RECONFIGURABLE ACCELERATOR PLATFORM WITH FPGAS

Mike Strickland, Director, Data Center Solution Architect
Intel® Corporation
November 2017
AGENDA

- Microsoft FPGA Acceleration Trend
- Data Analytics Acceleration
- Machine Learning Acceleration
- HPC Acceleration Examples
- FPGA Acceleration Platform and Card
- Summary
**MULTI-FUNCTION ACCELERATION: ALGORITHM + NETWORKING + DATA ACCESS**

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
- Lower and predictable FPGA latency for ranking vs. software
- Excellent FPGA inferencing with reduced precision floating point (FP8)
  - Intel® Stratix® 10 10SG280 FPGA has 90 Teraflops* (MSFT Hot Chips presentation (Oct 2017))

Note *: eight bit reduced precision floating point
FPGA-ACCELERATED DATA ANALYTICS
ACCELERATE BIG DATA ANALYTICS WITH INTEL FRAMEWORKS AND LIBRARIES WITH FPGAS

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

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

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.
FPGAs Offer Unique Value for Analytics or Streaming

Multi-function Accelerator

Integrate to Intel Frameworks and APIs
- Run unmodified customer applications
- Orchestration run time advantage: Intel® Xeon® processors or Intel FPGAs
- End-to-end security and virtualization framework

Significant Acceleration
- PCIe lookaside acceleration
- Networking + streaming + data access
  - Inline acceleration and protocol acceleration
- Compression, filtering, encryption
- Fast lookups/hashing
**DIFFERENT DATA STORE APPROACHES**

1. **Structured Data/Relational**
   - Traditional relational database for online transaction processing (OLTP), business intelligence, and online analytical processing (OLAP)

2. **Semi-structured Data/NoSQL**
   - NoSQL databases run on multiple servers with no single point of failure
   - Key-value store (KVS), column-oriented, field, range queries, regex, and hybrid
   - Hadoop designed to work on inexpensive hardware with redundancy
   - Hadoop usually with unstructured HDFS flat files; Spark resilient distributed datasets (RDD) in memory is faster

3. **Unstructured Data**
   - cassandra
   - mongoDB
ECOSYSTEM PARTNER SOLUTIONS
USING INTEL® FPGAS FOR DATA ANALYTICS

SQL over relational open databases
  • Traditional data warehousing acceleration
  • Real-time analytics acceleration

NoSQL
  • Memcached/KVS, Cassandra

Hadoop, Spark
  • Accelerate aggregation or “shuffle” phase
  • Data Ingest: FPGA does “inline” offload of extract, transform, and load (ETL)
  • Accelerate Deep Learning (BigDL)
  • Spark MLlib unstructured machine learning APIs – e.g. K-means, SVM, ...
**SWARM64 RELATIONAL DATABASE ACCELERATION**

**TWO WORKLOADS: TRADITIONAL DATA WAREHOUSING, REAL TIME DATA ANALYTICS**

Database accelerate with a plugin

**Acceleration Overview**

- 10X+ single table inserts/s for real time data analytics
  - With modest tuning, 15M PostgreSQL INSERT/s*
- 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.

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](http://www.intel.com/performance). Copyright © 2017, Intel Corporation

Source: Swarm64
SWARM64 RELATIONAL DATABASE ACCELERATION
SCALE UP DATA WAREHOUSING, REAL TIME DATA ANALYTICS, AND STORAGE COMPRESSION

Database accelerate with a plugin

Overview
- No customer application change
  - Storage engine plugin: PostgreSQL, MySQL, ...
- Query Engine accelerate INSERT, SELECT, ...
- Optimized indexing
- More IO bandwidth, mem depth from compression

Significant Acceleration
- Data access acceleration
- Compression, filtering, replication ...
- Memory mapped acceleration, data cache
- “Optimized Columns” indexing
NOSQL: SYSTEM & IO ACCELERATION OPPORTUNITY

[SOURCE: RENIAC]

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

System & IO: 75%

Business Logic: 25%
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

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

Note 1: https://www.exar.com/uploads/mwp-0002_a01_maximizinghadooppersandstoragecapacitywithultrahd.pdf
The only true in-line acceleration of big data or machine learning using Intel® FPGAs

- Up to 10X using Intel® Arria® 10 and Intel® Stratix® 10 FPGAs
  - Zero code changes
  - Cross platform: Spark*, Kafka*, TensorFlow* -- cloud or on-prem
- Intelligent and adaptive
  - Automatic partitioning of computation between CPU and FPGA
- Spark* SQL TPC-DS results
- Industry targets: FinServ/FinTech, AdTech, Healthcare
- Use cases: Spark* SQL analytics, ingest/ETL, EDW

Source: Bigstream
MACHINE LEARNING ACCELERATION
Step 1: Training
(In Data Center – Over Hours/Days/Weeks)

- Massive data sets: labeled or tagged input data
- Create "Deep neural net" math model
- Output Classification

Step 2: Inference
(At the Edge or in the Data Center - Instantaneous)

- New input from camera and sensors
- Trained neural network model
- Output Classification

FPGA Focus
FPGA MACHINE LEARNING ADVANTAGES

- 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

Image: Power efficiency - AlexNet

Graph showing a 5x improvement in images/sec/Watt between Xeon and Xeon w/ Arria 10 FPGA.
FPGAs control the data path.

- **CPU**
- **I/O**
- **FPGA**
- **Storage**

**Inline Data Flow Processing**
- Machine learning
- Object detection and recognition
- Advanced driver assistance system (ADAS)
- Gesture recognition
- Face detection

**Storage Acceleration**
- Machine learning
- Cryptography
- Compression
- Indexing

**Compute Acceleration/Offload**
- Workload agnostic compute
- FPGAaaS
- Virtualization
DEEP LEARNING ACCELERATOR (DLA) LIBRARY

- Common topologies in a graph loop architecture
- Support for streaming or memory mapped data input
- Static: Fixed architecture for maximum performance when reconfiguration not needed
- Dynamic: Run-time reconfigurable to different topologies without FPGA compile
- 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® DEEP LEARNING DEPLOYMENT TOOLKIT AND FPGA DEEP LEARNING ACCELERATOR (DLA) ARCHITECTURE

ML Framework (Caffe, TensorFlow)  Software Application

DLA Runtime Engine
- DLA Graph Compiler
- DLA Memory Manager

OpenCL Runtime and Drivers

DLA Architecture
- Compute Primitives
- Foundational Primitives

OpenCL BSP

Intel® Xeon®
FPGA FLEXIBILITY SUPPORTS ARBITRARY ARCHITECTURES

Many efforts to improve efficiency

- Batching
- Reduce bit width
- Sparse weights
- Sparse activations
- Weight sharing
- Compact network

\[ I \xrightarrow{W} O \]


- BinaryConnect [NIPS'15]
- TernaryConnect [ICLR'16]
- Spatially SparseCNN [CIFAR-10 winner '14]
- Pruning [NIPS'15]
- SparseCNN [CVPR'15]
- DeepComp [ICLR'16]
- HashedNets [ICML'15]
- SqueezeNet

Shared Weights

\[ 3 \]

\[ 2 \]
NARROW PRECISION INFERENCES ON FPGAS

FPGA Performance vs. Data Type

Impact of Narrow Precision on Accuracy

Source: Microsoft

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. Copyright © 2017, Intel Corporation
# Intel® AI Ecosystem Now Enabled for FPGA

## Experiences
- **Intel® Nervana™ Cloud & Appliance**
- **Intel® Nervana™ DL Studio**
- **Intel® Computer Vision SDK**
- **Intel® Movidius™ MDK**
- **Intel® Saffron™**

## Platforms
- **Intel® Nervana™ Cloud & Appliance**
- **Intel® Nervana™ DL Studio**
- **Intel® Computer Vision SDK**
- **Intel® Movidius™ MDK**
- **Intel® Saffron™**

## Frameworks
- **Spark**
- **BigDL**
- **TensorFlow**
- **MLE**
- **Neon**
- **MXNet**
- **Theano**
- **Caffe**
- **Chainer**
- **Keras**

## Libraries
- **Intel® Data Analytics Acceleration Library (DAAL)**
- **Intel® Nervana™ Graph**
- **Intel® Math Kernel Library (MKL, MKL-DNN)**
- **Intel Python Distribution**

## Hardware
- **Compute**
- **Memory & Storage**
- **Networking**

*Future
Other names and brands may be claimed as the property of others.*
OTHER HPC EXAMPLES
FPGA Acceleration in GATK

- Targets PairHMM full integration

## Latest Intel Benchmark

<table>
<thead>
<tr>
<th>Configuration</th>
<th>PairHMM</th>
<th>CPU Cores Used</th>
<th>Peak Perf (GCUPS)</th>
<th>Average Perf (GCUPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Socket Intel® Xeon® Processor E5 v4</td>
<td>AVX</td>
<td>1</td>
<td>0.699</td>
<td>0.676</td>
</tr>
<tr>
<td>2 Socket Intel® Xeon® Processor E5 v4</td>
<td>AVX</td>
<td>44</td>
<td>22.0</td>
<td>21.2</td>
</tr>
<tr>
<td>2 Socket Intel® Xeon® Processor E5 v4 + Intel® Arria® 10 FPGA</td>
<td>OpenCL</td>
<td>1</td>
<td>44.1</td>
<td>32.4</td>
</tr>
</tbody>
</table>

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](http://www.intel.com/performance). Copyright © 2017, Intel Corporation.
Ease of Use

- C++ and Python MKL api calls to an OpenCL kernel running on the FPGA
- FinLib Phase 1: ~ 95% of exchange-traded options

Performance

- FinLib can execute 3.2 billion option calculations/second using ~40% of an Arria10 1150 GX at 300MHz

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. Copyright © 2017, Intel Corporation
FPGA ACCELERATION PLATFORM AND CARD
ACCELERATION STACK FOR INTEL XEON® CPU WITH FPGAS

- Fast, hot-swapping of accelerator functions
- Accessible from virtual machines and containers
- Support for leading cloud orchestrators

Visibility and Access Levels:

- Orchestration / Rack Level Management
- User Applications
- Frameworks & SDKs
- Acceleration Libraries

Development Tools and Accelerator Functions:

- Platform Development Tools
- Open Programmable Acceleration Engine (OPAE)

Intel® Xeon® Processor with FPGA Platforms

Intel® FPGA Interface Manager (FIM)
ACCELERATION ENVIRONMENT
Common Developer Interface for Intel® FPGA Data Center Products

Intel® Acceleration Engine with OPAE Technology
FPGA Interface Manager (FIM)
Optimized and simplified hardware and software APIs provided by Intel

1OPAE = Open Programmable Acceleration Engine
2UPI = Intel® Ultra Path Interconnect
3HSSI = High Speed Serial Interface

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

Some names pending final approval and may change in the future.
INTEL® PROGRAMMABLE ACCELERATION CARD WITH INTEL® ARRIA® 10 FPGA
SAMPLING NOW, PRODUCTION 1H18

INTEL'S 1ST VERSATILE FPGA PCIe* ACCELERATION CARD
THAT OFFERS INLINE AND LOOK-ASIDE ACCELERATION FOR WORKLOADS REQUIRING UP TO 45W
SUMMARY

- Customers can run analytics workloads without change
  - on industry standard frameworks and Intel® APIs with FPGAs underneath
  - up to 3X server level acceleration for relational, NoSQL, and Hadoop/Spark
- Reduced precision floating point breakthrough for deep learning
  - Intel® Stratix® 10 FPGAs: 80+ FP8 TFLOPs
- Bioinformatics integrated to GATK standard and Financial integrated to Intel® MKL
- Intel branded PCIe low profile FPGA card in production in 1Q18
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, and Arria are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

© Intel Corporation