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.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks.

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/benchmarks.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate. Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

© 2018 Intel Corporation.
AGENDA

- Intro: Scale Out and Scale Up with FPGA
- Programmer Tool Chain
  - Financial Library API Example
- Libraries & Turnkey Solutions
MICROSOFT EXA-OP WITH FPGAS
(IGNITE SEPT 2016)

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

Link to video (~ minute 55)
Microsoft Scale Out FPGA Multi-Function Accelerator

- “Diversity of cloud workloads and ... rapid ... change” (weekly or monthly)
  - Search, SmartNIC, Machine Learning, Encrypt, Compress, Big Data Analytics,...
- Bing Search: 2X server level perf, 29% latency reduction, 10% increase in power (1)
- Networking Virtualization: 10X latency improvement, 2X perf many db and OLTP workloads (2)
- Machine Learning: Stratix 10 capable of 90 TFLOPs 8 bit floating point (3)
WHAT FPGA DOES WELL

- Custom Processing Pipelines
- Variable Precision Arithmetic
- Heterogenous Dataflows
- Diverse Memory Hierarchy

- Multiple Workloads
- High-bandwidth data caching
- Look-aside AND In-line Acceleration
- Parallel processing
Applications Acceleration Overview
Framework or API’s with OpenCL underneath option

**Intel® OpenVINO™**
- With FPGA acceleration option

**Data Analytics**
- Open Relational D/B
  - Data Warehouse, Real Time
- Cassandra NoSQL, ElasticSearch
- Hadoop/Spark

**HPC: Programmer API**
- Genomics, Financial
- Government pattern matching
- Video transcode
- Emerging: oil & gas
Infrastructure Acceleration Overview

**Networking & Data Access**
- In-line advantage over look-aside
- Compression, Encryption, Dedupe
- Virtualization or complete network stack
- Torus & inline acceleration

**NVMe over RoCE with Accelerators**
- Attala cpu offload & inline acceleration
SCALE OUT: NOVO-G 3D TORUS

Developed and deployed at CHREC
- Most powerful reconfigurable computer in research community (2009-present)
  - 448 (soon 512) high-end Altera FPGAs with 3.5TB (soon 4.5TB) of FPGA-attached SDRAM
- 2012 Alexander Schwarzkopf Prize for Technology Innovation @ NSF

App acceleration
- Computer vision, finance, bioinformatics, molecular dynamics, crypto, et al.

Hardware emulation
- Behavioral emulation of future apps and systems, up to Exascale

2015 - 2016
- 64 Altera Stratix-V D8 FPGAs
  - On Gidel ProcE-V PCIe boards
  - Additional 64 in development
- 3D torus interconnect (4x4x4)
  - 6 links per Stratix-V (40 Gb/s per link)

2009: 96 Altera Stratix-III E260 FPGAs, each with 4.25GB SDRAM, on Gidel ProcStar-III cards
2010: 96 Altera Stratix-III E260 FPGAs, each with 4.25GB SDRAM, on Gidel ProcStar-III cards
2011: 96 Altera Stratix-IV E530 FPGAs, each with 8.50GB SDRAM, on Gidel ProcStar-IV cards
2012: 96 Altera Stratix-IV E530 FPGAs, each with 8.50GB SDRAM, on Gidel ProcStar-IV cards
2014: 32 Altera Stratix-V D8 FPGAs, each with 16GB SDRAM and 3D torus i/f, on Gidel ProcE-V card
2015: 32 Altera Stratix-V D8 FPGAs, each with 16GB SDRAM and 3D torus i/f, on Gidel ProcE-V card

3D torus interconnect

Made possible by support from UF, Altera, Gidel, NSF, and DOE
**520N PCIe FPGA Board**

**Intel Stratix 10 FPGA**
- GX2800 F1760 NF43
- -1 or -2 SerDes and FPGA speed grades

**PCIe ¾ Length, Dual Width**
- Gen3 x16, standard height, ¾ length

**Four 100G QSFPs**
- Range of line rates include 4x 40/100G or 16x 10/25G

**4x DDR4 Banks**
- 2400MT/s, up to 32GB total

**OpenCL HPC BSP**

---

**SCALE OUT: STRATIX 10 2D TORUS**

- On-board System Manager (USB and FPGA access)
- On-board USB Blaster II and UART-to-USB
- FPGA configuration functions incl. JTAG, remote flash upgrade, reconfiguration, and fallback mechanism
Scale Your Innovation

**Scale Out: 3D-Torus Roadmap Concept**

**520N**

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

- **(4) QSFP28 Cages supporting 1G → 100G line rates**
  - Upgrade to **(6) 100G network ports using QSFP-DD**
  - Enables 3D-Torus network connectivity
INTEL® EMBEDDED MULTI-DIE INTERCONNECT BRIDGE (EMIB) TECHNOLOGY

Intel® Stratix® 10 FPGAs and SoCs with Intel EMIB

- **Intel® Stratix® 10**
- **Intel HyperFlex**
- **EMIB**
- **Flip-Chip Pitch > 100μm**
- **Microbump pitch 55mm**
Development Kit Contents

- Development board
  - Stratix® 10 MX FPGA, 2.1M LE, 8GB HBM2
- 2x QSFP28 cages
- 2GB DDR4 onboard
- Hi-Lo and DIMM Connectors for DDR4
- PCIe Gen3x16
  - Endpoint via edge connector
  - Rootport via slot

Intel.com page: [link]
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

Delivering Industry Leading Performance and Power
PROGRAMMER TOOL CHAIN
Removing the Barriers of Adoption

Hardware Developers

- Intel® Quartus® Prime Design Suite
- DCP Platform / Acceleration Framework

Software Developers

- Parallel Compilers
- HLS Compiler
- LLVM Compiler
- DSP Builder
- High-Level Design Backend Compiler

Libraries

- Primitives
- DLA
- OpenVINO™

Boost Productivity

Boost Performance

 SCALE YOUR INNOVATION
COMMON HETEROGENOUS PROGRAMMING ENVIRONMENT

Common environment for heterogenous programming

- Plugin to Intel® System Studio and Intel Parallel Studio
- CPU, GPU, FPGA, ...

Easy path to FPGA

- Already familiar environment
- Intel® Developer Zone
Software Programmers
Need Logic and Data Management
- By writing lines of code

OpenCL™ Compiler Benefits
- Ease of use
- Scalable
- Heterogeneous
- Leverage existing libraries
- Vendor choice w/open standards
- Foundation for HLS & SPIR

Channels/Pipe Extension
- Kernel → Kernel
- External IO → Kernel
- Mix ‘n Match HDL & Kernels

Evaluated by
CPU

FPGA

Compile Code
Create Data
and Arguments
Execute

Context

DDRx Global Memory Buffer

I/O → Kernel → Kernel → Kernel → I/O

Scale Your Innovation
LEVERAGING SOFTWARE DEVELOPMENT ENVIRONMENT

1. Modify kernel.cl
2. x86 Emulator (sec)
3. Optimization Report (sec)
4. Profiler (hours)

DONE!

Hardware performance met?
Functional Bugs?
Stall-free pipeline? Memory coalesced?
API CALL + OPENCL OPTION

FinLib Example
- High level C++ APIs
- OpenCL implementation
- “Quants” can use high level APIs
- Similar approach used for PairHMM
FPGA IN COMPUTING MAINSTREAM
INGREDIENTS NEEDED TO MAKE FPGA IN COMPUTING “MAINSTREAM”

- User Application
- Data Framework (e.g. Apache Spark*)
- Scalable Functions (e.g. PDE Solver)
- Library Primitives (Mathematics, Statistics)
- Developer SDKs (e.g. OpenCL™)
- Acceleration Stack, Drivers, BSPs & Interface IP
- Boards and Platforms
- FPGA Silicon & FPGA Design Tools

Ecosystem Partners & Integrators

OEM Qualification
ACCELERATION STACK INGREDIENTS: OVERVIEW

- Application
- Libraries
- Drivers
- Software
- Accelerator Functional Unit (AFU)
- Signal Bridge and Management
- FPGA Platforms (Programmable Acceleration Cards)
- FPGA Hardware
- User, Intel, and 3rd Party (Tuning Expert)
- Open Programmable Acceleration Engine (OPAE) Provided by Intel
- User, Intel, or 3rd-Party IP Plugs into AFU Slot (tuning Expert)
- Qualified and Validated for volume deployment Provided by OEMs

Developed by User (Domain Expert)

PCle* Drivers Provided by Intel

FPGA Interface Manager Provided by Intel

Provided by Intel

Intel® Xeon®

User, Intel, or 3rd Party IP

Software

Libraries

Drivers

Application

Intel® Xeon®

FPGA

Hardware

FPGA Platforms (Programmable Acceleration Cards)

User, Intel, or 3rd Party IP

Plugs into AFU Slot (tuning Expert)

Qualified and Validated for volume deployment Provided by OEMs

User, Intel, and 3rd Party (Tuning Expert)

Open Programmable Acceleration Engine (OPAE) Provided by Intel

User, Intel, or 3rd Party IP

Plugs into AFU Slot (tuning Expert)

Qualified and Validated for volume deployment Provided by OEMs

User, Intel, and 3rd Party (Tuning Expert)
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

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

FPGA Hardware + Interface Manager
FPGA Driver
(Physical Function – PF)
FPGA Driver
(Common – AFU, Local Memory)
FPGA Driver
(Virtual Function – VF)
OS
Hypervisor
Bare Metal
Virtual Machine

Applications, Frameworks, Intel® Acceleration Libraries

**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 GX FPGA (AKA “RUSH CREEK”)

Arria® 10 GX FPGA [10AX115N2F40E2LG]
High-perf, multi-gigabit SerDes transceivers up to 15 Gbps
1150K logic elements available (-2L speed grade)
53 Mb of embedded memory

On-board Memory
8 Gbytes DDR4 Memory Banks with ECC (2 banks), 2400 Mbps
1Gb Mbit (128 MB) Flash

Interfaces & Dimensions
PCIe x8 Gen3 electrical, x16 mechanical *
USB 2.0 interface for debug and prog FPGA and Flash
1x QSFP with 4x 10GbE or 40GbE support
½ Length, ½ Height, 1RU

Software
Acceleration Stack for Intel® Xeon® CPU with FPGAs
FPGA Interface Manager Installed

Board Management Controller
Voltage, current, temperature monitoring
Power sequencing and reset
Platform Level Data Model (PLDM)

Power
70W TDP, 45W FPGA
650 LFM at Tja 55°C – Passively Cooled

Specifications preliminary and are subject to change
INTEL® PROGRAMMABLE ACCELERATION CARD
WITH INTEL® STRATIX® 10 FPGA GX (AKA “DARBY CREEK”

Versatile Workload Acceleration
- Customizable Hardware Architecture using Intel® Stratix® 10 FPGA GX

High Performance with Intel® Stratix® 10 FPGA GX
- 2.8M logic elements available with 229Mb of embedded memory
- 32GB DDR4 Memory with ECC (4 banks), 2400 Mbps

High Data Ingestion and Lower Latency
- PCIe* x16 v3 with SRIOV support
- 2x QSFP with 100GbE support

PCIe* Form Factor Compliant
- Dual slot, three-fourths length, full height
- 225W TDP – Passively Cooled
- Intel Max® 10 based Board Management Controller
  - Configuration, telemetry, and remote update

Darby Creek offers 1,2:
- 2.5X more logic elements
- 2X higher PCIe bandwidth
- 4X more system memory density
- 2.5X faster network interface

1 Rush Creek: 1.1M LE A10, PCIe Gen 3 x8, 8G DDR3, 40GbE
2 Darby Creek: 2.8M LE S10, PCIe Gen 3 x16, 32G DDR4, 100GbE
FINANCIAL LIBRARY API EXAMPLE
How much should the bank charge the Investor as **Option Price**?

**DNT Option Terms:**
- Currencies / Asset
- Start/Expiry Date
- Barrier levels
- Payout price

**Option Pricing Engine**

**PDE Solver**

**Investor**

**Should I buy?**

DNT = Double No-Touch
PDE = Partial Differential Equation
CREATING A PDE SOLVER IN FPGA

FPGA used to provide a solver for a particularly computationally challenging workload

- **Intent:** Improve time to results
  - More results (Present Value of Options (PVs)) in a given amount of time or compute resource
- **Starting point:** C-model implementation of PDE Solver created (880 lines of C code)
- **End point:** Optimised OpenCL™ implementation (920 lines of OpenCL)

<table>
<thead>
<tr>
<th>Task</th>
<th>Dev. Time</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convert C model to OpenCL</td>
<td>2 weeks</td>
<td>142 PV/s**</td>
</tr>
<tr>
<td>Optimise pipeline</td>
<td>1 week</td>
<td>174 PV/s**</td>
</tr>
<tr>
<td>New C Code + Open CL optimisations</td>
<td>2 weeks</td>
<td>387 PV/s**</td>
</tr>
<tr>
<td>Scale infrastructure (4 x FPGA Cards)</td>
<td>1 week</td>
<td>1511 PV/s**</td>
</tr>
</tbody>
</table>

** See appendix for server configuration (8)
Throughput scales linearly to the number of FPGAs
INTEL LIBRARIES & TURNKEY EXAMPLES
INTEL AI FOR COMPUTE

**GENERAL AI**
- Mainstream AI
- Flexible Acceleration

**TRAINING**
- Data Center/Workstation
  - Mainstream Training
  - Intensive Training

**INFERENCE**
- Data Center/Workstation
  - Mainstream Inference
  - Intensive Inference
  - Streaming Inference

**GATEWAY/EDGE**
- Mainstream Inference
- Higher Inference Throughput
- Vision 1-20W
- Speech/Audio 1-100+mW
- Autonomous driving
- Custom Inference

**DEEP LEARNING**
- Scale Your Innovation
- Intel AI for Compute
- Streaming Inference
- Mainstream Inference
- Intensive Training
- Intensive Inference
- Higher Inference Throughput
- Vision 1-20W
- Speech/Audio 1-100+mW
- Autonomous driving
- Custom Inference
WHY FPGAs WIN IN DEEP LEARNING

First to market to accelerate evolving AI workloads
- Adversarial Networks
- Reinforcement Learning
- Neuromorphic computing

Flexible system level functionality with deterministic latency
- AI+I/O ingest
- AI+Networking
- AI+Security
- AI+Pre/Post processing

Low latency memory constrained workloads
- RNN
- LSTM
- Speech WL

RNN – Recurrent Neural Network
LSTM – Long Short-Term Memory
PUBLIC INTEL FPGA MACHINE LEARNING SUCCESS

**Microsoft:** Microsoft has revealed that Intel FPGAs have been installed across every Azure cloud server, creating what Microsoft is calling the world's first AI supercomputer.

**NEC:** To create the NeoFace Accelerator, the engine software IP was integrated into an Intel Arria 10 FPGA, which operate in Xeon processor–based servers.

**JD.COM:** Arria®10 FPGA can achieves significant improvement in the performance of LSTM accelerator card compared to GPU.

**Inspur/iFlytech:** Server vendor Inspur Group and Intel launched a speech recognition acceleration solution based on Intel's Arria® 10 FPGAs and DNN algorithm from iFLYTEK.
EVOLVING DEEP LEARNING REQUIREMENTS

**2017**
- Convolutional Neural Network (CNN)
- Floating Point: FP32

**2018**
- Recurrent Neural Network (RNN)
- Floating Point: FP32, FP16, FP11, FP9, BFLOAT
INTEL® FPGA DEEP LEARNING ACCELERATION SUITE

- Supports common software frameworks (Caffe, TensorFlow)
- Intel DL software stack provides network optimizations
- Intel FPGA Deep Learning Acceleration Suite provides turn-key or customized CNN acceleration for common networks

User Visibility via OpenVINO™ toolkit

Support at DLA API level is extremely costly

Optimized Acceleration Engine

Standard ML Frameworks

Intel® Xeon® Scalable Processor

Intel® Arria® FPGA

Intel DLA Network Bitstreams

Heterogenous CPU/FPGA Deployment

Intel DLA SW API

Inference Engine

Model Optimizer

Intel Deep Learning Deployment Toolkit

- Intel Pre-Trained Models
- DLA SW
- Caffe
- TensorFlow

Pre-compiled Network Architectures

- GoogleNet Optimized Template
- ResNet Optimized Template
- SqueezeNet Optimized Template
- VGG Optimized Template
- Additional, Generic CNN Templates

Hardware Customization Supported *

* Encrypted DLA source code license required, sold separately
CPU + FPGA ACCELERATE AI INFERENCE

COMPARISON OF FRAMES PER SECOND (FPS)

GET AN EVEN BIGGER PERFORMANCE BOOST WITH INTEL® FPGA

1* Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. Performance results are based on testing as of June 13, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Configuration: Testing by Intel as of June 13, 2018. Intel® Core™ i7-6700K CPU @ 4.00GHz fixed, GPU GT 2 1.0GHz fixed Internal only testing. Test vs.11.21.1 - Ubuntu 16.04, openVINO 2018 R4. Intel® Xeon® 10 FPGA 1150CK. Tests based on various parameters such as model used (these are public), batch size, and other factors. Different models can be accelerated with different Intel hardware solutions, yet use the same Intel software tools.

Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110801.
AI FUTUREPROOFING WITH BFLOAT16 SUPPORT

- **FP32**: 8 bit exponent
- **FP16**: 5 bit exponent
- **BFLOAT16**: 8 bit exponent
FINLIB IS IN INTEL® QUARTUS® 17.0!

- Option pricing
  - European, American, Equities, Average rate, Spread normal, Spread lognormal
- Statistical functions
  - Norm_std(), norm_cdf(), norm_icdf(),...
- Working Demo accessible from the Intel labs
## GENOMICS – GATK ACCELERATION

**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 (Note 7)</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 (Note 7)</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 (Note 7)</td>
<td>OpenCL</td>
<td>1</td>
<td>44.1</td>
<td>32.4</td>
</tr>
</tbody>
</table>

Performance results are based on testing as of March 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. System Configuration: [Click for more information](#)
PARTNER LIBRARIES & TURNKEY EXAMPLES
FALCON ACCELERATED GENOMICS PIPELINE

Adaptive cross-hardware platform to provide a path to efficient & cost-effective genome analysis

> 10X Speed-up on GATK4 & GATK3.x

- End-to-end solution with performance optimization
- Accelerates GATK best practices; **No proprietary pipelines**
- Supports multiple GATK versions (3.8 & 4.0)
- Both germline and somatic best practices

**Available on public & private clouds or on-premise**

**Falcon Accelerated Genomics Pipelines**

**Performance Comparison for GATK WGS from Alignment to Variant Calling**

- Intel Xeon E5-2676v3 40 cores, 160GB Mem., HDD 64.7
- Intel Xeon E5-2699v3 36 cores, 256GB Mem., HDD 36.1
- Intel Xeon Platinum 8180 56 cores, 512GB Mem., SSD, Intel Arria 10 PAC card 10.5
- Intel Xeon Gold 8148 40 cores, 512GB Mem., SSD, Intel PAC FPGA 4.59

**Falcon solution is so fast! What had taken me over a week to do on my computer cluster, I was able to do with the Falcon-accelerated Genomics pipeline in a few hours**

Amy Cummings, MD, UCLA Medicine
FPGAs Offer Unique Value for Analytics/Streaming

Single Multi-function Accelerator

Offloads algorithm, networking, and data access processing

Moderate Acceleration is common
- PCIe lookaside acceleration (two data copies)

Significant Acceleration requires FPGA
- Multifunction and inline w/single FPGA
- Relational: 2.3X TPC-H w/Swarm64⁴
  - PostgreSQL, MariaDB, MySQL, ...
- NoSQL: 4X Cassandra⁵ w/rENIAC (80/20 R/W)
- Hadoop/Spark: 3X+ data streaming⁶ w/Bigstream, Megh
DIFFERENT DATA STORE APPROACHES

Structured Data/ Relational

Semi-structured Data/NoSQL

Unstructured Data
Database accelerate with a plugin

Acceleration Overview
- 20X+ single table inserts/s for real time data analytics
- With modest tuning, 15M PostgreSQL INSERT/s
- 2.3X* TPC-H data warehousing on Arria 10
- 3X+ TPC-H for many CSP hosting configs
- 3X+ storage compression
- Data & tables managed by Swarm64

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

Note *: TPC-H SF1000, Dual Intel® Xeon® Gold 6130, 2.10 GHz, (12) 32GB DDR4-2166, (4) 960GB SSD RAID0 HPE MK000960GWJPP

Source: Swarm64
NOSQL: SYSTEM & IO ACCELERATION OPPORTUNITY
SOURCE: RENIAC CEO

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

System & IO: 75%

Business Logic: 25%
RENIAC DISTRIBUTED DATA ENGINE/SWITCH (RDS)
4X+ CASSANDRA (80% R/20% W), POCs OF 2X, GOING TO 4X BY FEB

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
Ingest/Apache Kafka*: Extract, transform, load and filtering (BigStream, Megh)

- SQL over Apache Spark (BigStream)
- BigDL: Deep learning acceleration (Megh)
- Machine learning MLlib: e.g. ALS (Megh)
- Hadoop/Spark: Shuffle phase (A3Cube)
Frictionless acceleration: Arria 10 and Stratix 10
- Zero code changes
- Cross platform: Spark, Kafka, TensorFlow
- Cloud or on-prem

Intelligent and adaptive
- Automatic profiling and partitioning of computation
  - Between CPU and FPGA
- Overlay dataflow execution on FPGA

Kafka consumer speedup up to 13X

Spark SQL TPC-DS results
- 4X average speedup for 26 of the queries with Arria 10

Industry targets: FinServ/FinTech, AdTech, Healthcare

Use cases: Spark SQL analytics, ingest/ETL, EDW
REAL TIME ANALYTICS STACK OPTIMIZED FOR HETEROGENEOUS CPU+FPGA PLATFORM

Data Sources
- Internal Data Sources
- External Data Sources

ETL
- kafka
  - Connector

Data Processing
- Streams
- ML
- DL

FPGA
- Packet RX
- Accelerated Functions
- Packet TX

Data Stores
- mongoDB
- Hazelcast

Application

Admin
- Microsoft Azure
- Amazon Web Services

BI
- Tableau

CPU+FPGA platform for 1 in-line processing of streaming analytics and 2 off-load processing of ML and DL to deliver >5x performance efficiency and provide actionable operational insights.

SCALE YOUR INNOVATION
INTEL APACHE YARN SUBMISSION: [https://issues.apache.org/jira/browse/YARN-5983](https://issues.apache.org/jira/browse/YARN-5983)
SUMMARY
SUMMARY

- FPGAs are ready for scale out and scale up
- Intel ® Acceleration Stack: driver, FPGA Interfaces, virtualization, security, etc.
- Variety of Interfaces: OpenCL™, library call, framework level
- AI, Genomics, and Financial acceleration options
- Data Analytics acceleration with no change to application required
  - Relational DB, NoSQL, SPARK* shuffle phase, Kafka* streaming, BigDL (deep learning)
- FPGA advantage of multiple concurrent functions and inline acceleration
(1) A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services, sec 2.3

(2) Microsoft's Production Configurable Cloud (Mark Russinovich) (Slide 26)
https://www.slideshare.net/ChrisGenazzio/microsofts-configurable-cloud

(3) Accelerating Persistent Neural Networks at Datacenter Scale

(4) TPC-H SF1000, Dual Intel® Xeon® Gold 6130, 2.10 GHz, (12) 32GB DDR4-2166, (4) 960GB SSD RAID0 HPE MK000960GWJPP, CentOS 7.4.1708, Kernel 3.10.0-693.21.1.e17.x86_64, Docker 18.03.0.ce, Swarm64 DB 1.4.1-PREVIEW, PostgreSQL 10.3

(5) Cassandra Stress Test (80% R/20% W)
   ▪ “Dual Xeon E52670, 2.6 Ghz, 32 cores total, 64GB ram, 1 TB NVMe, Centos 7.4”

## Configuration Details for ‘PairHMM Comparison - Xeon/FPGA’

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 www.intel.com/benchmarks.

![INTEL](https://www.intel.com)

<table>
<thead>
<tr>
<th>INTEL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>Intel® Xeon CPU E5-2699, v4, 2.20 GHz</td>
</tr>
<tr>
<td><strong>FPGA</strong></td>
<td>Intel® Arria® 10 GX, 10AX11552F45I1SG2</td>
</tr>
<tr>
<td>ALM</td>
<td>427,200</td>
</tr>
<tr>
<td>Memory</td>
<td>53.0 Mb</td>
</tr>
<tr>
<td>DSP Blocks</td>
<td>1,518</td>
</tr>
<tr>
<td><strong>Systolic Array</strong></td>
<td>208 Processing Elements (PEs)</td>
</tr>
<tr>
<td><strong>FPGA Resource Usage</strong></td>
<td></td>
</tr>
<tr>
<td>Logic</td>
<td>55%</td>
</tr>
<tr>
<td>Memory</td>
<td>50%</td>
</tr>
<tr>
<td>DSP Blocks</td>
<td>99%</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>230 MHz</td>
</tr>
<tr>
<td><strong>Input Data</strong></td>
<td>Chromosome 21 from 30x WGS NA12878</td>
</tr>
</tbody>
</table>
**SYSTEM CONFIGURATION FOR PERFORMANCE TESTING**

Server configuration:
Dell PowerEdge R740
2 x Intel® Xeon® Gold 6132 @ 2.6 GHz
192GB (12 x 16GB) RDIMM, 2666MT/s, Dual Rank

Operating System:
Red Hat Enterprise Linux: Release 7.5 with Linux kernel 3.10.0-862.el7.x86_64

FPGA:
Intel Programmable Acceleration Card with Intel Arria® 10 GX FPGA, Acceleration Stack version 1.0

Test performed during August 2018.
OpenCL code was developed within Intel® Programmable Solutions Group.
Functional correctness was verified by comparison with single-precision floating point results from CPU, using the “==” operator in C/C++

Tests were performed with pre-production, proof-of-concept code.

Not all capabilities are part of shipping products.