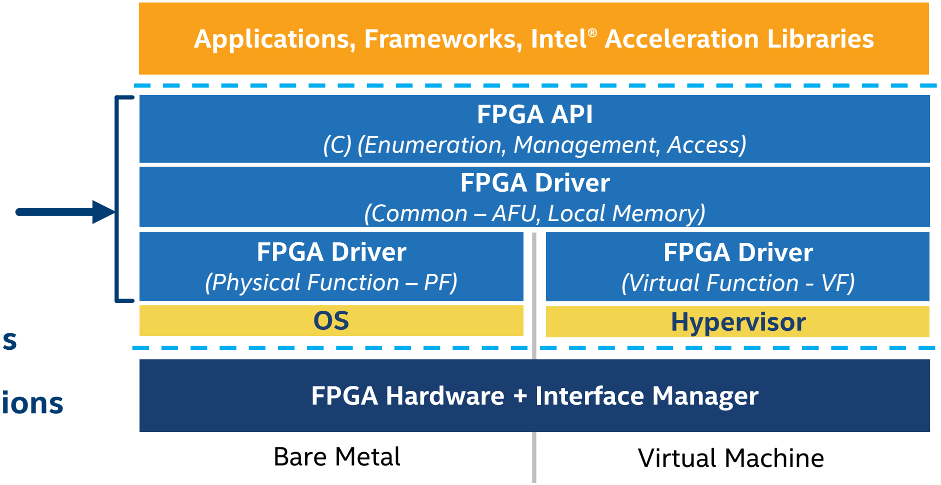
- License: FPGA API (BSD), FPGA driver (GPLv2)

FPGA driver being upstreamed into Linux* kernel

Supports both virtual machines and bare metal platforms

Faster development and debugging of Accelerator Functions with the included AFU Simulation Environment (ASE)**

Includes guides, command-line utilities and sample code



Applications, Frameworks, Intel® Acceleration Libraries
FPGA API
(C) (Enumeration, Management, Access)
FPGA Driver
(Common – AFU, Local Memory)
FPGA Driver
(Physical Function – PF)
FPGA Driver
(Virtual Function - VF)
OS
Hypervisor
FPGA Hardware + Interface Manager
Bare Metal
Virtual Machine

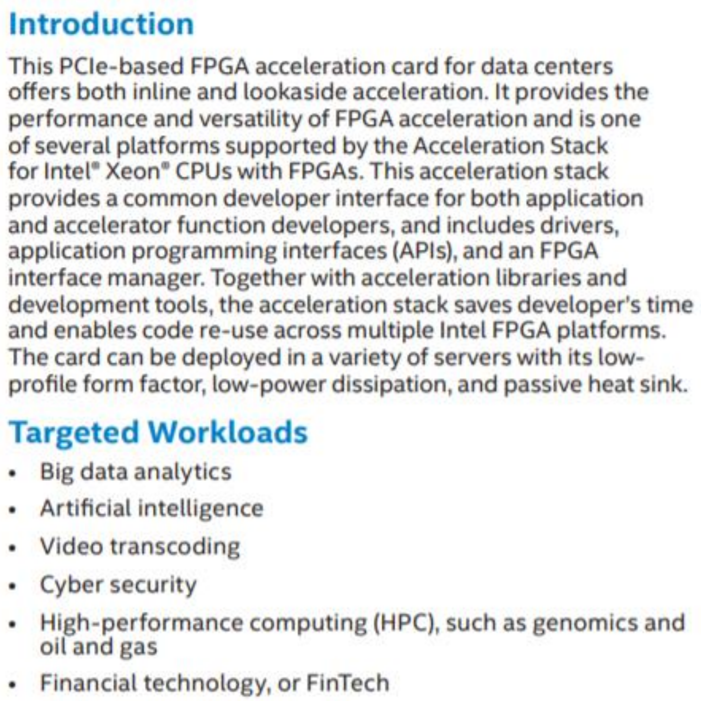
Start developing for Intel® FPGAs with OPAE today: http://github.com/OPAE

***ASE requires Acceleration Functions written in RTL and a properly installed RTL simulator: Synopsys* VCS-MX, Mentor Graphics* ModelSim-SE*/QuestaSim*

Some names pending final approval and may change in the future.

Supports: Red Hat Enterprise Linux* 7.3 w/ kernel 4.7, Intel® Xeon® Processors v4 or newer

INTEL® PROGRAMMABLE ACCELERATION CARD WITH INTEL ARRIA® 10 FPGA (SAMPLING NOW, PRODUCTION 2Q18)

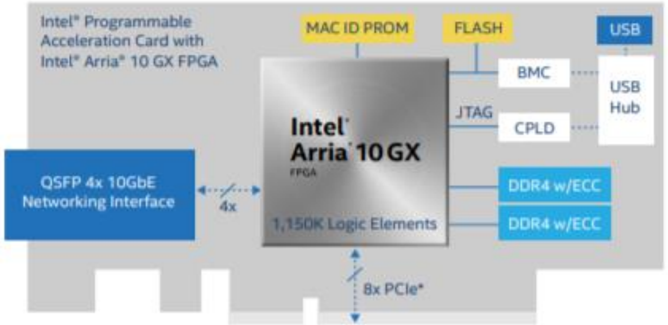


Introduction

This PCIe-based FPGA acceleration card for data centers offers both inline and lookaside acceleration. It provides the performance and versatility of FPGA acceleration and is one of several platforms supported by the Acceleration Stack for Intel® Xeon® CPUs with FPGAs. This acceleration stack provides a common developer interface for both application and accelerator function developers, and includes drivers, application programming interfaces (APIs), and an FPGA interface manager. Together with acceleration libraries and development tools, the acceleration stack saves developer's time and enables code re-use across multiple Intel FPGA platforms. The card can be deployed in a variety of servers with its low-profile form factor, low-power dissipation, and passive heat sink.

Targeted Workloads

- Big data analytics
- Artificial intelligence
- Video transcoding
- Cyber security
- High-performance computing (HPC), such as genomics and oil and gas
- Financial technology, or FinTech



Summary

- Intel has comprehensive standards-based frameworks and APIs for ML & data analytics
- Run deep learning on FPGAs with the same framework used for Xeons
  - Stratix 10 FPGA sampling with 80-90 TFLOPs reduced precision FP8 floating point
- Customers can run analytics workloads without change on these frameworks and Intel® APIs with FPGAs underneath
- A single FPGA per server can deliver multifunction acceleration for one or more workloads
- Broad and developing data analytics ecosystem partner solutions and POCs
- Intel branded PCIe low profile FPGA card in production in 2Q18

intel®

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at www.intel.com.

Tests measure performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

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

Intel, the Intel logo, Intel Inside, the Intel Inside logo, Xeon, Arria, Stratix, and Intel Xeon Phi are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

Microsoft White Papers

- A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services
  - https://www.microsoft.com/en-us/research/publication/a-reconfigurable-fabric-for-accelerating-large-scale-datacenter-services/
- Microsoft's Production Configurable Cloud (Mark Russinovich)
  - https://www.slideshare.net/ChrisGenazzio/microsofts-configurable-cloud
- Accelerating Persistent Neural Networks at Datacenter Scale
  - https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.60-NeuralNet1-Pub/HC29.22,622-Brainwave-Datacenter-Chung-Microsoft-2017_08_11_2017.pdf