Intel® HPC Solutions Update Focus on FPGA and ML

Dr. Jean-Laurent PHILIPPE, PhD
EMEA HPC Technical Sales Specialist

With Dell
Amsterdam, October 27, 2016

Legal 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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].

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.

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

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

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 https://www-ssl.intel.com/content/www/us/en/high-performance-computing/path-to-aurora.html.

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.

Intel, the Intel logo, Xeon, Intel Xeon Phi, Intel Optane and 3D XPoint are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

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

Agenda

Introduction

Intel® Xeon® Processor and FPGA

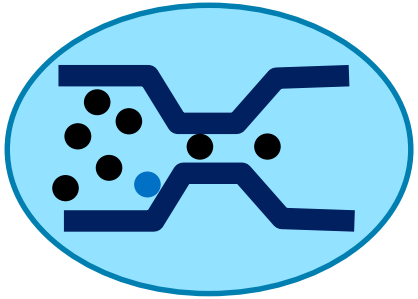
Machine Learning

Conclusion

Backup

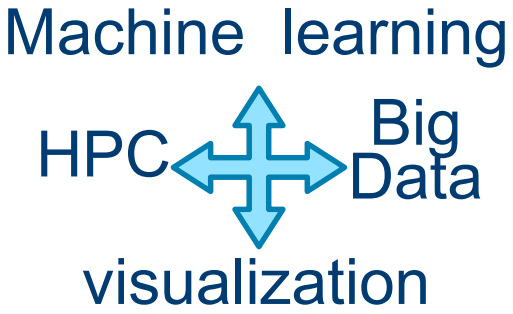
Growing Challenges in HPC

System Bottlenecks
"The Walls"



Memory | I/O | Storage Energy Efficient Performance Space | Resiliency | Unoptimized Software

Divergent Workloads



Resources Split Among Modeling and Simulation | Big Data Analytics | Machine Learning | Visualization

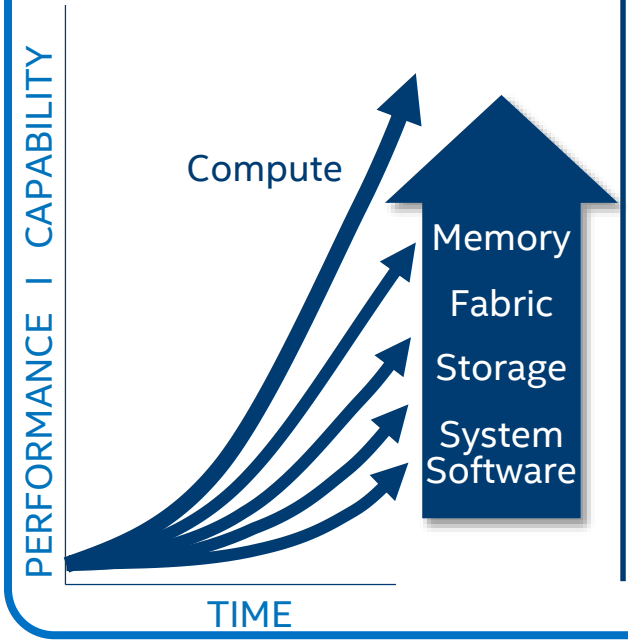
Barriers to Extending Usage

Democratization at Every Scale | Cloud Access | Exploration of New Parallel Programming Models

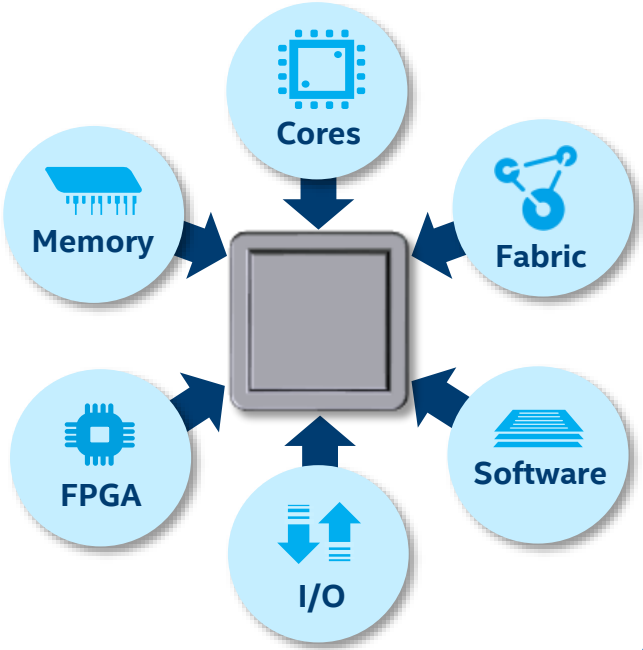
A Holistic Architectural Approach is Required

System

Innovative Technologies

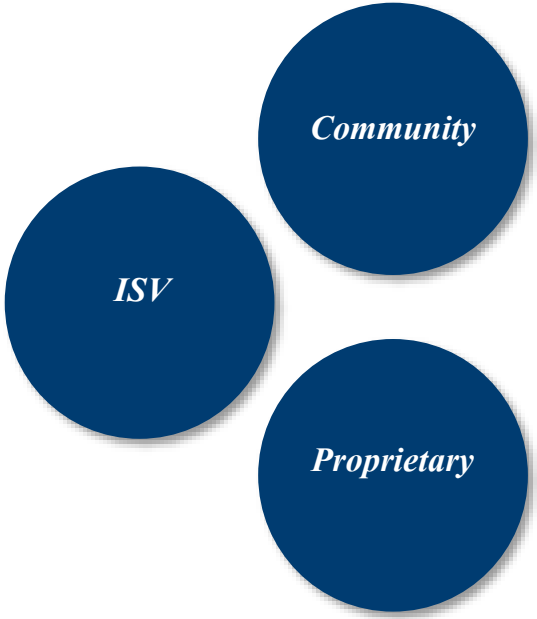


Tighter Integration



Application

Modernized Code



Intel® Scalable System Framework

Modeling & Simulation

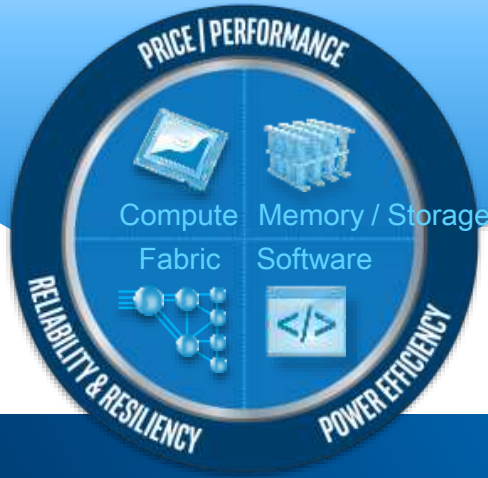
HPC Data Analytics

Machine Learning

Visualization

Many Workloads – one Framework

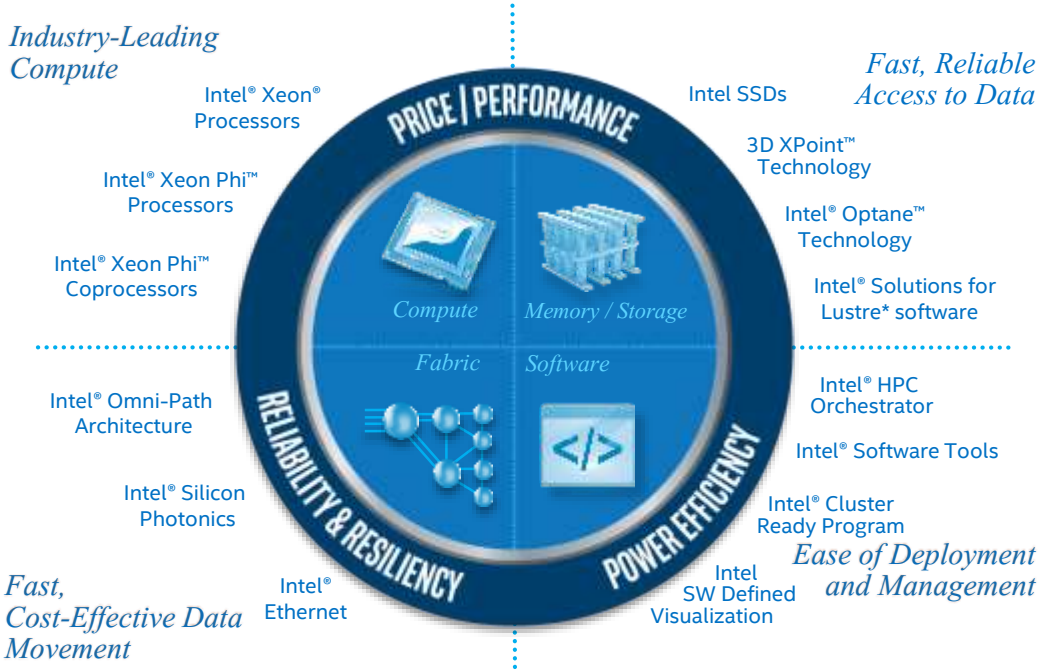
A Flexible Framework for Today & Tomorrow



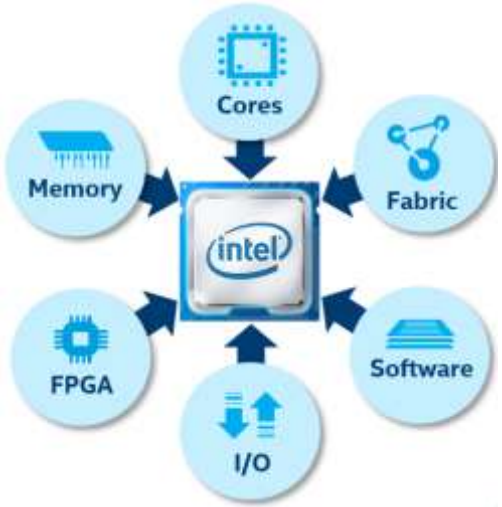
Delivering Breakthrough System Performance

How Intel® Scalable System Framework Works

Innovative Technologies



Tighter Integration
and Co-Design



Benefits

Compatibility

Bandwidth

Density

Latency

Power

Cost

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

Agenda

Introduction

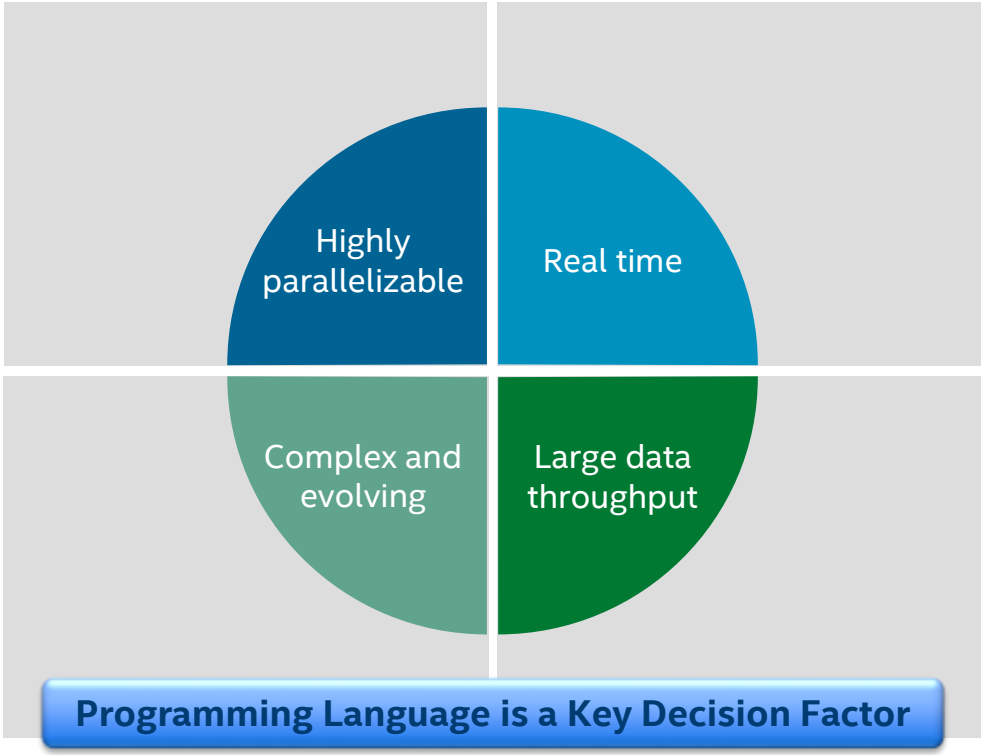
Intel® Xeon® Processor and FPGA

Machine Learning

Conclusion

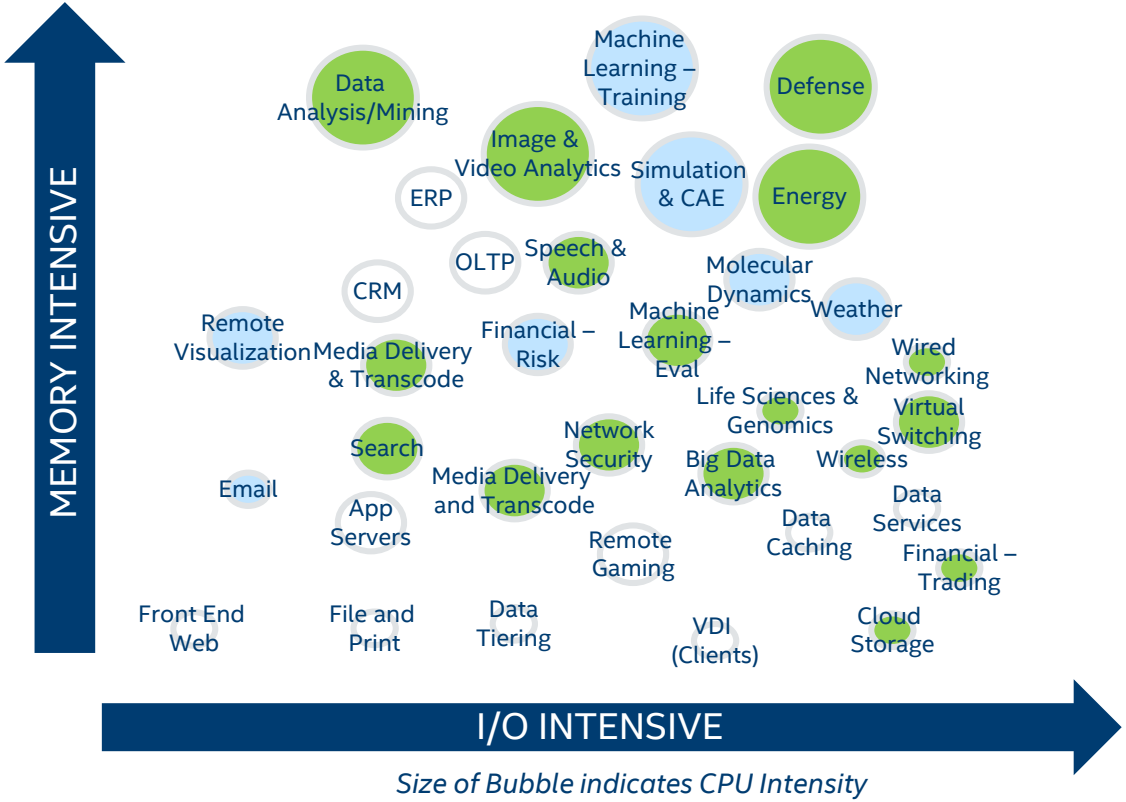
Backup

Suitable Workloads for FPGAs



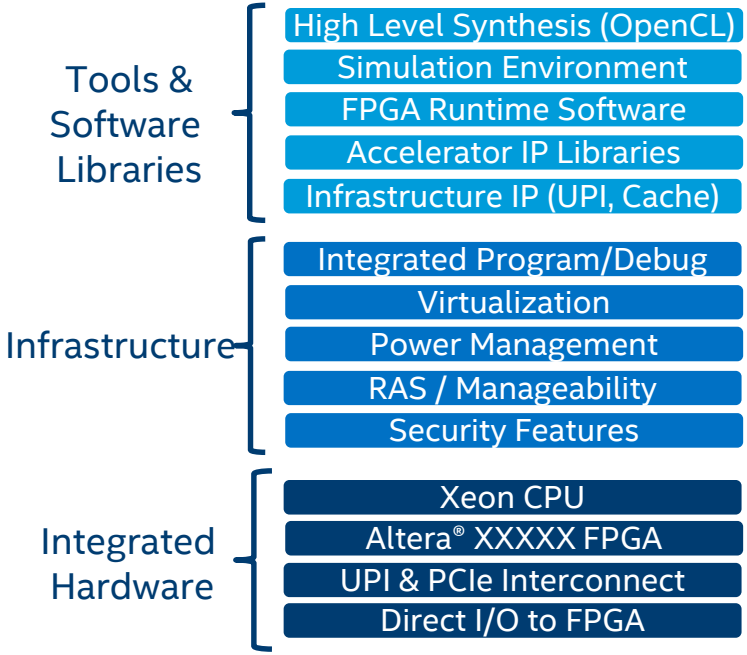
Source: Bain FPGA market research (October 2015) survey of 400 developers

Data Center Workloads for FPGA



Very Applicable
Applicable
Less Common

Intel® Xeon® processor + FPGA Value Proposition



High Performance

- Accelerate inline data stream processing without look aside memory copy
- Up to 2x lower latency & up to 2x higher bandwidth vs discrete FPGA
- Cache-coherent access to FPGA & shared system memory with CPU

Low Total Cost of Ownership

- Deploy FPGA in socket with CPU in server form factor
- Resource sharing for platform components (memory, power, Ethernet)

Ease of Deployment

- Pre-loaded Infrastructure IP for CPU-to-FPGA interaction (UPI, Cache, I/O)
- Accelerator IP libraries (Intel & 3rd party) aligned to key usages
- Program FPGA with existing Altera® tool suite in OpenCL or RTL

Delivered through combined hardware & software features of Xeon + FPGA

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Xeon + FPGA Target Workloads

FPGA Activity	Workload Examples
Compute intensive algorithms	▪ Visual Understanding/Deep Learning classification ▪ Compression/decompression ▪ Video Motion Estimation ▪ Genomics (Pair HMM, Smith Waterman) ▪ Memory copy routines
Latency sensitive pre-filtering & processing for CPU	▪ Bump in the wire network processing ▪ FSI market data pre-filtering ▪ HPC Radar data pre-processing ▪ Automotive video input ▪ Security appliance, targeted Vswitch
Evolving algorithms or stable algorithms on low latency and inline interconnect	▪ New compression algorithms ▪ High compression ratios ▪ Custom crypto algorithms

Xeon + FPGA Use Case Examples by Segment

	Cloud SPs	Comm SPs	Enterprise IT	Tech Computing
Example End user	SaaS/IaaS provider	NFV adopter	Database, Big Data Analytics user	FSI user
Workload accelerated on FPGA	Visual Understanding	VM-to-VM Packet Processing	Database Compression	Trading algorithms
Sample FPGA IP Libraries	Convolutional Neural Network algorithms	Parse, Lookup, Modify	Compression, Sort, Join algorithms	Proprietary

Differences between Discrete FPGA & Xeon + FPGA

	Discrete FPGA	Xeon + FPGA
Workload type best suited for	Coarse-grained acceleration offload: FPGA works on independent task, returns result to CPU when complete	Fine-grained workload acceleration: CPU/FPGA jointly working on task, access shared data set
Where to deploy FPGA	- On PCIe card - On motherboard	In server form factor inside CPU socket (up to 2 sockets)
FPGA Options	- Option of any FPGA to deploy - Option to deploy multiple FPGAs together	- 1 FPGA option available - 1 FPGA integrated with CPU as multi-chip package
Memory Options	- Option for memory local to FPGA - System memory access is via PCIe & not cache-coherent	- FPGA shares system memory with CPU - System memory access is low latency & cache-coherent
Power Options	FPGA powered separately from CPU	FPGA & CPU share socket TDP
Tools & Programming	- Same Altera tool suite for discrete & integrated FPGA - Program FPGA with OpenCL or RTL	

Choice between the reconfigurable accelerators will depend on workload demands & deployment environment

Agenda

Introduction

Intel® Xeon® Processor and FPGA

Machine Learning

Conclusion

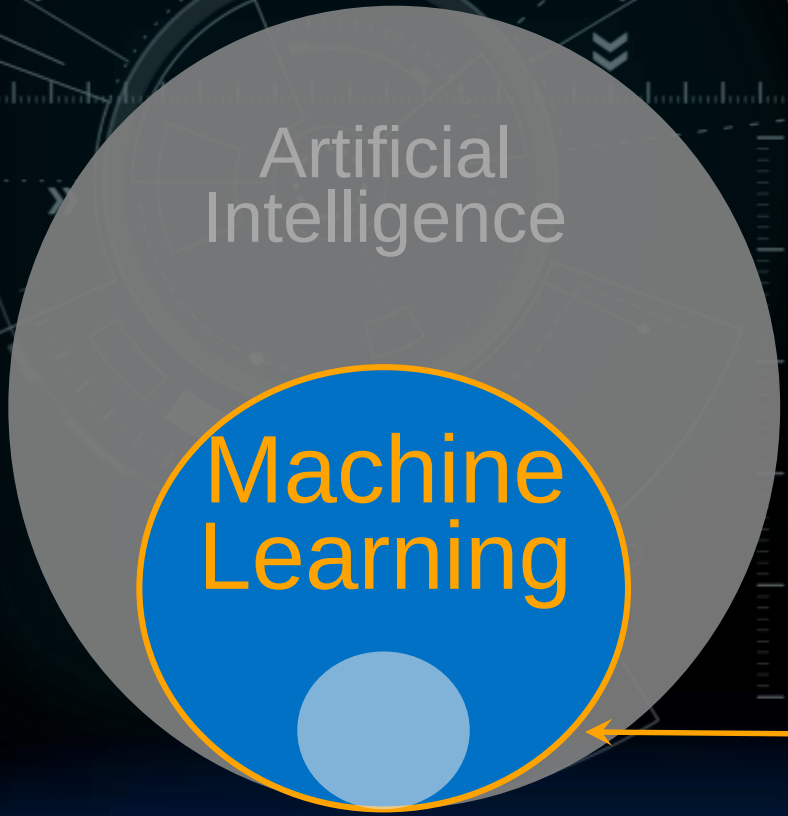
Backup

Artificial

Intelligence

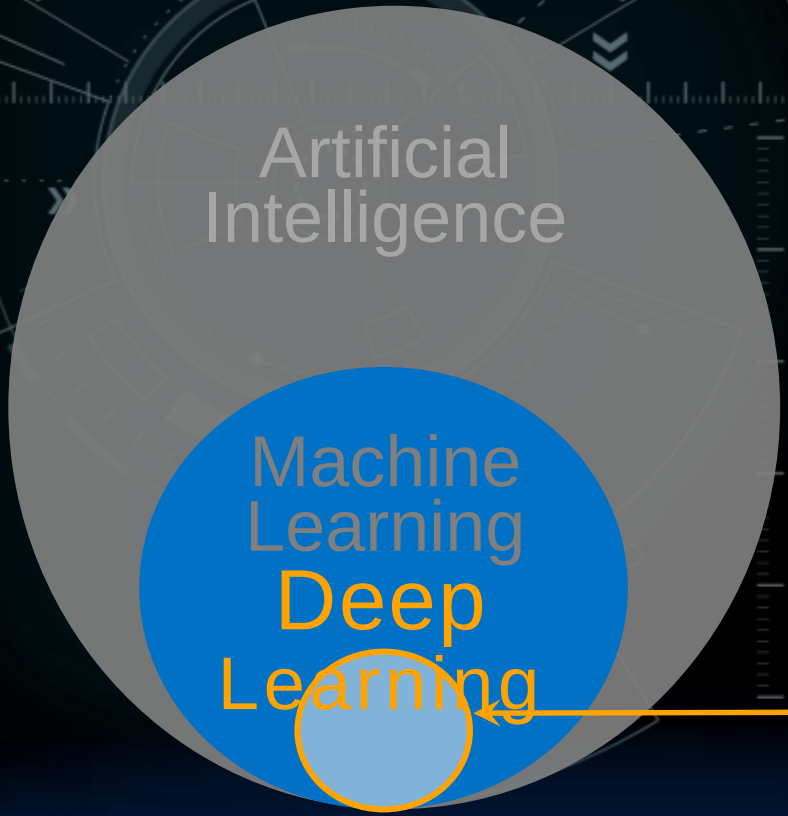
is

Human Intelligence
Exhibited by Machines



Machine Learning
is a small, but fast growing workload

- **Training:** Simple math applied at massive scale to analyze & create a model
- **Scoring:** Trained models are applied to new data to generate predictions
  - *Future: Autonomous computation methods that learn from experience*



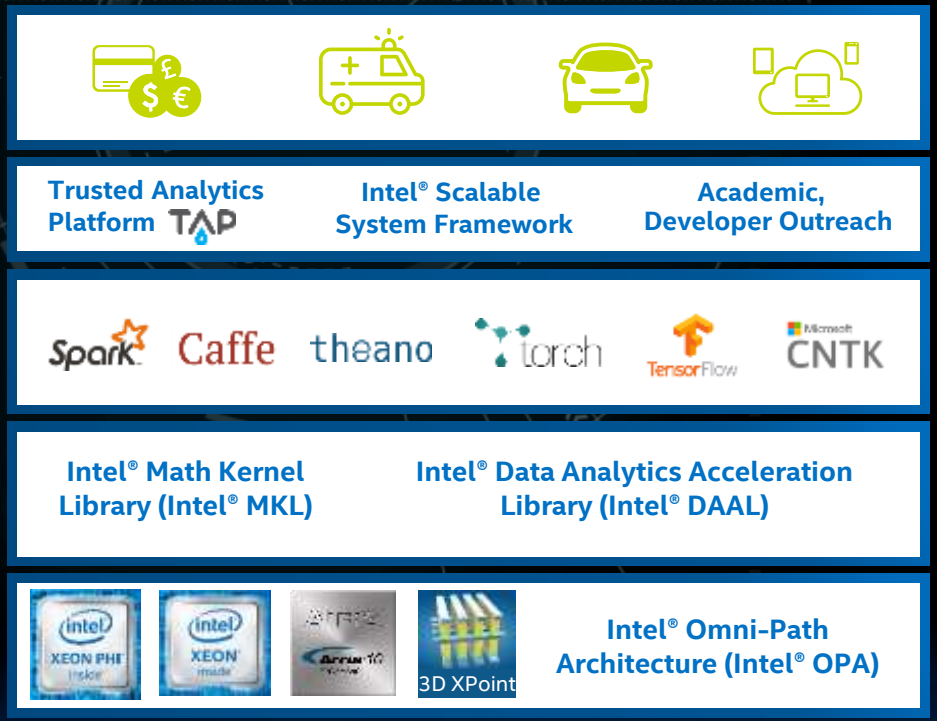
Deep

Learning

is

One Branch of Machine Learning

Intel's AI Framework



Fuel the development of vertical specific solutions

Accelerate adoption of analytics platforms

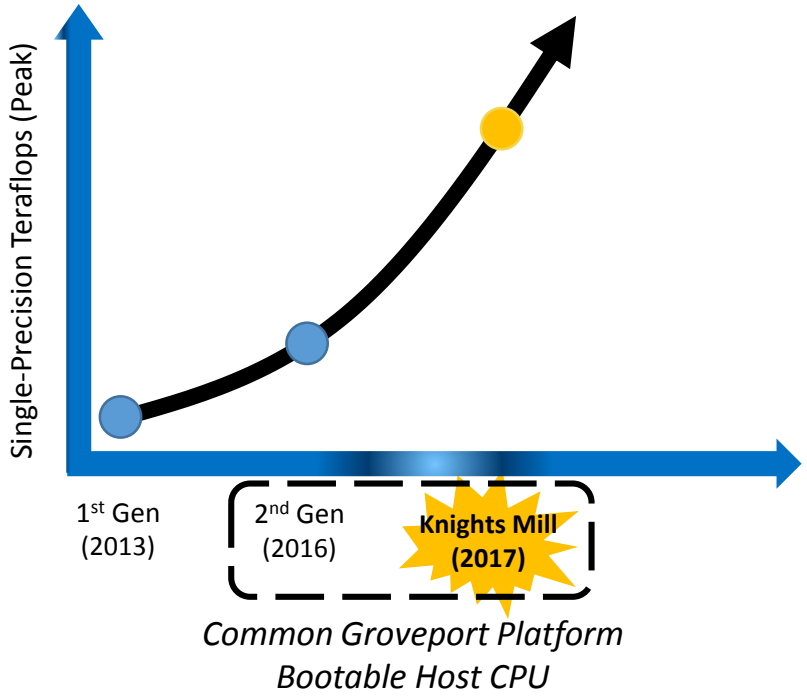
Drive CPU optimizations across open source machine learning frameworks

Enable maximum performance with Intel libraries

Deliver best single node and multi-node performance

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

Knights Mill: Optimal Deep Learning Throughput



Faster Time to Train Machines

- Provides High Single Precision Peak performance
- Provides High Variable Precision QVNNI performance
- Bootable Host-CPU avoids PCIe latency & bottlenecks
- Efficient Scaling with Multi-node optimizations for top ML frameworks
- High memory bandwidth for seamlessly training Complex Neural Network datasets

Agenda

Introduction

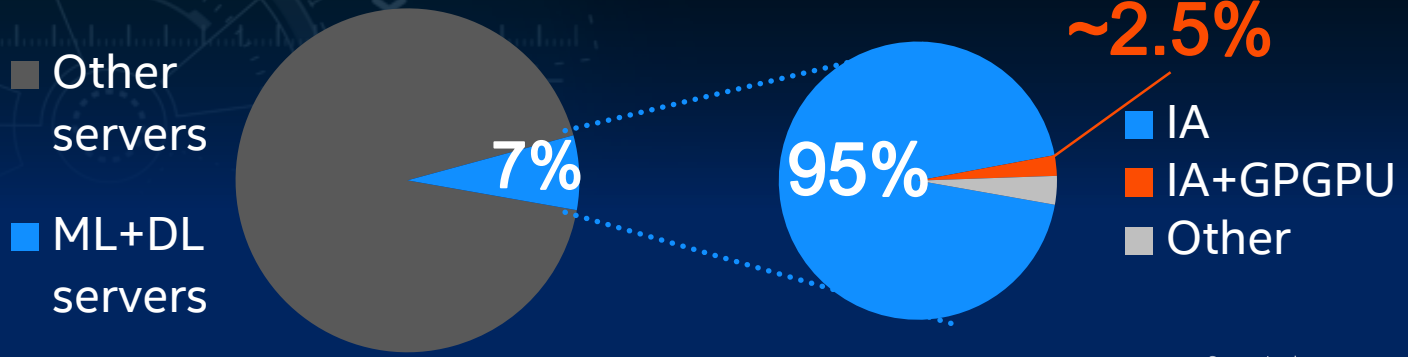
Intel® Xeon® Processor and FPGA

Machine Learning

Conclusion

Backup

IA for AI: Better Hardware Today & Tomorrow



Source: Intel

Intel Platform for Training

Best Performance

Maximum scalability

Intel Platform for Scoring

Best Performance Per TCO

Preeminent Scoring Solution

Best Performance Per Watt

Maximum Flexibility

Thank you …

… Q&A ?


intel
experience
what's inside™

Backup. Dell offering

Dell PowerEdge C4130 Accelerator platform

6.01
double-precision teraFLOPS per unit

10.04
Single-precision teraFLOPS per unit

33% improved density versus competition

PCIe Gen3 switch option for inter-Phi communication

Industry-leading accelerator density and unmatched flexibility

1U height

4x Xeon Phi

2x Xeon CPU

16x DDR4 DIMMs

DELL EMC | intel

Dell PowerEdge C6320 high-performance platform

Flexible and independent server nodes

144 processing **cores** in a 2U chassis

4 TFLOPS **performance** per chassis

iDRAC 8 system management

Performance optimized

2U	4x	24x	12x
height	server sleds	2.5” drives	3.5” drives

DELL EMC intel

Dell PowerEdge R930 Large-Memory platform

Up to **6TB RAM**

72 Intel Xeon **Compute Cores**

24 x 2.5” HDD
8 x NVMe SSD

iDRAC 8 system management

Designed for the most demanding **HPDA** applications

4U height

10x PCIe slots

8x PCIe Flash SSD

16x 3.5” drives

DELL EMC | intel

The Dell H-Series Omni-Path Architecture

The **next-generation** of High-Performance Computing fabrics

100 Gbs

.9us Latency

Software

Open Source
Host Software and Fabric Manager

Cables

Passive Copper Active Optical

Up to **17% lower latency** than InfiniBand

Up to **17% lower latency** than InfiniBand

24 or **48 Port** Edge Switches

24 or **48 Port** Edge Switches

DELL EMC intel

Dell Intel® EE Lustre* Software

Limitless
Endless Scalability

Parallel
For ultimate scale-out

Hadoop
Converged Platform

11GB/s
Peak Read
per building block

7GB/s
Peak Write
per building block

Up to **3PB** usable in a single rack

Up to **44GB/s** in a single rack

The **Ultimate** HPDA File System

DELL EMC intel