ACCELERATING WORKLOADS USING THE ACCELERATION STACK FOR INTEL® XEON® CPU WITH FPGAS

David Munday
Sr. Manager, Data Center Acceleration Engineering
FPGA Software Solutions
Programmable Solutions Group, Intel Corporation
BY 2020...

<table>
<thead>
<tr>
<th>Category</th>
<th>Data/Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVG. INTERNET USER</td>
<td>1.5 GB of traffic/day</td>
</tr>
<tr>
<td>AUTONOMOUS VEHICLES</td>
<td>4 TB of data/day</td>
</tr>
<tr>
<td>SMART FACTORY</td>
<td>1 PB of data/day</td>
</tr>
<tr>
<td>CLOUD VIDEO PROVIDERS</td>
<td>750 PB of video/day</td>
</tr>
</tbody>
</table>

THE COMING FLOOD OF DATA

Source: Amalgamation of analyst data and Intel analysis.
THE WORLD OF COMPUTE MUST ACCELERATE

Autonomous Driving
- 1GB/s real-time processing
- Sense, understand, react safely in <1 second

Networking & 5G
- 1,000x increase in bandwidth
- 100x more connected devices
- 1ms end-to-end round trip delay

Cloud Computing
- 2013: 4.4ZB → 2020: 44ZB data on the planet
- Need for scale, throughput, compute efficiency and lowest latency

HETEROGENEOUS ARCHITECTURES MEET DATA DEMANDS

- Acceleration of compute means heterogenous compute
- Dedicated accelerators for maximum compute efficiency
- Flexible accelerators for maximum compute flexibility
- Shift towards software-defined hardware
AUTONOMOUS DRIVING ACCELERATION

Acceleration of image and sensor processing

- ASSP accelerates vision processing
- FPGA accelerates fusion, machine learning and security
SMART NIC AND CLOUD ACCELERATION

Acceleration of networking, compute and storage

NIC ASIC accelerates networking functions

FPGA accelerates compute such as transcoding, encryption and custom networking functions

HOST

CPU

NIC ASIC

FPGA

Network

SMART NIC
Many silicon artificial intelligence (AI) accelerators ... require ... ‘batching’ ... to achieve high performance ... Intel® Stratix® 10 devices demonstrated more than 39 teraflops of sustained performance on a single request ... a new level of cloud performance for real-time AI computation, with record low latency, ...
Database acceleration with a plug-in

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

2X+ queries for data warehousing
- Using industry standard TPC-DS benchmark

3X+ storage compression
- Data and tables managed by Swarm64

>40% Total Cost of Ownership Savings

Source: Swarm64
1. Based on average Giga-Cell Updates per Second (GCUPS) performance using Intel® Arria® 10 GX FPGA acceleration compared to Intel Xeon® processor single-core with Intel Advanced Vector Extensions (AVX) technology.
2. Overall genomics pipeline performance improvement with Intel Arria® 10 GX FPGA acceleration for PairHMM algorithm compared to Intel® Xeon® processor single-core with Intel Advanced Vector Extensions (AVX) technology.

System configuration: Intel Xeon processor E5-2697 v2 @ 2.70 GHz, 2 sockets/12 cores per socket, 128 GB RAM, 2 TB Seagate HDD ST2000DM001 with Intel Arria 10 GX FPGA compared to Intel Xeon processor E5-2699 v4 @ 2.20 GHz, 2 sockets/22 cores per socket, 256 GB RAM, 2 TB Intel SSD Data Center P3700 Series with single-core Intel Advanced Vector Extensions (AVX) technology.

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
SOLVING REAL-WORLD PROBLEMS: STORAGE NVME OVER FABRIC

FPGA-based hardware implementation “cut-through” design provides
58 to 72% lower read/write latency\(^1\)
than network interface card (NIC) + software solution

---

1. Read/write latency performance for Attala Systems Development Host NVMe Adapter compared with Mellanox ConnectX* 4 Lx EN RNIC with Linux* OS on Intel® Xeon® processor E5-2600 v3. System configuration: Quanta D51B with Intel® Xeon® processor E5-2600 v3, Attala Systems Development Host NVMe Adapter with Intel® Arria® 10 FPGA, Intel SSD DC P3700 STD compared to Quanta D51B with Intel Xeon processor E5-2600 v3, Mellanox ConnectX-4 Lx EN RNIC, Intel SSD DC P3700 STD.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYMark 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.
FPGAS CRITICAL TO HETEROGENEOUS ARCHITECTURES

Delivering the performance of hardware with the programmability of software

FLEXIBLE
REPROGRAMMABLE
INHERENTLY PARALLEL
LOW LATENCY
HIGH PERFORMANCE
ENERGY EFFICIENT
“I don't speak FPGA!

What is the programming model, and where are the compilers, libraries and tools I am used to?”
SOFTWARE DEVELOPERS ARE THE NEW FPGA DEVELOPERS

- Functional Abstraction
- Platform compatibility
- Eco-System Richness
EMPOWERING DEVELOPERS WITH A COMPLETE SOLUTIONS STACK

Software Developers

- Application Software
- Software Development Kits
- Reference Platforms
- Integrated Frameworks and Tools
- Language Support, Compilers and Libraries/IP

Hardware Developers

- Platform Development Tools

Ecosystem:
Open Source Community Partners
FPGA ACCELERATION STACK WITH LIBRARIES

User Application
Software Frameworks
Library Orchestration
OpenCL™ Runtime

User Design
Library Infrastructure
Library kernels
Compute Primitives
Infrastructure Primitives

Intel® FPGA SDK for OpenCL
Board Support Package (BSP)

Gap: Creating full-stack accelerated applications on FPGA can be difficult and time consuming.

Library Orchestration
Increase Abstraction
Increase User Base

Driver
FPGA DEVELOPMENT AND PROGRAMMING TOOLS

Software Developer

1. Intel® SoC FPGA Embedded Design Suite (EDS)
2. Intel® FPGA SDK for OpenCL™

Algorithm Designer

1. DSP Builder for Intel® FPGAs

IP Library Developer

1. Intel® HLS Compiler

HDL Designer

1. Intel® Quartus® Prime Software Suite
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
DEPLOYING FPGA ACCELERATION IN THE CLOUD

Public and Private Cloud Users

Launch Workload

Workload

Orchestration Software (FPGA-Enabled)

Resource Pool

Storage

Network

Compute

Examples:
- Machine Learning
- Encryption, Compression
- Big Data Analytics
- NFV, vSwitch

End-User Developed Functions

Intel-Developed Functions

3rd-Party Developed Functions

Examples:

- Big Data Analytics
- Encryption, Compression
- Machine Learning
- NFV, vSwitch

VM = Virtual Machine
AF = Accelerator Function
INTRODUCING: INTEL® PROGRAMMABLE ACCELERATION CARD WITH INTEL ARRIA® 10 GX FPGA

Enabled by Acceleration Stack for Intel Xeon® CPU with FPGAs
For broad adoption in the data center

Sampling: NOW
General availability: 1H 2018
THIS IS JUST THE BEGINNING...

INTEL® PROGRAMMABLE ACCELERATION CARDS (PAC) WITH INTEL ARRIA® 10 GX FPGA

SAMPLING TODAY
GENERAL AVAILABILITY 1H2018

INTEL XEON® PROCESSOR SCALABLE PLATFORM WITH INTEGRATED FPGA

SAMPLING TODAY
GENERAL AVAILABILITY 2H2018

NEXT GENERATION PACS AND PLATFORMS

• HIGHER PERFORMANCE
• INCREASED CONNECTIVITY
• MORE INTEGRATION OPTIONS

Application and Intellectual Property (IP) Migration to Multiple Platforms
ECOSYSTEM - ACCELERATION STACK FOR INTEL® XEON® CPU WITH FPGAS

EASING DEVELOPMENT AND DATA CENTER DEPLOYMENT OF INTEL FPGAS FOR WORKLOAD OPTIMIZATION
The flood of data and new user experiences drive up compute needs.

FPGAs are critical for acceleration in heterogenous compute platforms.

Programming models must evolve to enable FPGAs for the software developer.

Intel offers a comprehensive solutions stack including FPGA-optimized libraries, compilers, tools, frameworks and SDK integration, and an FPGA-enabled ecosystem.

Use Intel® architecture performance-optimized tools and frameworks

Contact your Intel representative for more information

Find out more online at Intel.com and Intel FPGA website

Want to evaluate on FPGA? Contact us and we can get you started today
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.

Swarm64 system configuration: Supermicro® SuperServer 2028U-TR4+ with Super X10DRU-i+ Mainboard, 2X Intel® Xeon® E5-2695 v4 CPUs, 8X Samsung® 32GB DDR4-2400 ECC RAM. Note: This is SQL to relational database, not SQL to semi/unstructured data. Data Warehousing tested with data taken from the TPC-DS benchmarking suite and tested against the 99 TPC-DS query templates.

Swarm64 Total Cost of Ownership (TCO) projection derived from example pricing calculation based on Amazon Web Services (AWS) pricing as follows: Linux on f1.2xlarge FPGA Instance, 3 year no upfront convertible, pricing for US East / US Standard (Virginia), accessed via https://calculator.s3.amazonaws.com/index.html on Sept 26 2017.

Broad Institute system configuration: Intel® Xeon® processor E5-2697 v2 @ 2.70 GHz, 2 sockets/12 cores per socket, 128 GB RAM, 2 TB Seagate HDD ST2000DM001 with Intel Arria 10 GX FPGA compared to Intel® Xeon® processor E5-2699 v4 @ 2.20 GHz, 2 sockets/22 cores per socket, 256 GB RAM, 2 TB Intel® SSD Data Center P3700 Series with single-core Intel® Advanced Vector Extensions (AVX) technology.

NVME over Fabric system configuration: Quanta D51B with Intel® Xeon® processor E5-2600 v3, Attala Systems Development Host NVMe Adapter with Intel Arria* 10 FPGA, Intel SSD DC P3700 STD compared to Quanta D51B with Intel Xeon processor E5-2600 v3, Mellanox ConnectX®-4 Lx EN RNIC, Intel SSD DC P3700 STD.

Tests document 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 http://www.intel.com/performance.

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.

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.

This content contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Statements in this document that refer to Intel’s plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel’s results and plans is included in Intel’s SEC filings, including the annual report on Form 10-K.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

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.

© 2017 Intel Corporation. Intel, the Intel logo, and others are trademarks of Intel Corporation in the U.S. and/or other countries.

Altera, Arria, Cyclone, Enpirion, Max, Megcore, Nios, Quants and Stratix, words and logos are trademarks of Altera and registered in the U.S. Patent and Trademark Office and in other countries.

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