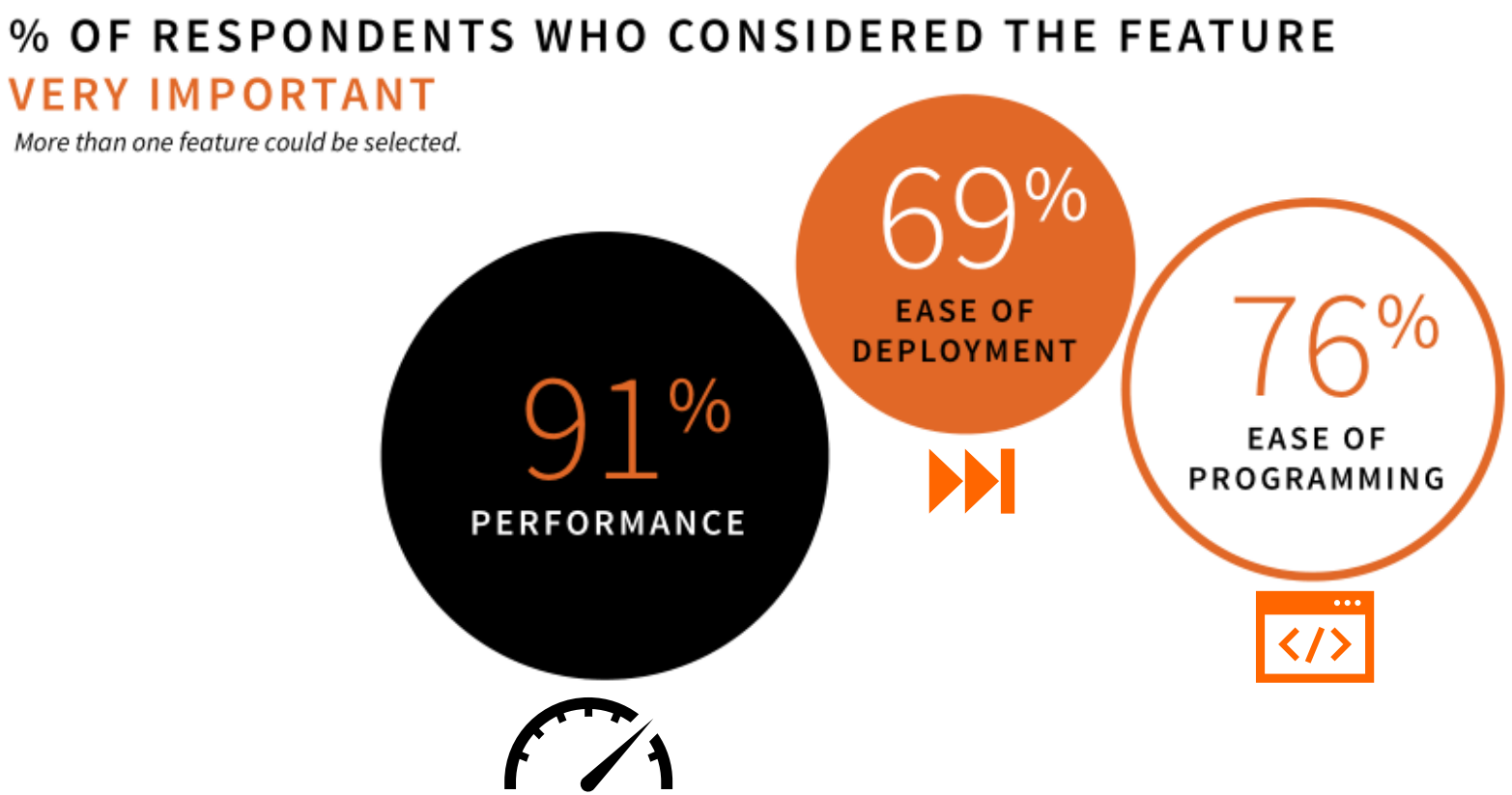
Accelerated ML on cloud FPGAs

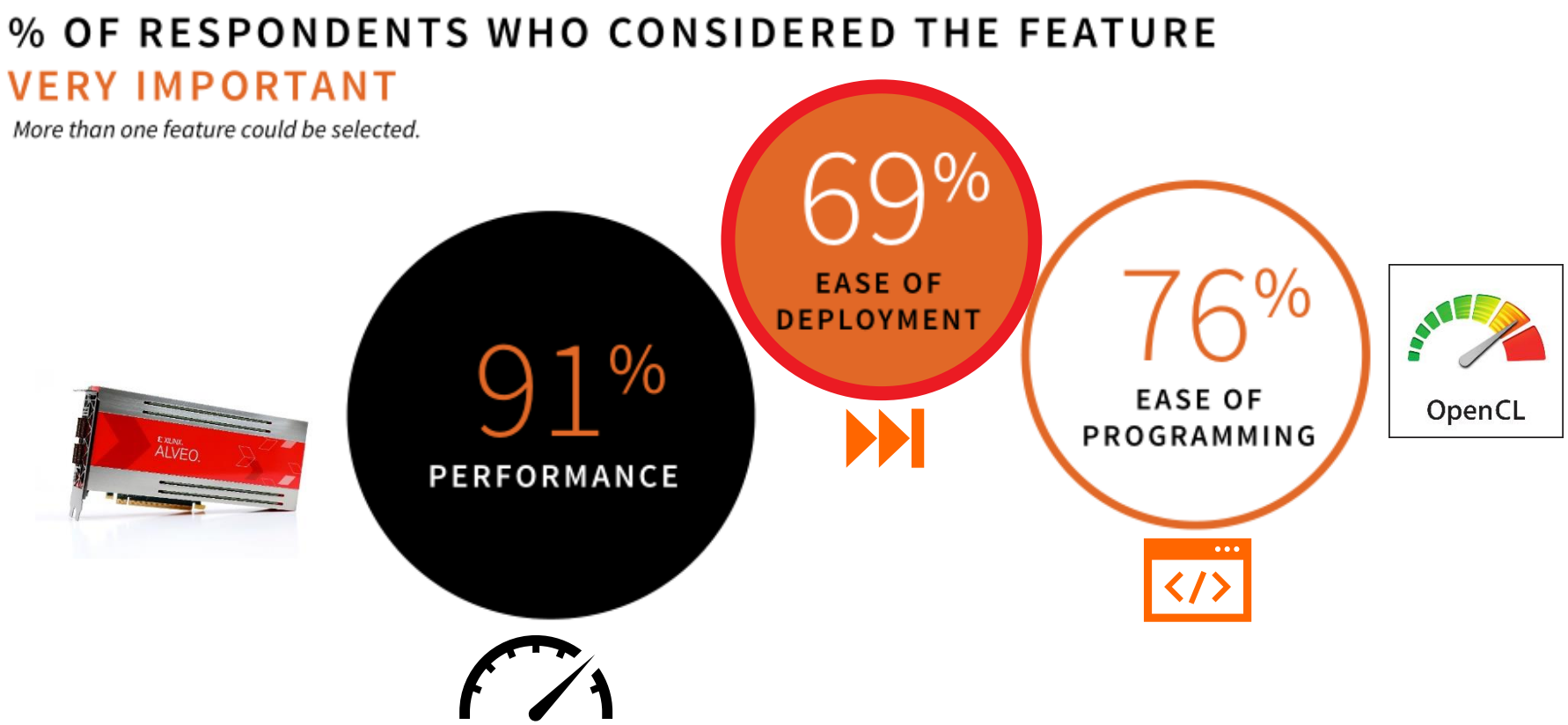
Christoforos Kachris
kachris@microlab.ntua.gr

What software developers/users want



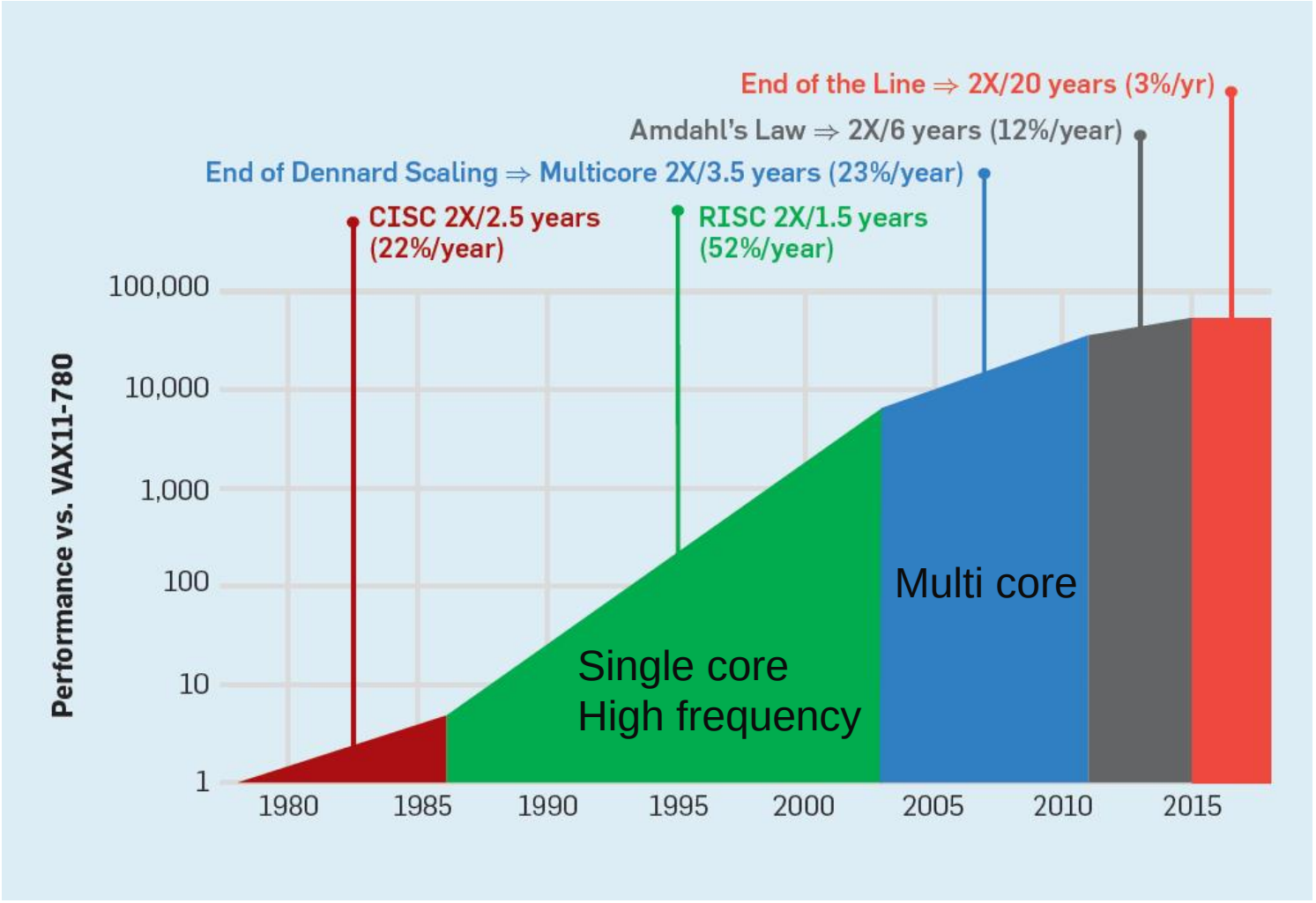
Source: Databricks, Apache Spark Survey 2016, Report

What software developers/users want



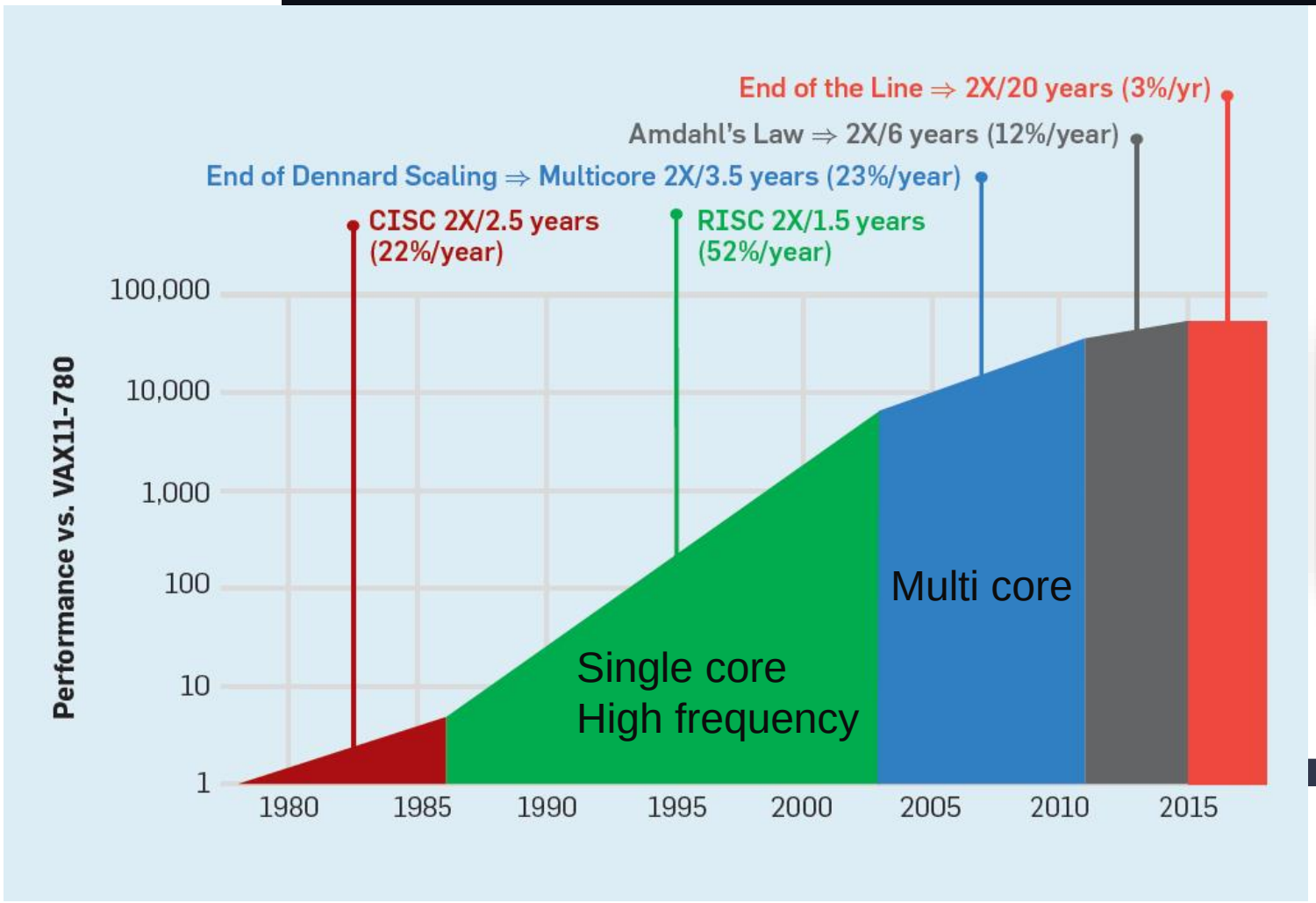
Source: Databricks, Apache Spark Survey 2016, Report

A short history of computing performance



Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018
A domain-specific architecture for deep neural networks
Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson

What's left for faster computing?



What's Left?

Only path left is *Domain Specific Accelerators (DSA)*

- Just do a few tasks, but extremely well

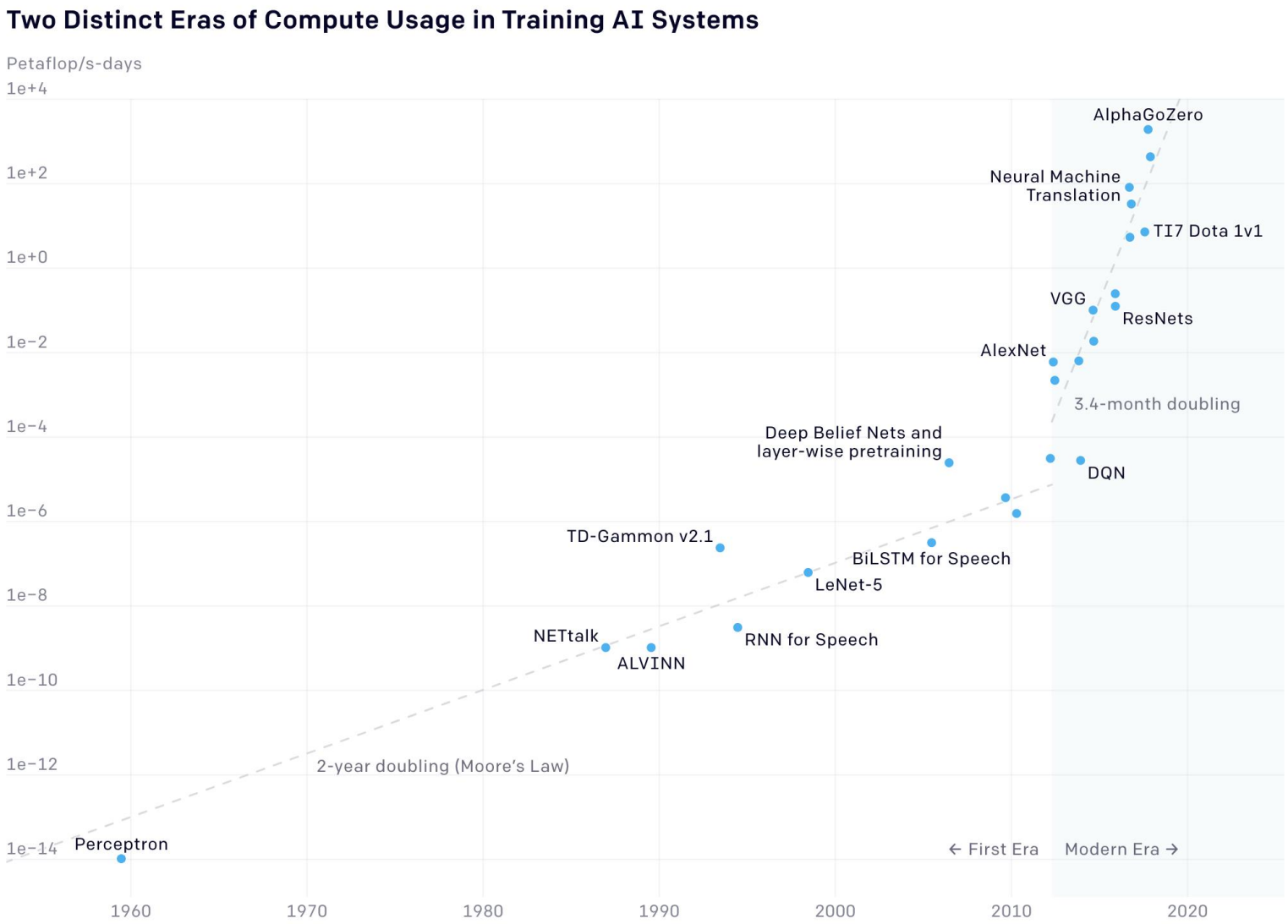
Standard computer for legacy software + accelerators to improve performance per Watt of critical computation

David Patterson, 2019

→ ?

Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018
A domain-specific architecture for deep neural networks
Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson

Computing power to train a model



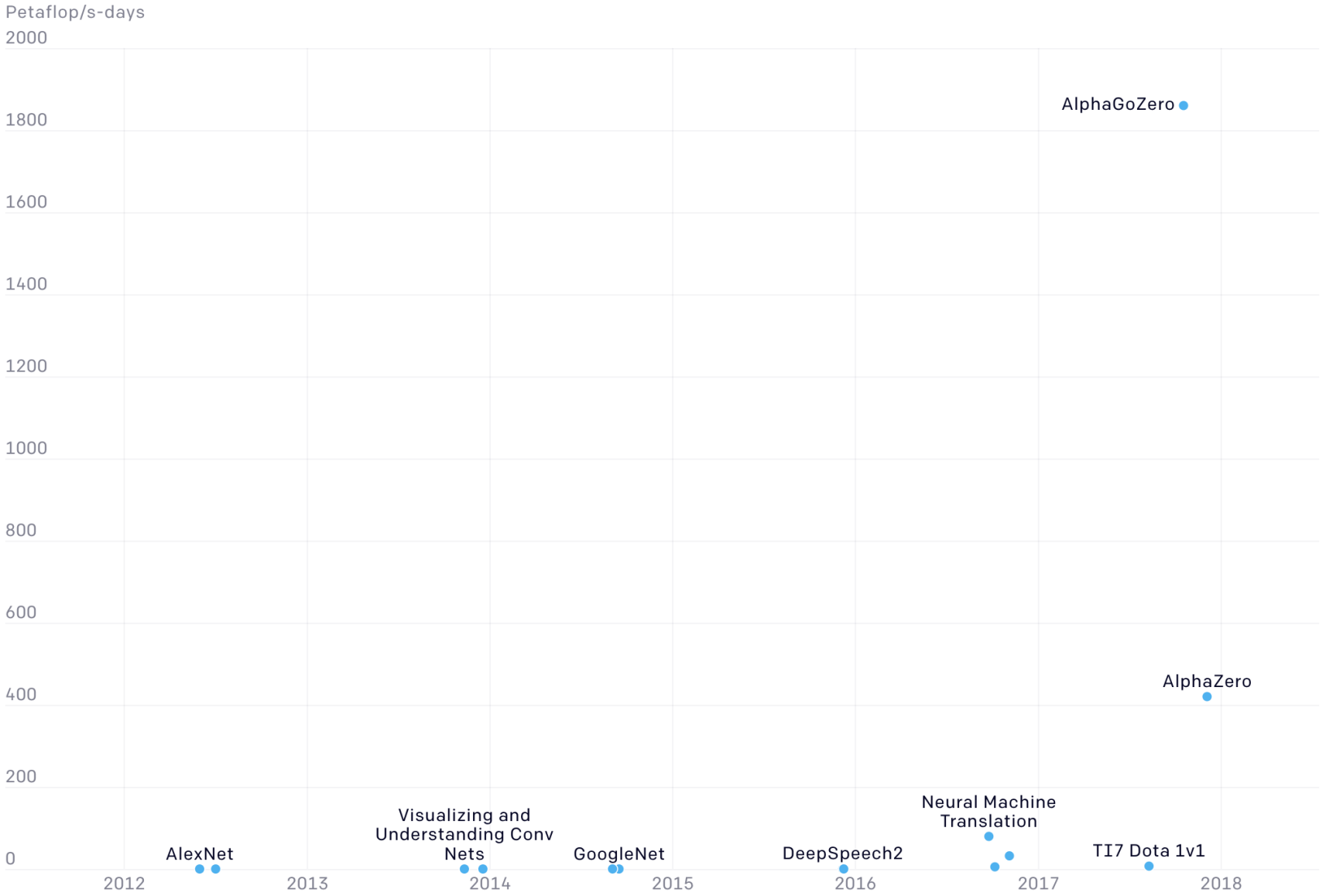
In 2018, OpenAI found that the amount of computational power used to train the largest AI models had doubled every 3.4 months since 2012.

https://www.technologyreview.com/s/614700/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/

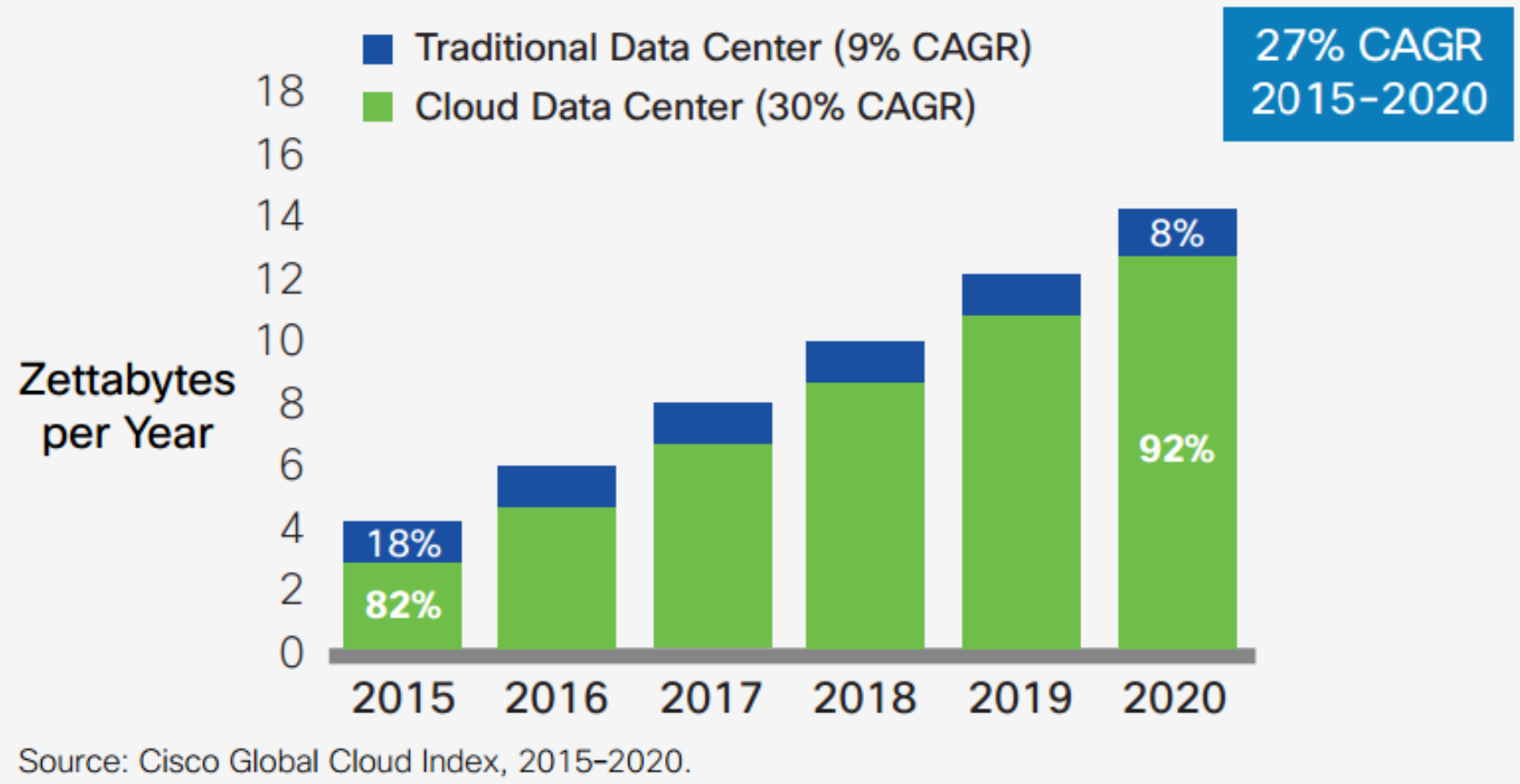
Open AI

https://openai.com/blog/ai-and-compute/#addendum

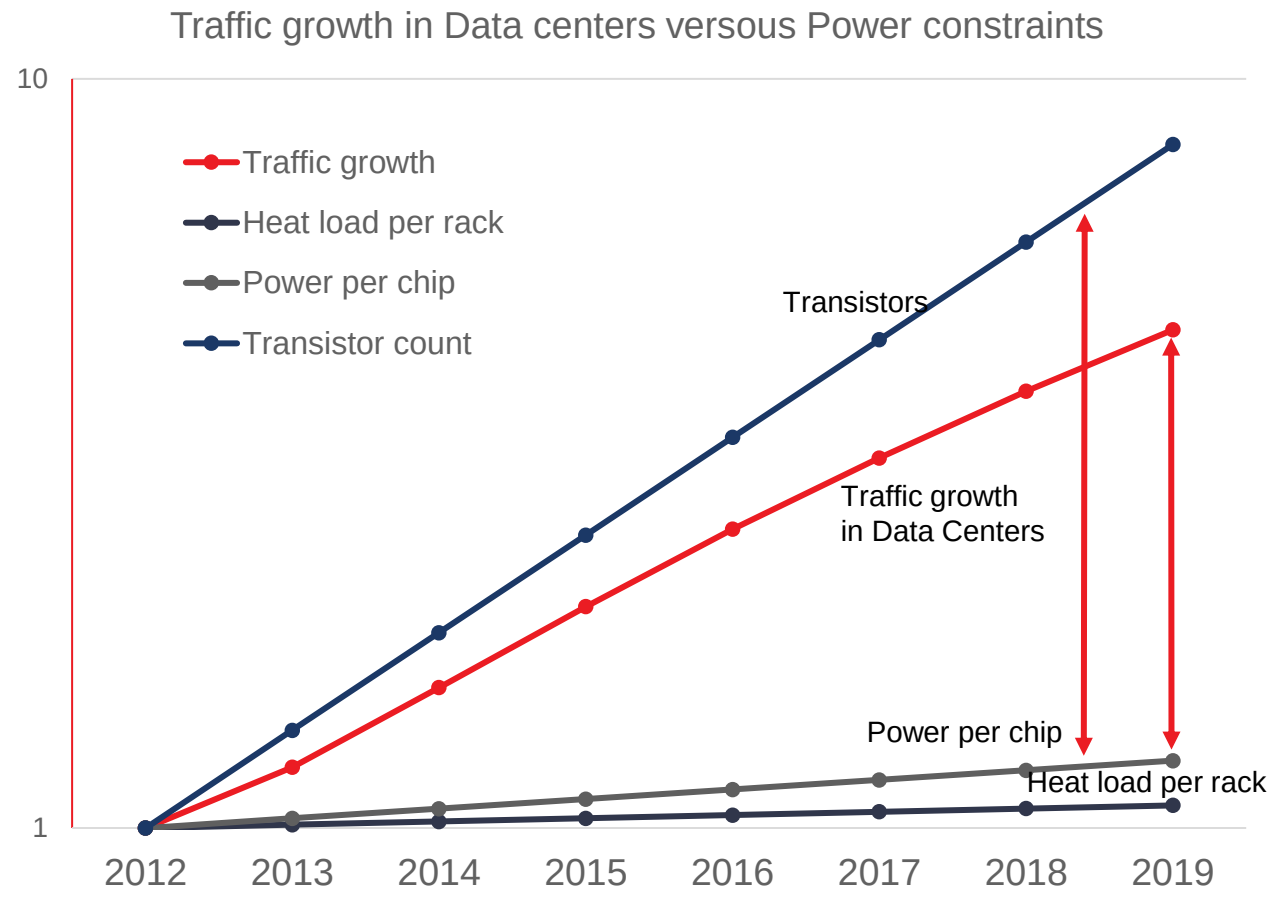
Processing requirements in DNN



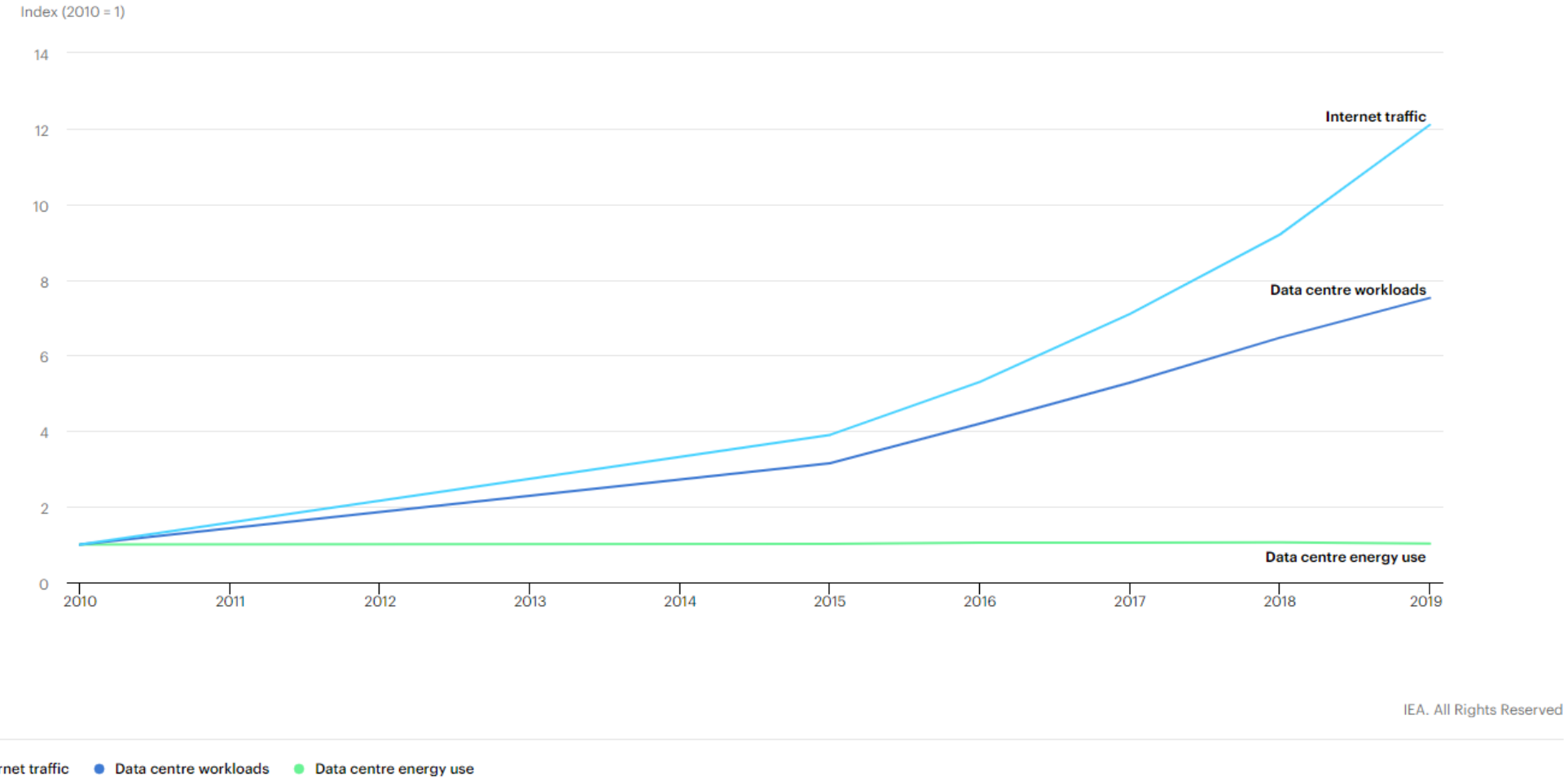
Data Center traffic



Data Center Requirements

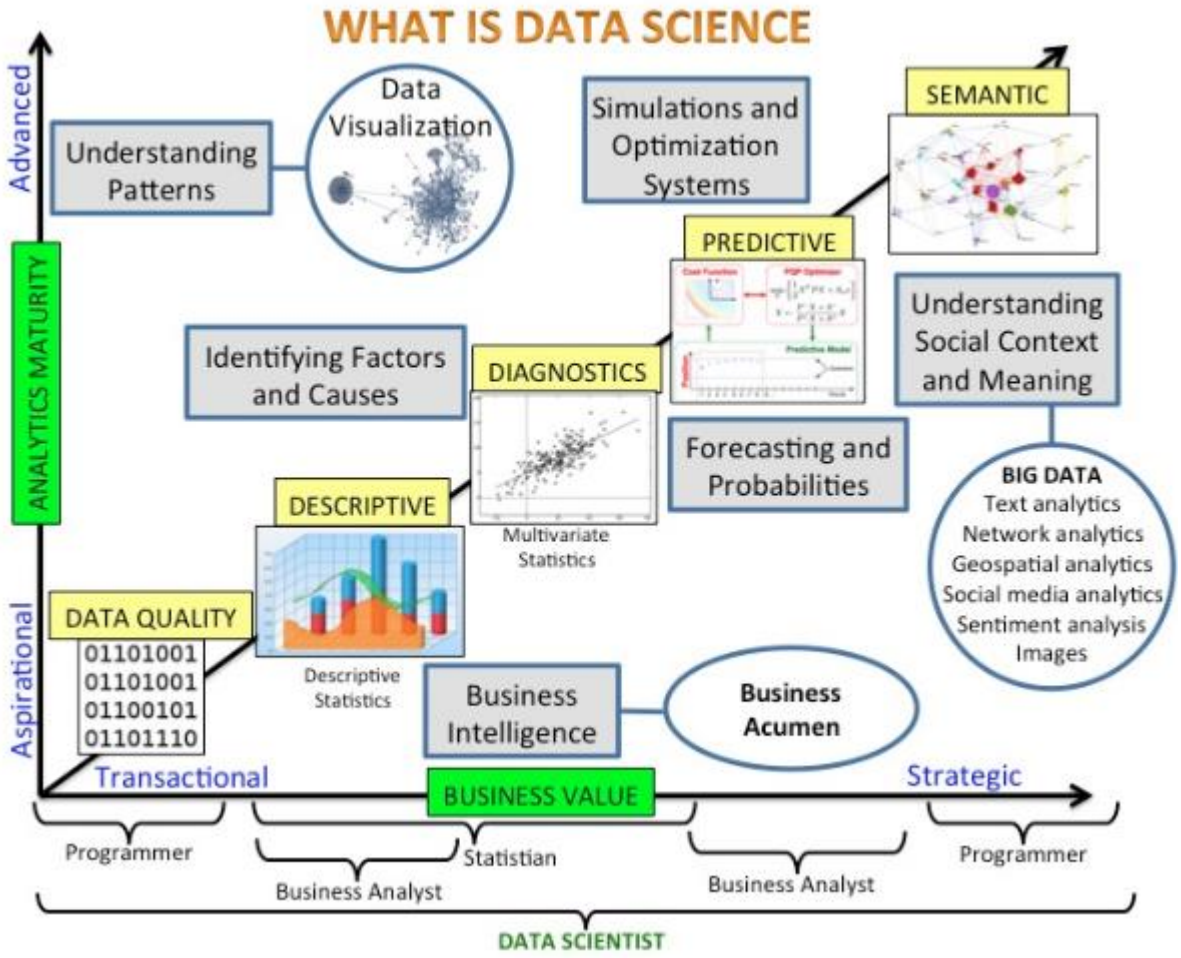


> **Traffic requirements increase significantly in the data centers but the power budget remains the same (Source: ITRS, HiPEAC, Cisco)**

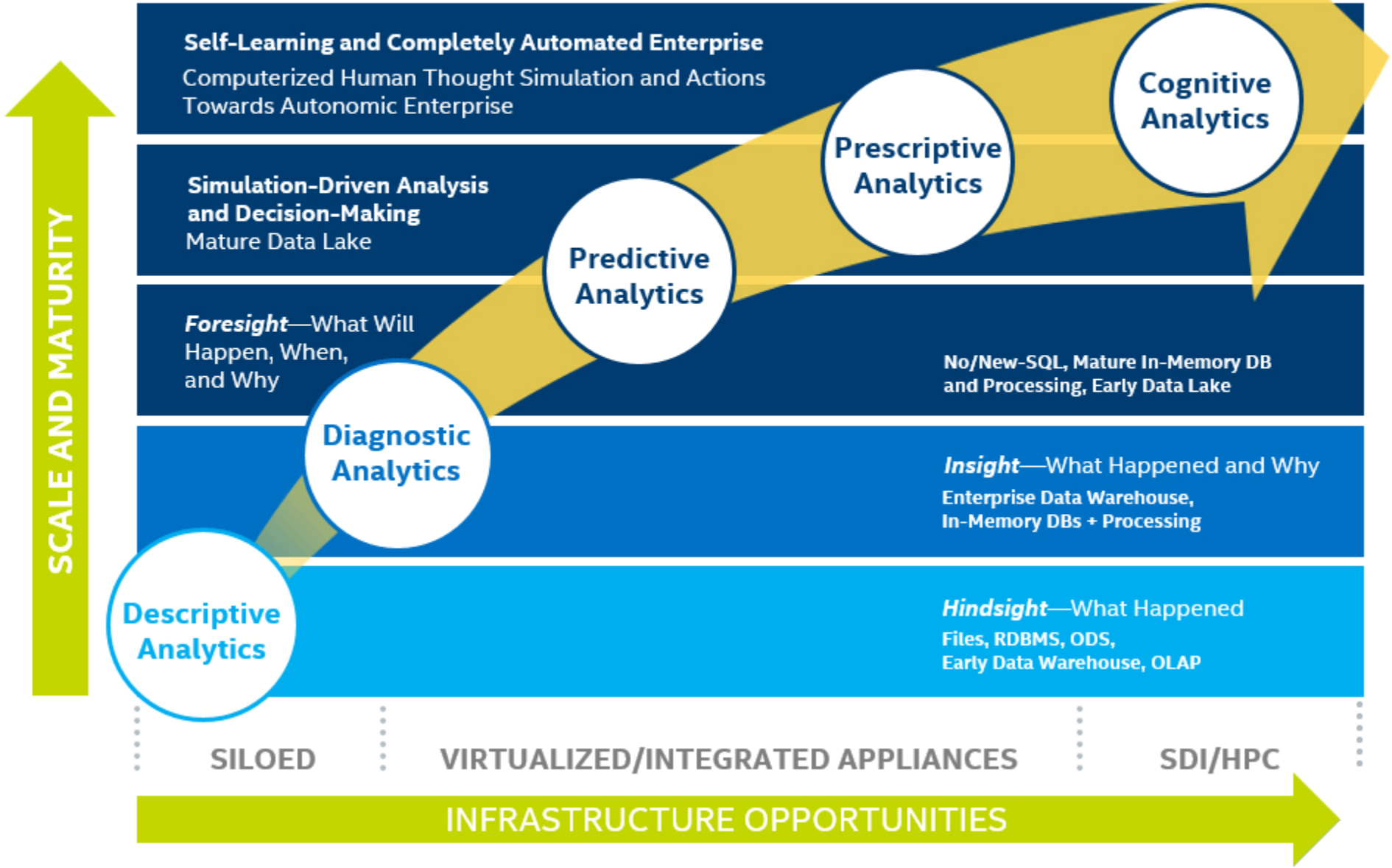


https://www.iea.org/reports/data-centres-and-data-transmission-networks

Data Science: need for high computing power

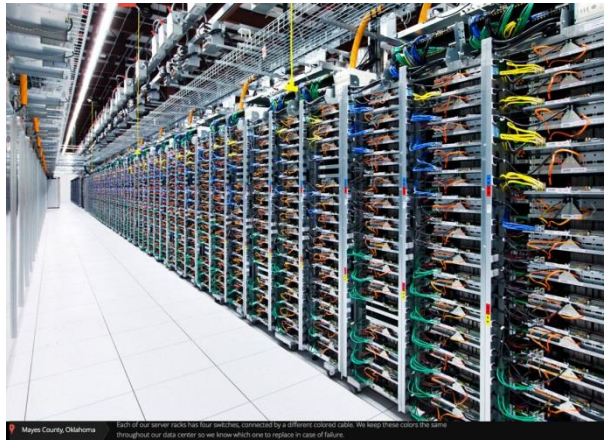


Advanced Analytics Maturity Path:
Moving to Real-Time Enterprise



How Big are Data Centers

Data Center Site	Sq ft
Facebook (Santa Clara)	86,000
Google (South Carolina)	200,000
HP (Atlanta)	200,000
IBM (Colorado)	300,000
Microsoft (Chicago)	700,000



[Source: “How Clean is Your Cloud?”, Greenpeace 2011]

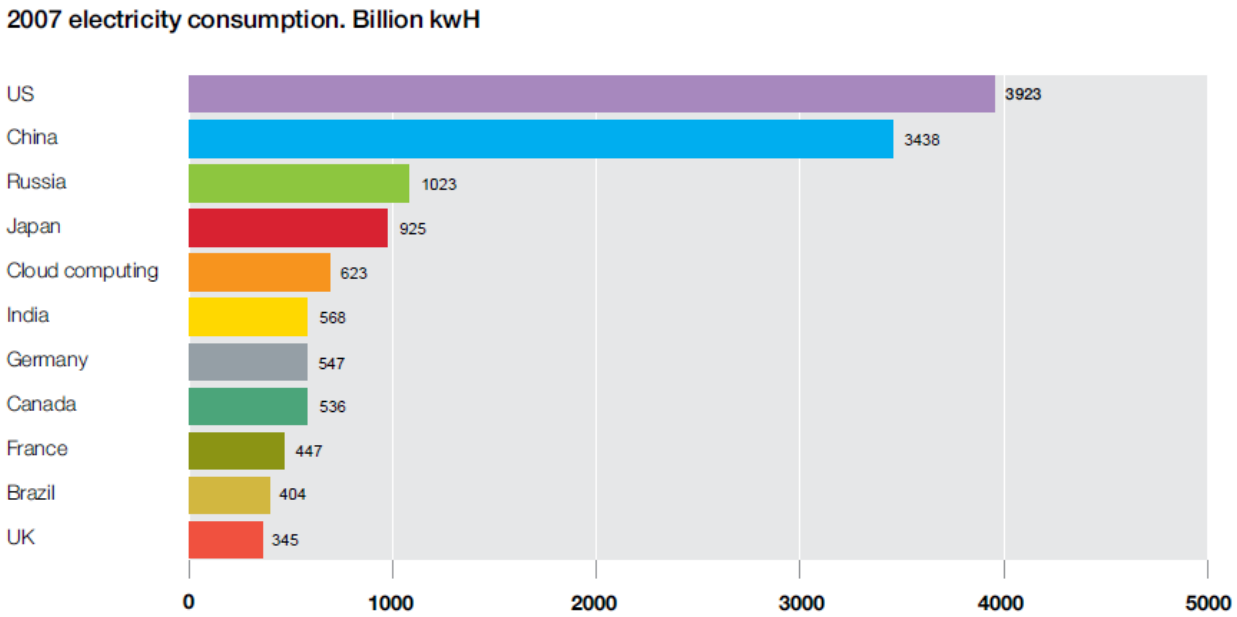
Wembley Stadium:**172,000** square ft

Google data center

Data Centers Power Consumption

- Data centers consumed **330 Billion KWh** in 2007 and is expected to reach **1012 Billion KWh** in 2020

	2007 (Billion KWh)	2020 (Billion KWh)
Data Centers	**330**	**1012**
Telecoms	293	951
Total Cloud	**623**	**1963**



[Source: How Clean is Your Data Center, Greenpeace, 2012

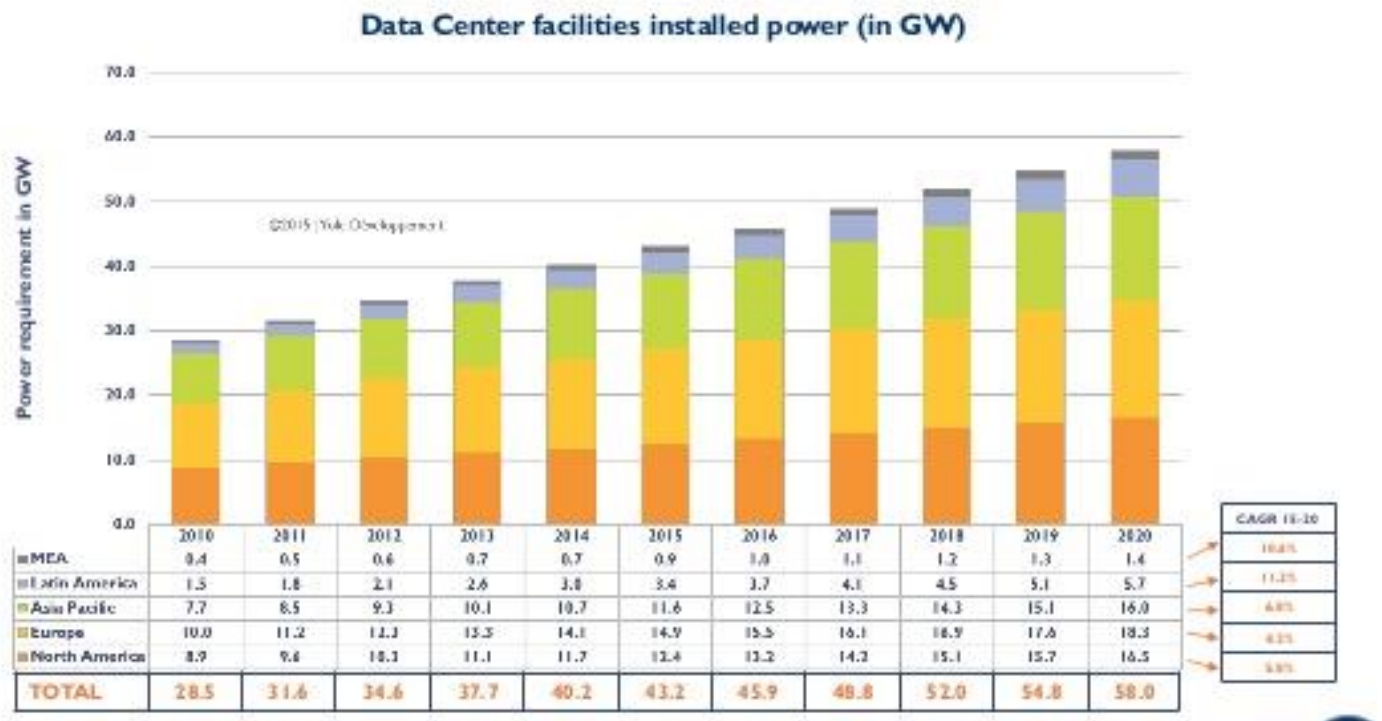
Soon we are going to need a power plant next to the Data Centers

Data Center power consumtion

HOW MUCH OF WORLDWIDE POWER IS NEEDED BY DATA CENTERS?

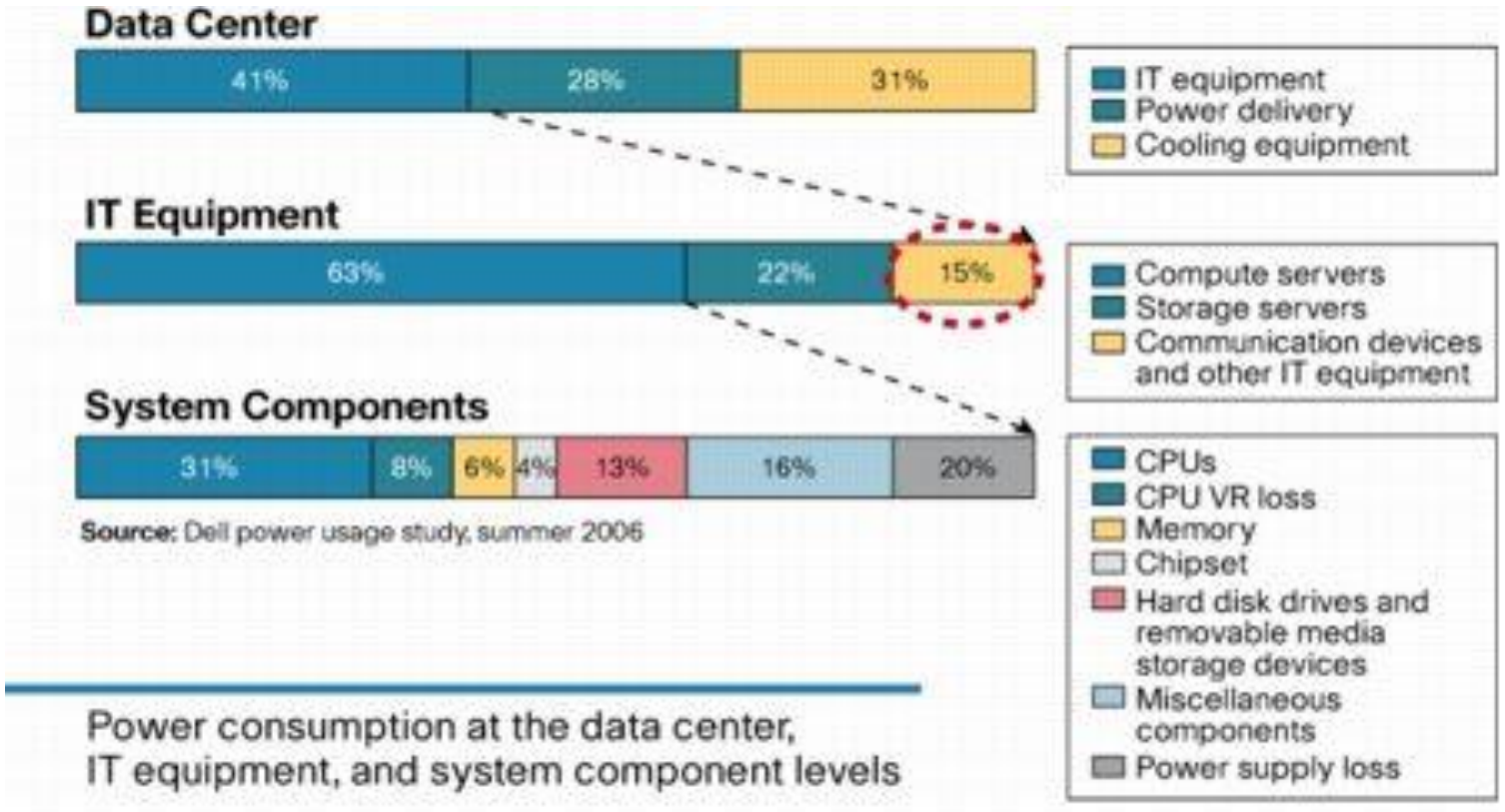
Required power evolution for the period 2010–2020

In 2015, Data Center power needs represent **1.62%** of worldwide production. In 2020, it will be **1.9%**.



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	CAGR 15-20
MEA	0.4	0.5	0.6	0.7	0.7	0.9	1.0	1.1	1.2	1.3	1.4	10.6%
Latin America	1.5	1.8	2.1	2.6	3.0	3.4	3.7	4.1	4.5	5.1	5.7	11.2%
Asia Pacific	7.7	8.5	9.3	10.1	10.7	11.6	12.5	13.3	14.3	15.1	16.0	6.8%
Europe	10.0	11.2	12.3	13.5	14.1	14.9	15.5	16.1	16.9	17.6	18.3	4.3%
North America	8.9	9.6	10.3	11.1	11.7	12.4	13.2	14.2	15.1	15.7	16.5	5.8%
TOTAL	28.5	31.6	34.6	37.7	40.2	43.2	45.9	48.8	52.0	54.8	58.0	

Power consumption



Power consumption at the data center, IT equipment, and system component levels

Hardware acceleration

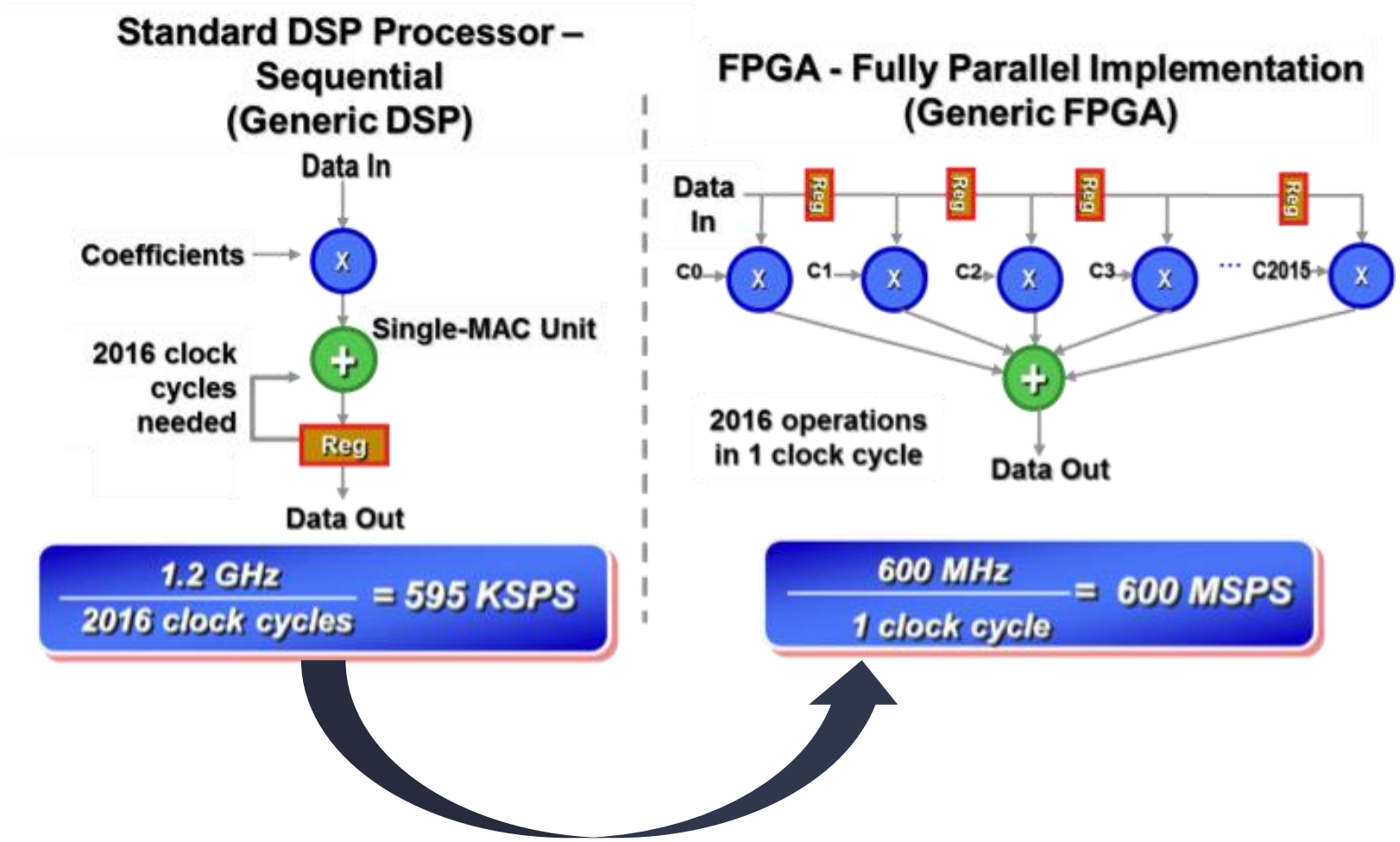
Hardware acceleration is the use of specialized hardware components to perform some functions faster (10x-100x) than is possible in software running on a more general-purpose CPU.

- **Hardware acceleration can be performed either by specialized chips (ASICS) or**
- **By programmable specialized chips (FPGAs) that can be configured for specific applications**

Data Center applications

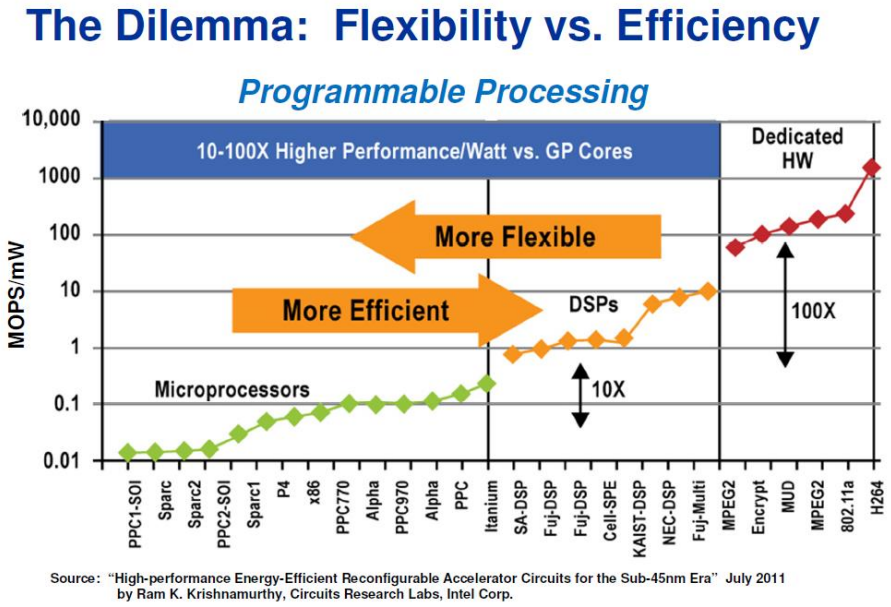
Hardware Accelerators – Why is it faster?

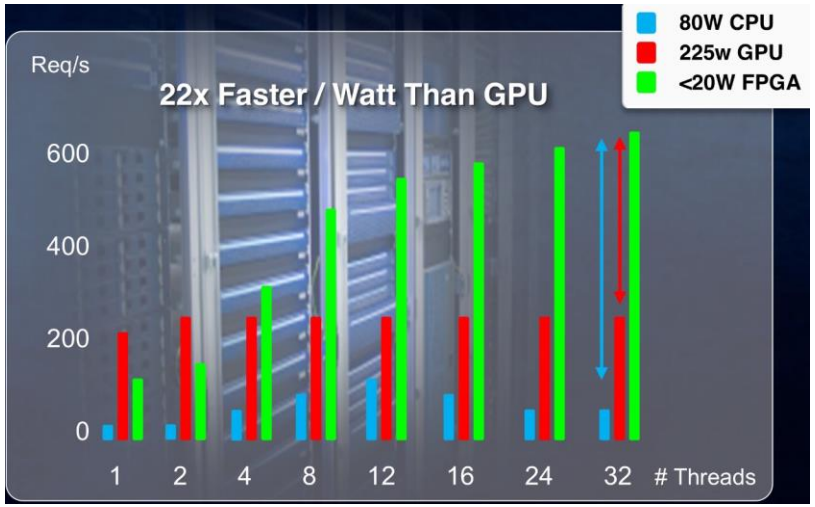
Switch from sequential processing to parallel processing



Hardware accelerators

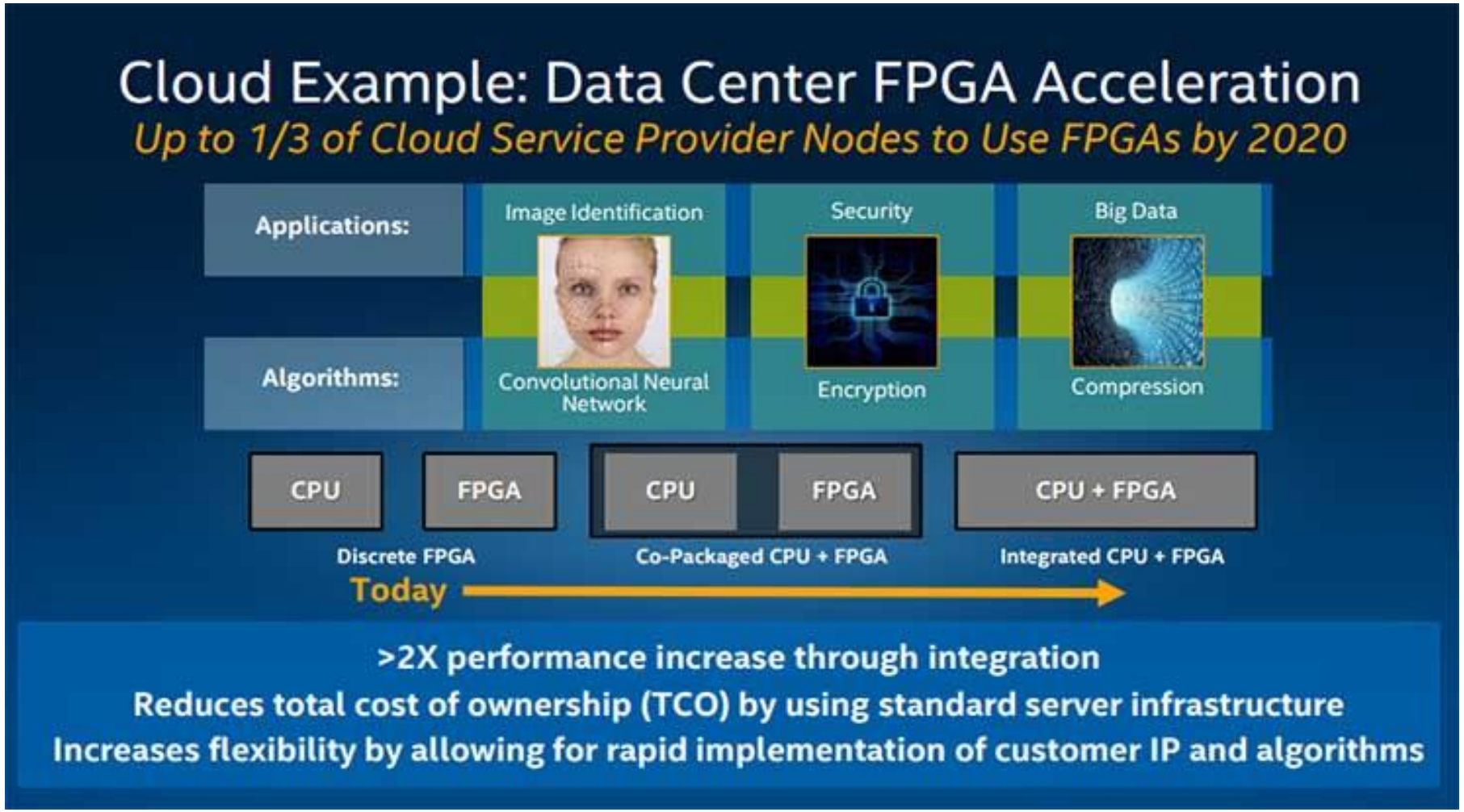
- HW acceleration can be used to reduce significantly the execution time and the energy consumption of several applications (10x-100x)





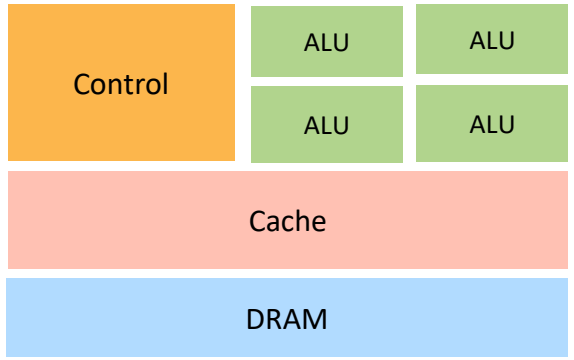
[Source: Xilinx, 2016]

FPGAs in the data centers

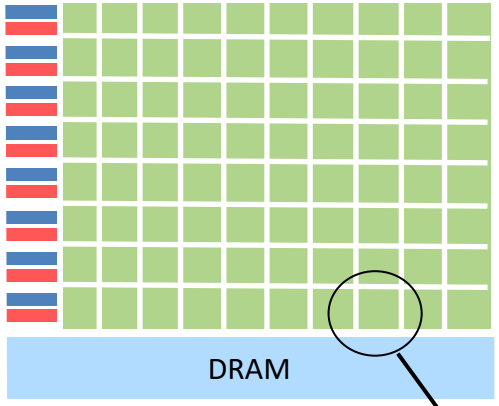


CPU vs GPU vs FPGA

A GPU is effective at processing the same set of operations in parallel – single instruction, multiple data (SIMD).

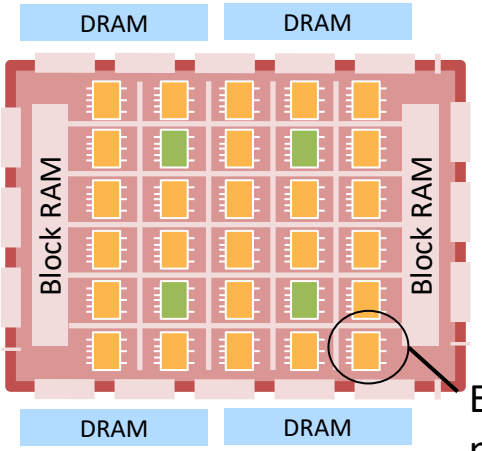


CPU
(one core)



GPU

Each GPU has 2880 of these cores



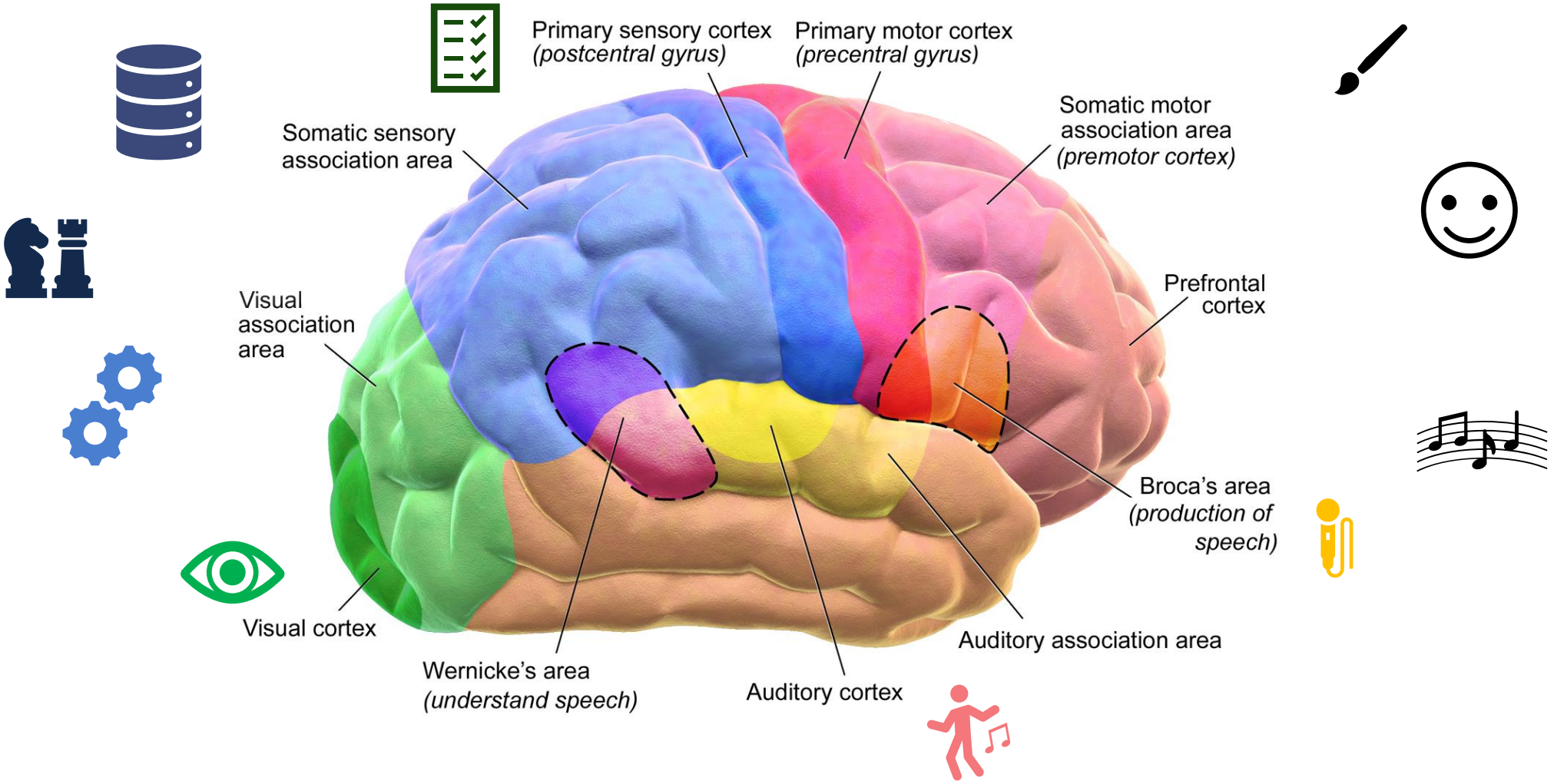
FPGA

Each FPGA has more than 2M of these cells

An FPGA is effective at processing the same or different operations in parallel – multiple instructions, multiple data (MIMD). Specialized circuits for functions.

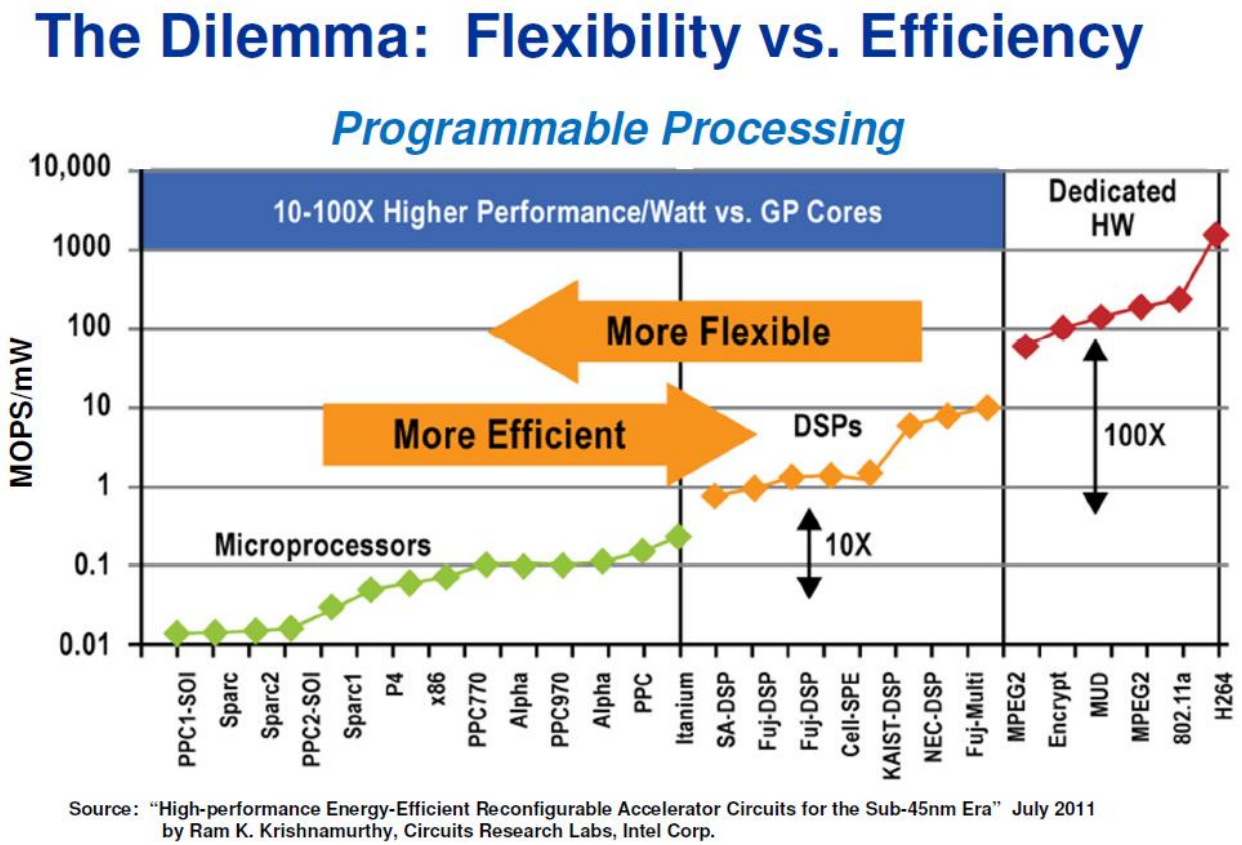
Specialization

One of the most sophisticated systems in the universe is based on specialization



Processing Platforms

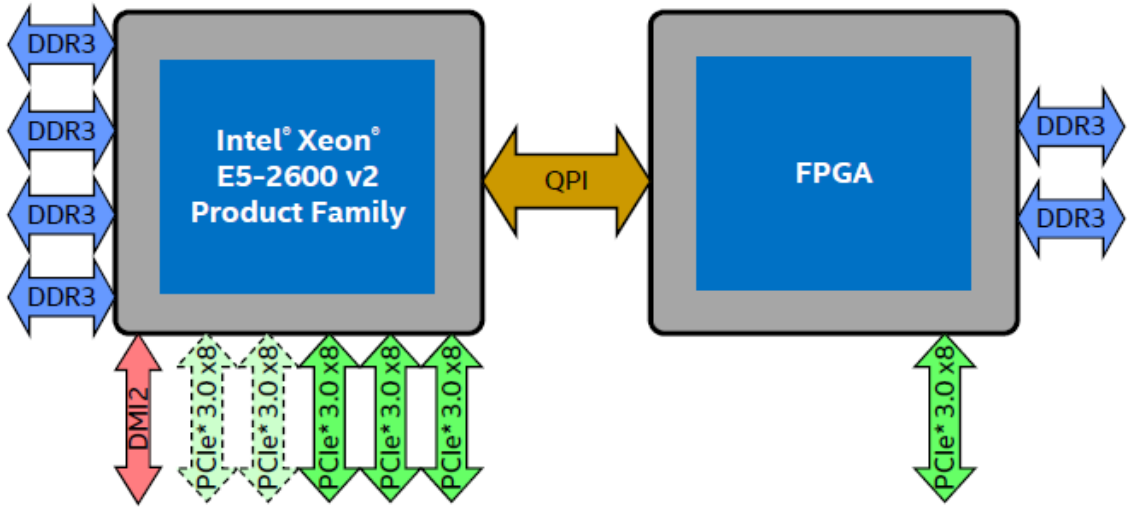
- **HW acceleration can be used to reduce significantly the execution time and the energy consumption of several applications (10x-100x)**



Source: "High-performance Energy-Efficient Reconfigurable Accelerator Circuits for the Sub-45nm Era" July 2011 by Ram K. Krishnamurthy, Circuits Research Labs, Intel Corp.

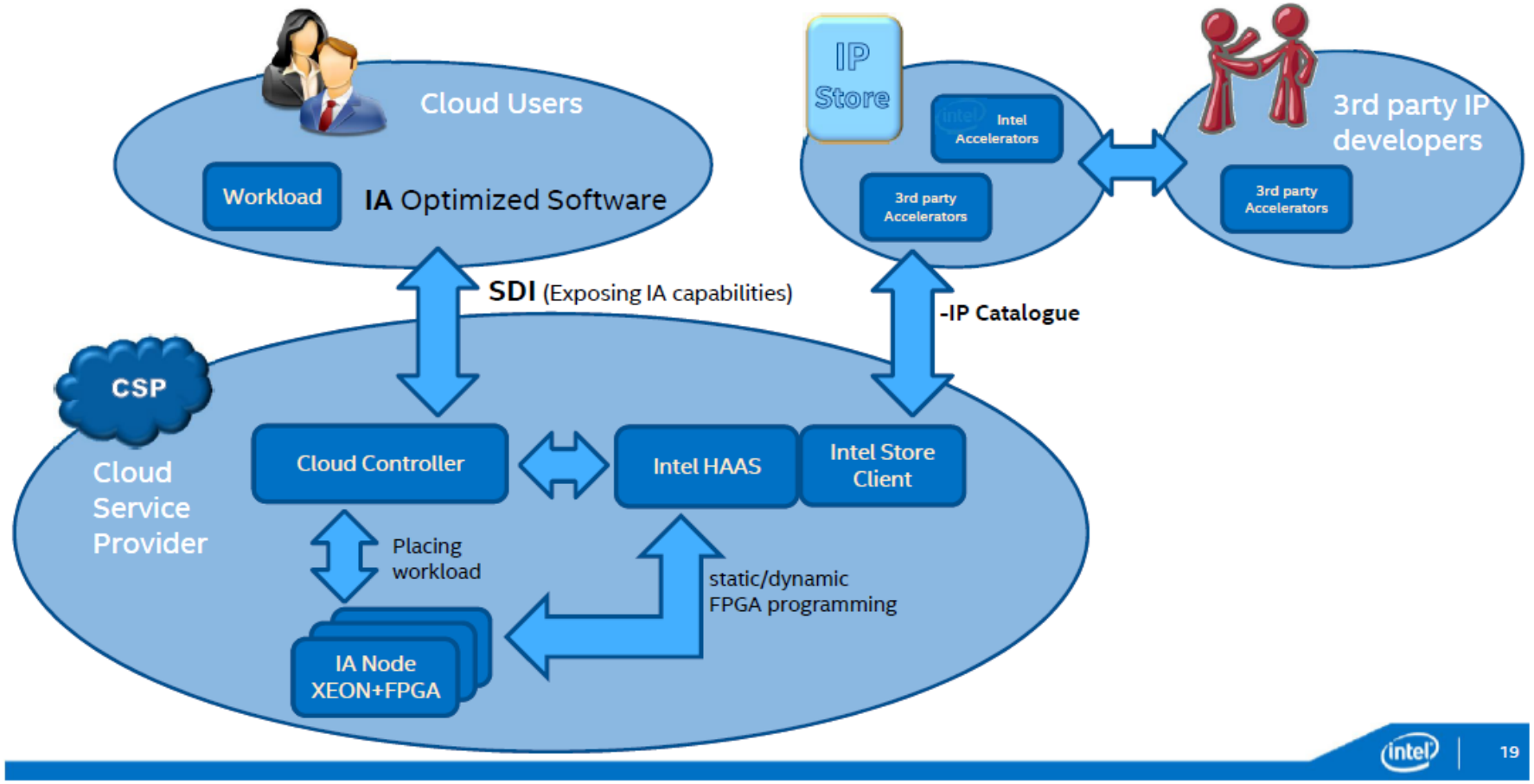
Intel Xeon + FPGAs

Software Development for Accelerating Workloads using Xeon and coherently attached FPGA in-socket



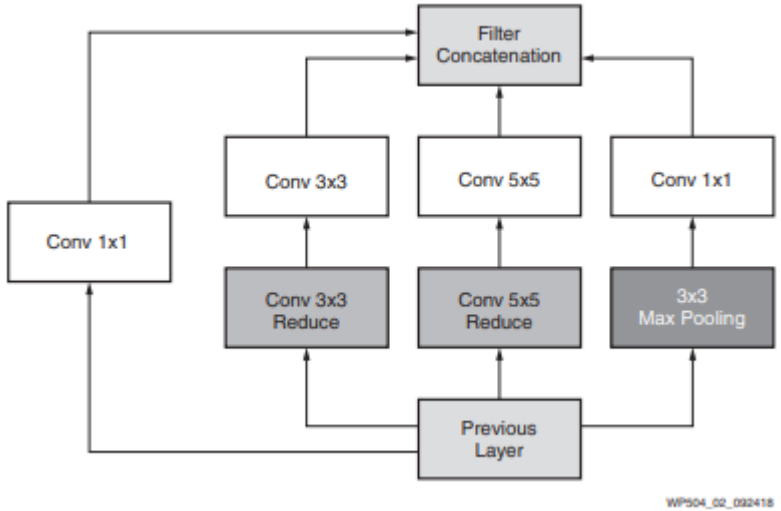
Processor	Intel® Xeon® E5-26xx v2 Processor
FPGA Module	Altera Stratix V
QPI Speed	6.4 GT/s full width (target 8.0 GT/s at full width)
Memory to FPGA Module	2 channels of DDR3 (up to 64 GB)
Expansion connector to FPGA Module	PCIe 3.0 x8 lanes - maybe used for direct I/O e.g. Ethernet
Features	Configuration Agent, Caching Agent,, (optional) Memory Controller
Software	Accelerator Abstraction Layer (AAL) runtime, drivers, sample applications

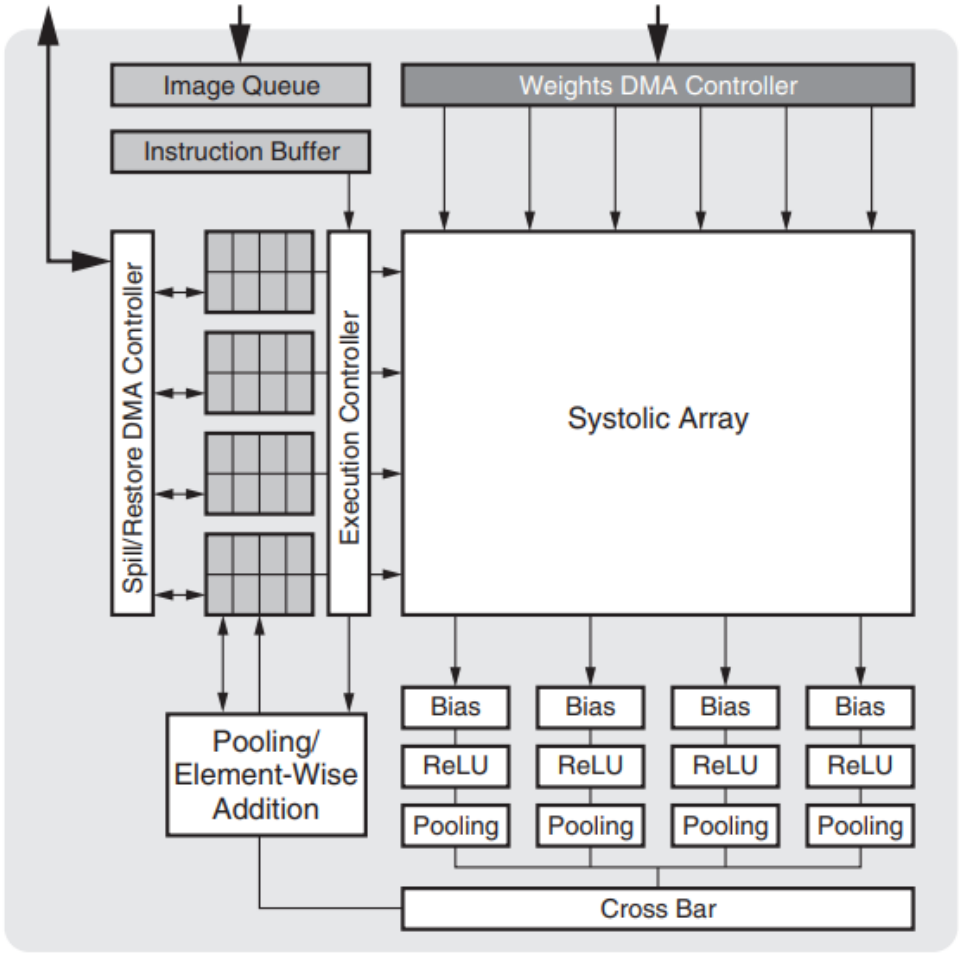
Xeon and FPGA in the Cloud



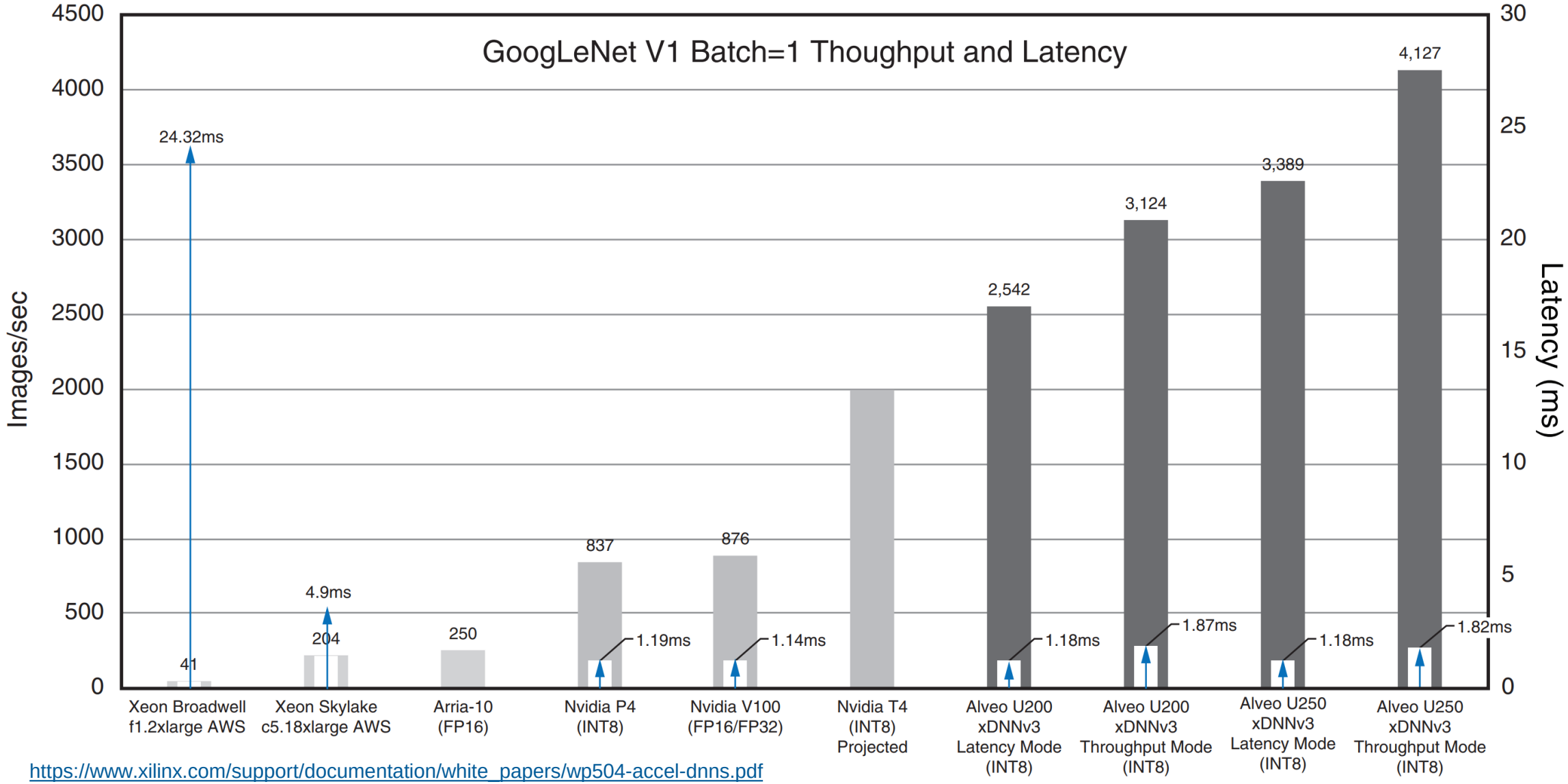
FPGAs for DNN

- The xDNN processing engine has dedicated execution paths for each type of command (download, conv, pooling, element-wise, and upload). This allows fc convolution commands to be run in parallel with other commands if the network graph allows it



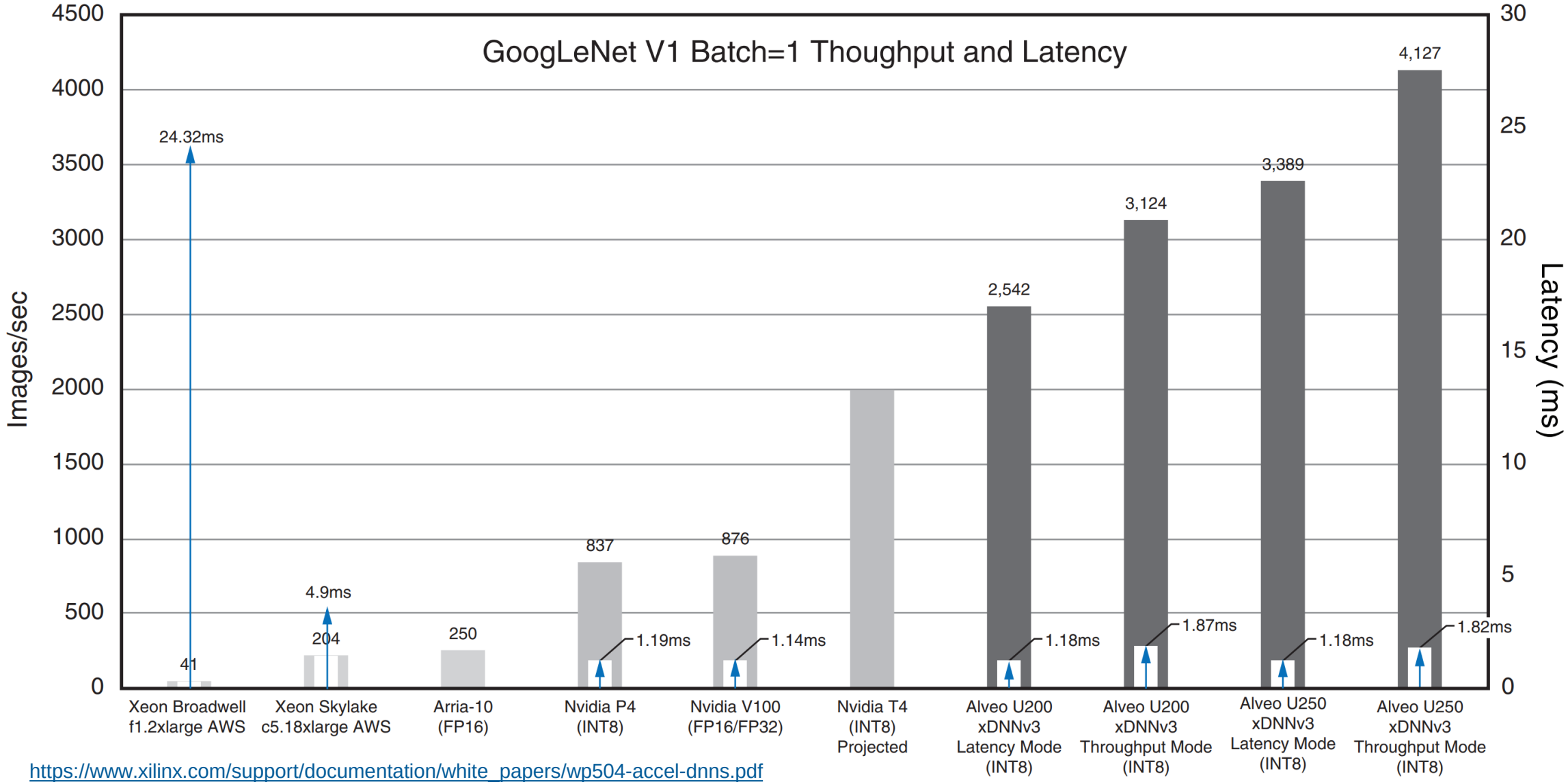


FPGAs for DNN – Throughput & Latency



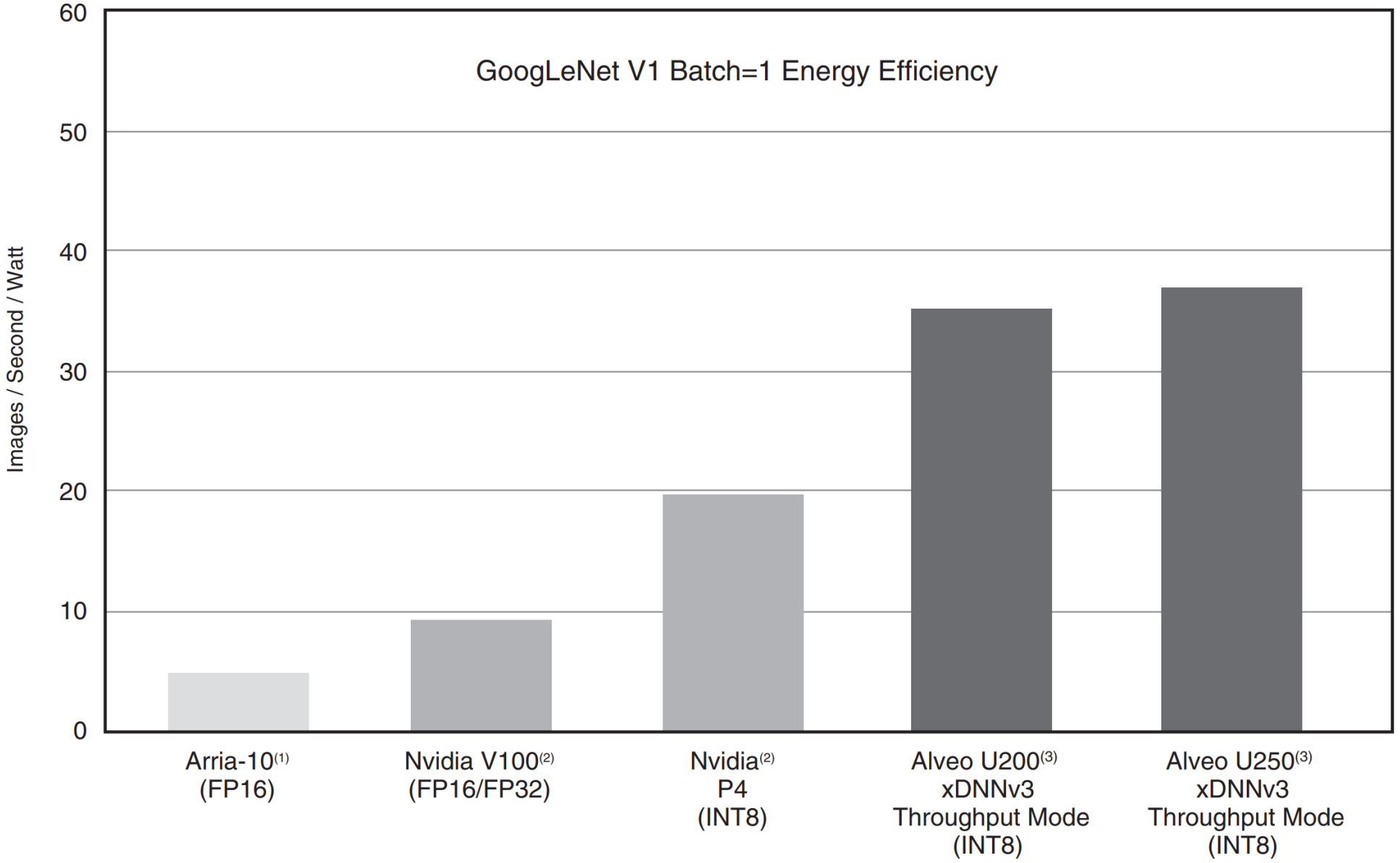
https://www.xilinx.com/support/documentation/white_papers/wp504-accel-dnns.pdf

FPGAs for DNN – Throughput & Latency

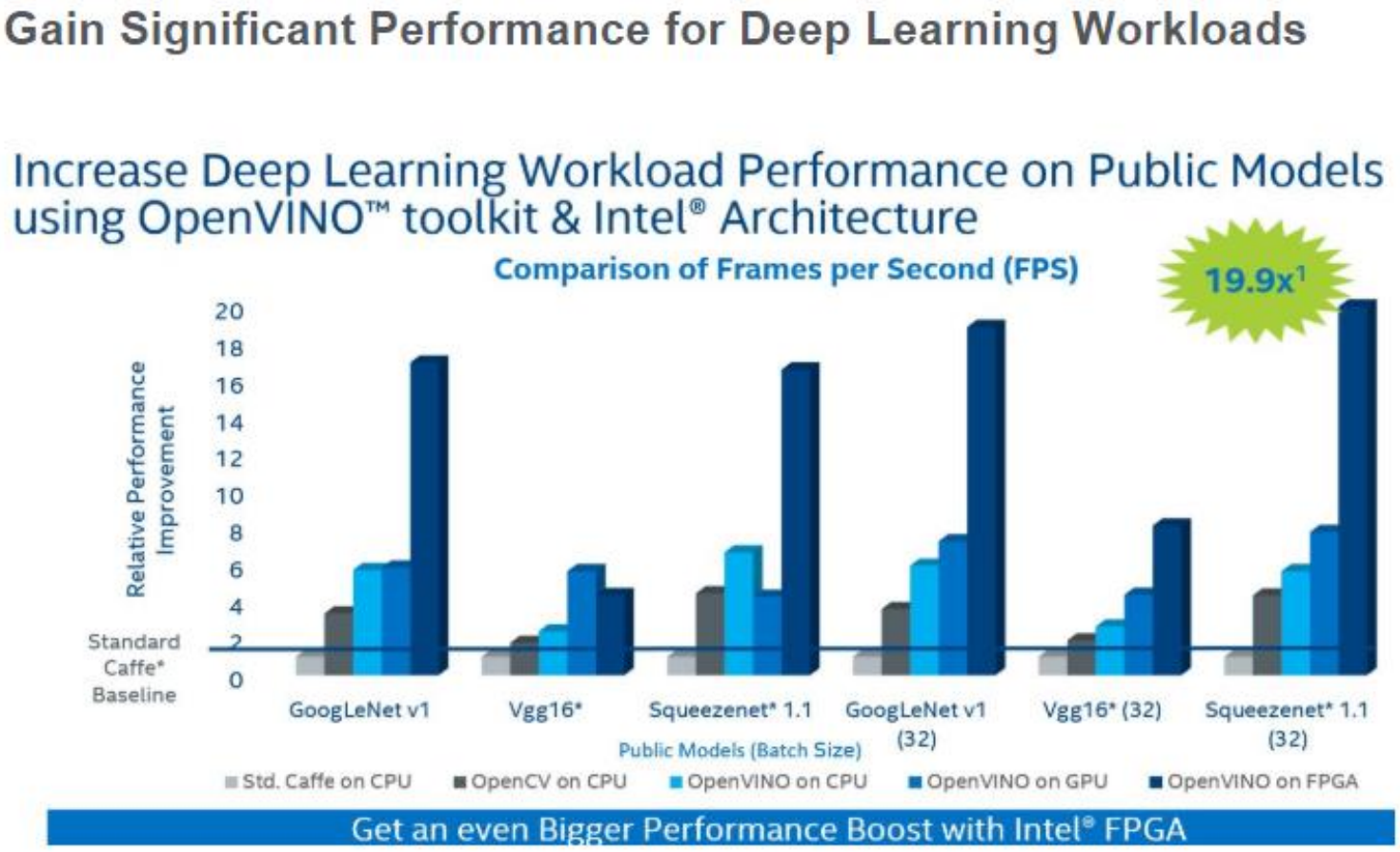


https://www.xilinx.com/support/documentation/white_papers/wp504-accel-dnns.pdf

FPGAs for DNN – Energy efficiency



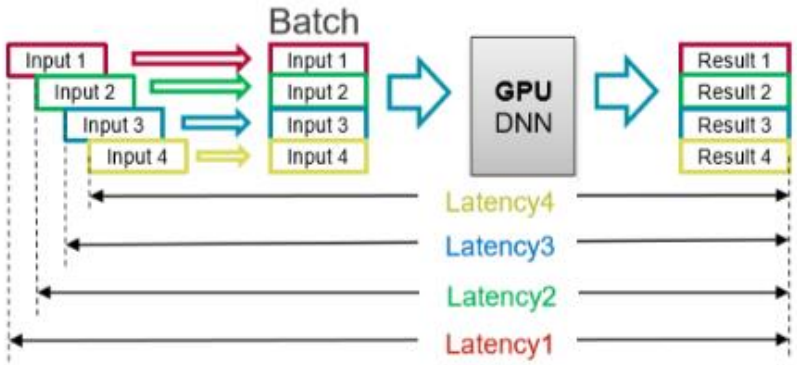
Intel FPGAs for DNN

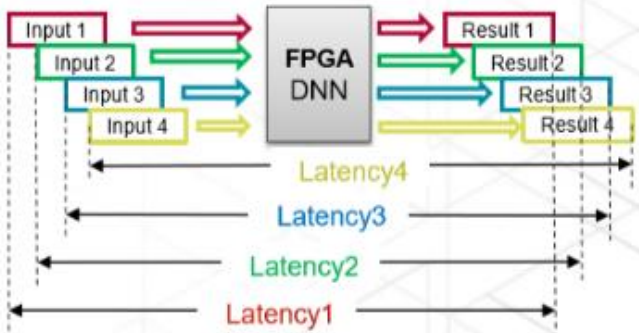


https://software.intel.com/content/www/us/en/develop/blogs/accelerate-computer-vision-from-edge-to-cloud-with-openvino-toolkit.html

FPGAs vs GPUs in DNN

FPGA Benefits: Low Latency, High Throughput





- Inference with batches
  - Require parallel batch of data for SIMD
  - High batch => high latency, higher throughput
  - Lower compute efficiency at low batch
- "Batch-less" inference
  - Low and deterministic latency
  - High throughput regardless of batch size
  - Consistent compute efficiency

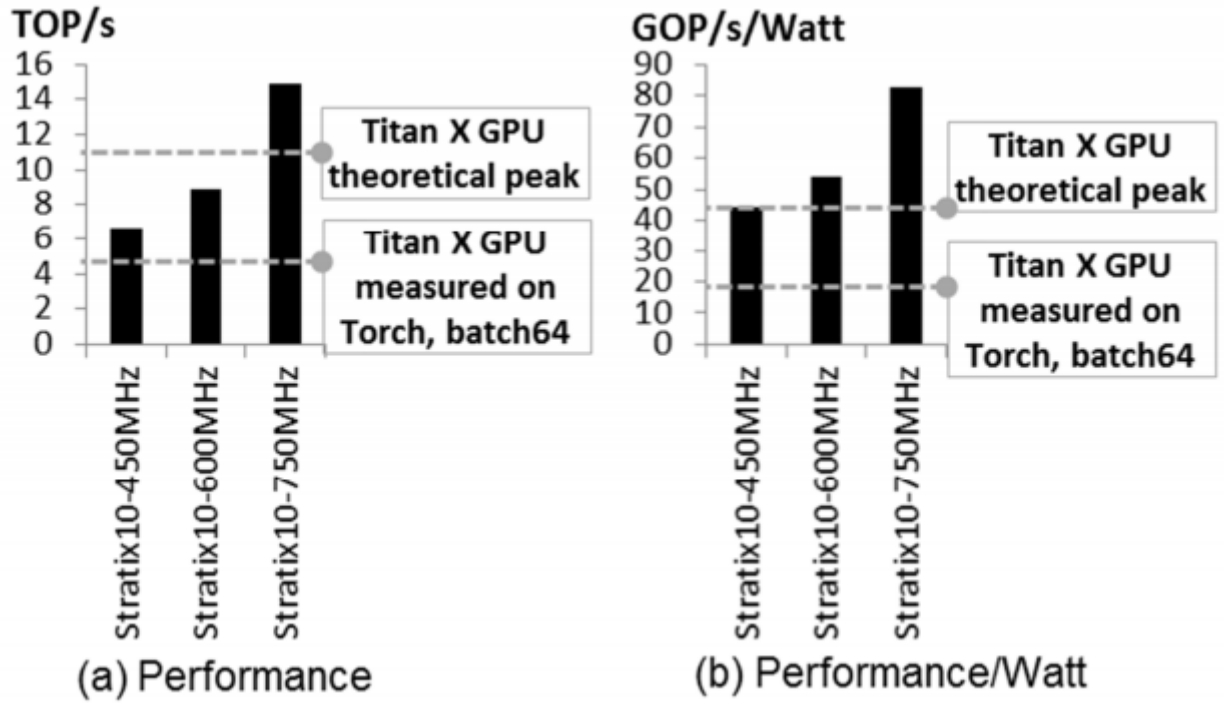
Customers, from edge to Cloud, require low latency inference (batch=1)

GPU vs FPGA for DNN

Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?

Eriko Nurvitadhi[1], Ganesh Venkatesh[1], Jaewoong Sim[1], Debbie Marr[1],
Randy Huang[2], Jason Gee Hock Ong[2], Yeong Tat Liew[2],
Krishnan Srivatsan[3], Duncan Moss[3], Suchit Subhaschandra[3], Guy Boudoukh[4]

[1]Accelerator Architecture Lab, [2]Programmable Solutions Group, [3]FPGA Product Team, [4]Computer Vision Group
Intel Corporation



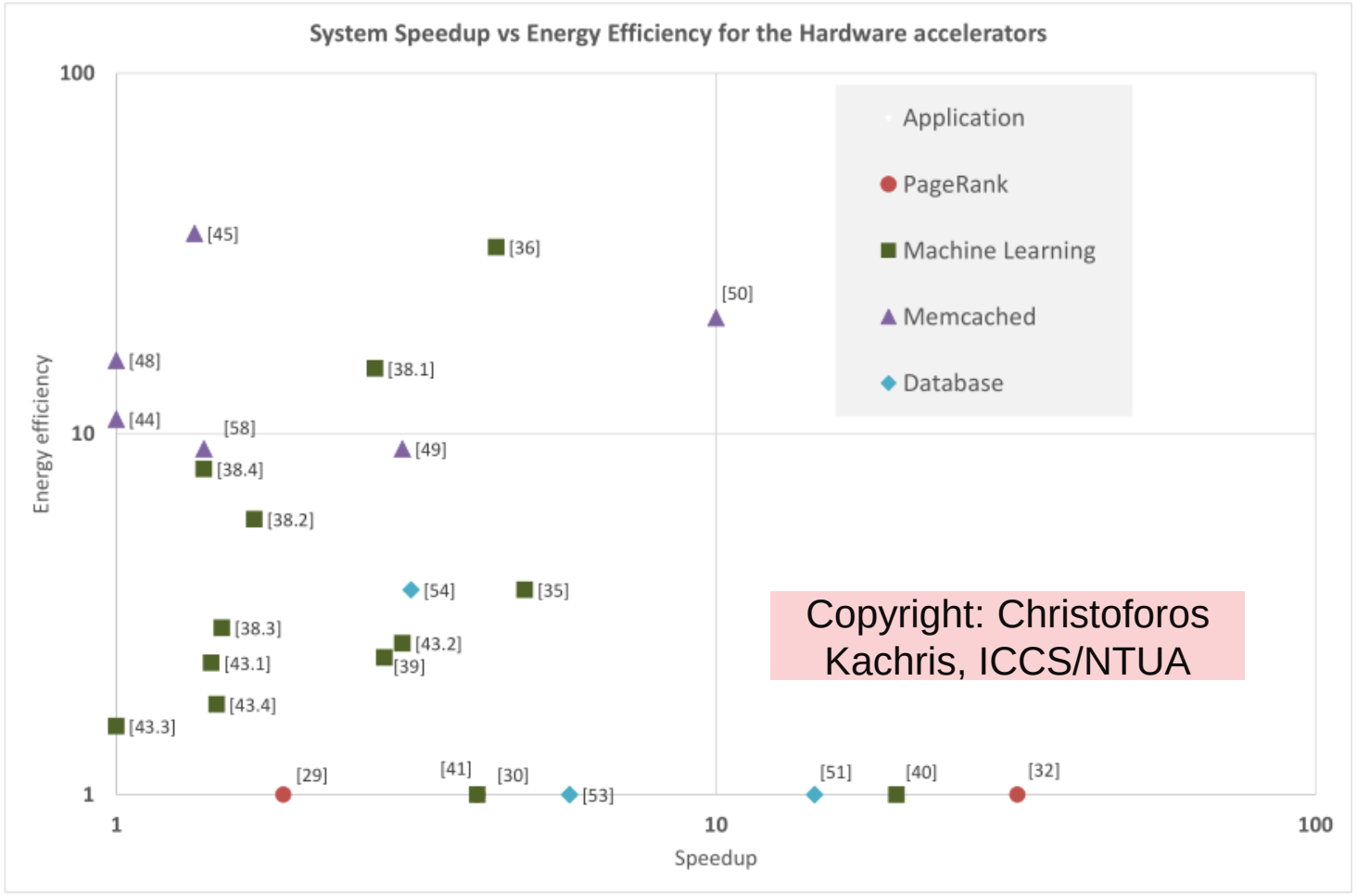
HW Accelerators for Cloud Computing

A Survey on Reconfigurable Accelerators for Cloud Computing, FPL 2016 Kachris

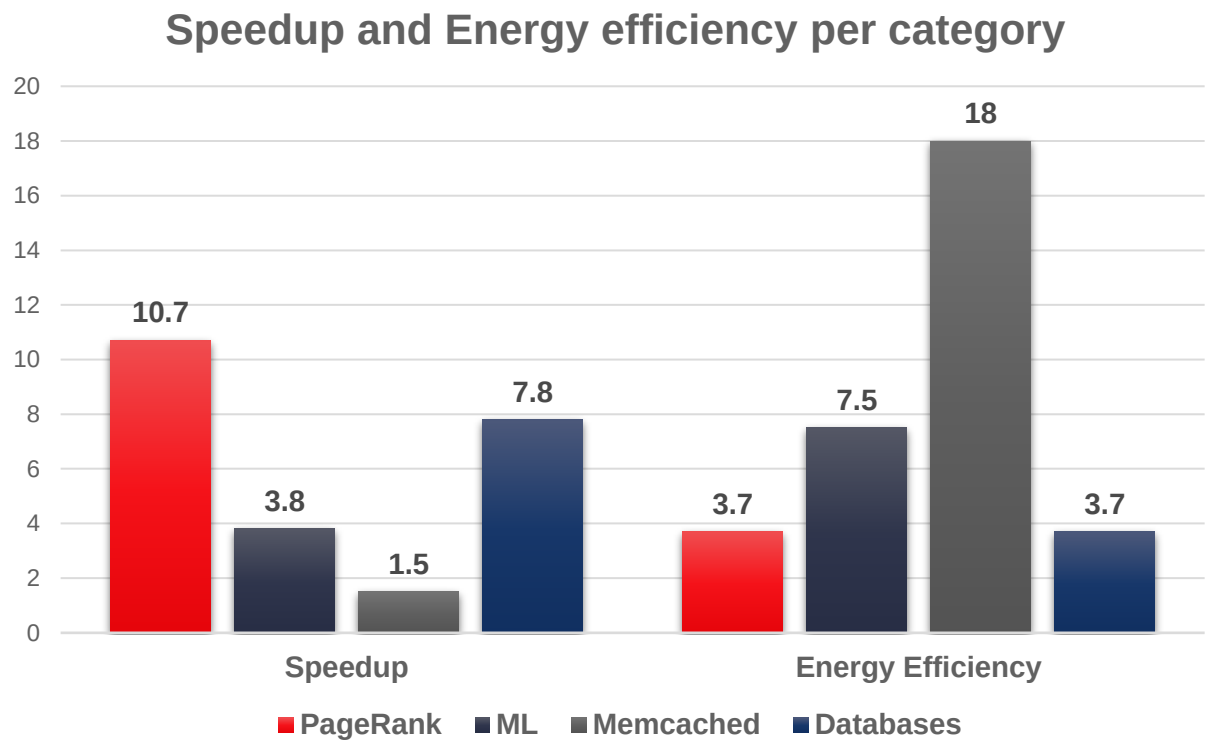
Paper	Institute	Application	Type		Speedup	Energy	Interface	Design	Integration
			Batch	Stream					
[29]	Microsoft	Search engine		•	1.95x	–	PCIe	HDL	Coprocessor
[30]	NUDT	RankBoost (MapReduce)	•		4x	–	Ethernet	HDL	Coprocessor
[32]	TU, Microsoft	RankBoost (MapReduce)	•		31.8x	–	PCIe	HLL	Coprocessor
[35]	NTT	MapReduce (Sort, Grep)	•		4.8x	3.7x	PCIe	C	Coprocessor
[36]	DUTh, NTUA	ML (average)	•		4.3x	33x	AXI4	HDL-HLL	Coprocessor
[38].1	GMU, UCLA	ML (K-Means)	•		2.7x	15.2x	AXI4	HLL	Coprocessor
[38].2	GMU, UCLA	ML (KNN)	•		1.7x	5.8x	AXI4	HLL	Coprocessor
[38].3	GMU, UCLA	ML (SVM)	•		1.5x	2.9x	AXI4	HLL	Coprocessor
[38].4	GMU, UCLA	ML (Naive Bayes)	•		1.4x	8x	AXI4	HLL	Coprocessor
[40]	HKU	ML (K-Means,MapReduce)	•		20x	–	PCIe	HDL	Coprocessor
[39]	UCLA	DNA Sequencing	•		2.8	2.4x	PCIe	HLL	Coprocessor
[41]	Toronto U	ML (K-Means - Spark)	•		4x	–	PCIe	HLL	Coprocessor
[43].1	UCLA-8Zynq	ML (K-Means - Spark)	•		1.44x	2.32x	Ethernet	HLL	Coprocessor
[43].2	UCLA-Virtex7	ML (K-Means - Spark)	•		3x	2.63x	Ethernet	HLL	Coprocessor
[43].3	UCLA-8Zynq	ML (LogRegr.- Spark)	•		1x	1.55x	PCIe	HLL	Coprocessor
[43].4	UCLA-Virtex7	ML (LogRegr.- Spark)	•		1.47x	1.78x	PCIe	HLL	Coprocessor
[44]	HP, UML	Memcached		•	1x	10.9x	Ethernet	HDL	Standalone
[45]	Xilinx	Memcached		•	1.35x	36x	Ethernet	HDL	Standalone
[48]	HP, ARM, Facebook	Memcached		•	0.7x	16x	Custom	HDL	Coprocessor
[49]	UTAustin	Memcached		•	3x	9.15x	Custom	HDL	Coprocessor
[58]	Berkeley	Memcached		•	1.4x	–	PCIe	HDL	Coprocessor
[50]	AlgoLogic	Memcached		•	10x	21x	PCIe	HDL	Coprocessor
[51]	IBM	Database		•	14.6x	–	PCIe	HDL	Coprocessor
[53]	Stanford	Database		•	5.7x	–	PCIe	OpenSPL	Coprocessor
[54]	EPFL,HP,UE,Google	Database		•	3.1x	3.7x	Custom	HDL	Coprocessor

Speedup vs Energy efficiency

A Survey on Reconfigurable Accelerators for Cloud Computing, FPL 2016 Kachris



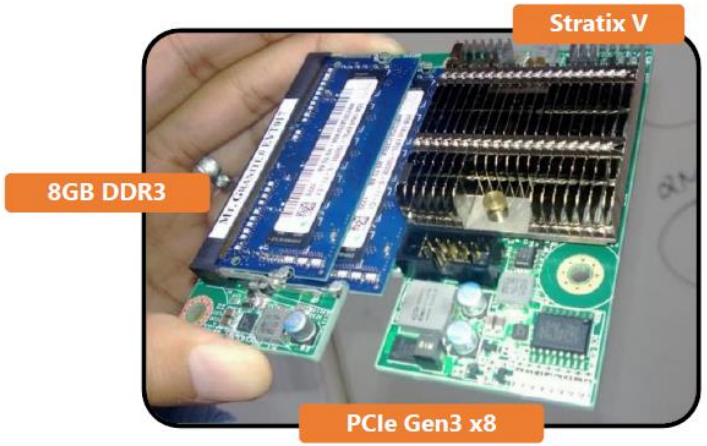
Speedup per category



- **Page Rank applications achieve the higher speedup**
- **Memcached application achieve higher energy efficiency**

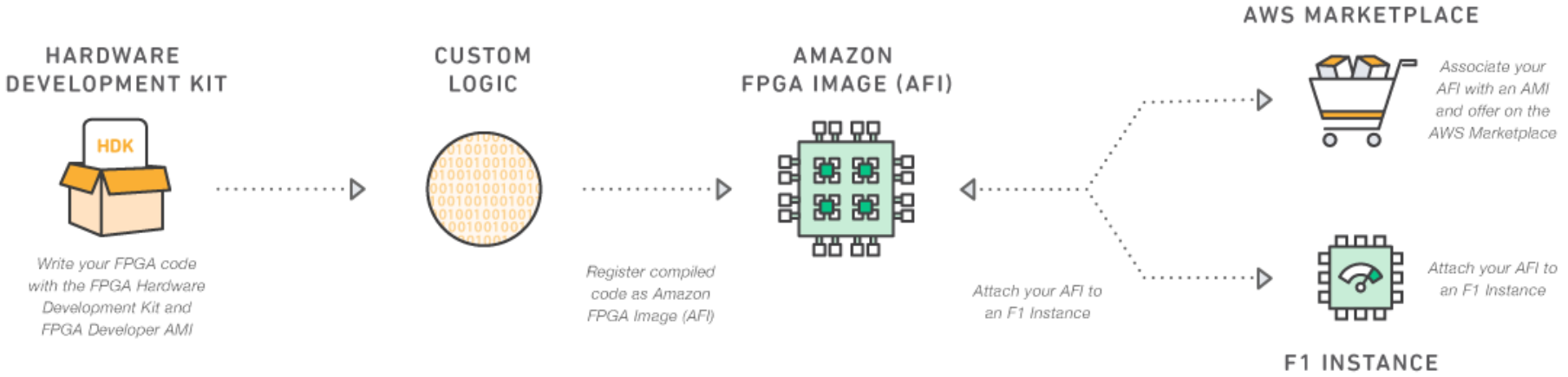
Catapult FPGA Acceleration Card

- Altera Stratix V D5
- 172,600 ALMs, 2,014 M20Ks, 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network



FPGA as a Service

- Amazon EC F1's Xilinx FPGA



Is there a market?

Official At Last: Intel Completes $16.7 Billion Buy of Altera

Up to 1/3 of Cloud Service Provider Nodes to Use FPGAs by 2020 — Intel

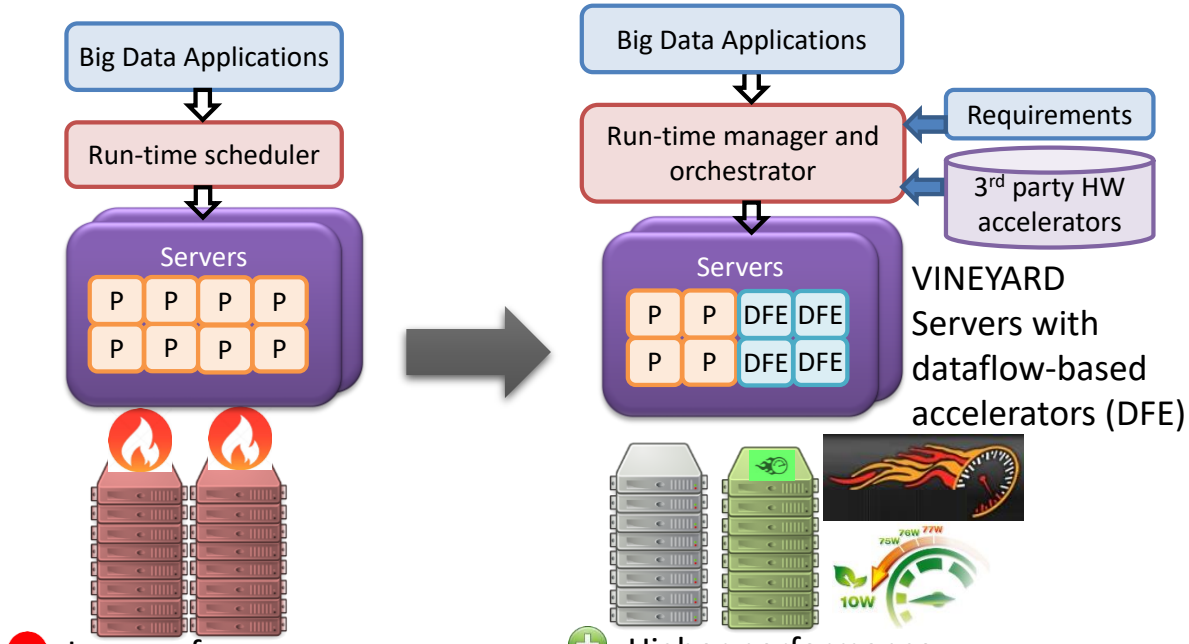
Available FPGAs

Microsoft's Bing search engine uses FPGA chips to provide more intelligent answers

- The global **Data Center Accelerator market** size is expected to reach **35 billion $ by the end of 2025** [1].
- The market for FPGA is expected to grow at the highest rate owing to the increasing adoption of FPGAs for acceleration of enterprise workloads [1]

[1] https://www.marketwatch.com/press-release/at-387-cagr-data-center-accelerator-market-size-is-expected-to-exhibit-35020-million-usd-by-2025-2019-10-15

Heterogeneous DCs for energy efficiency



- Low performance
- High power consumption
- Best effort

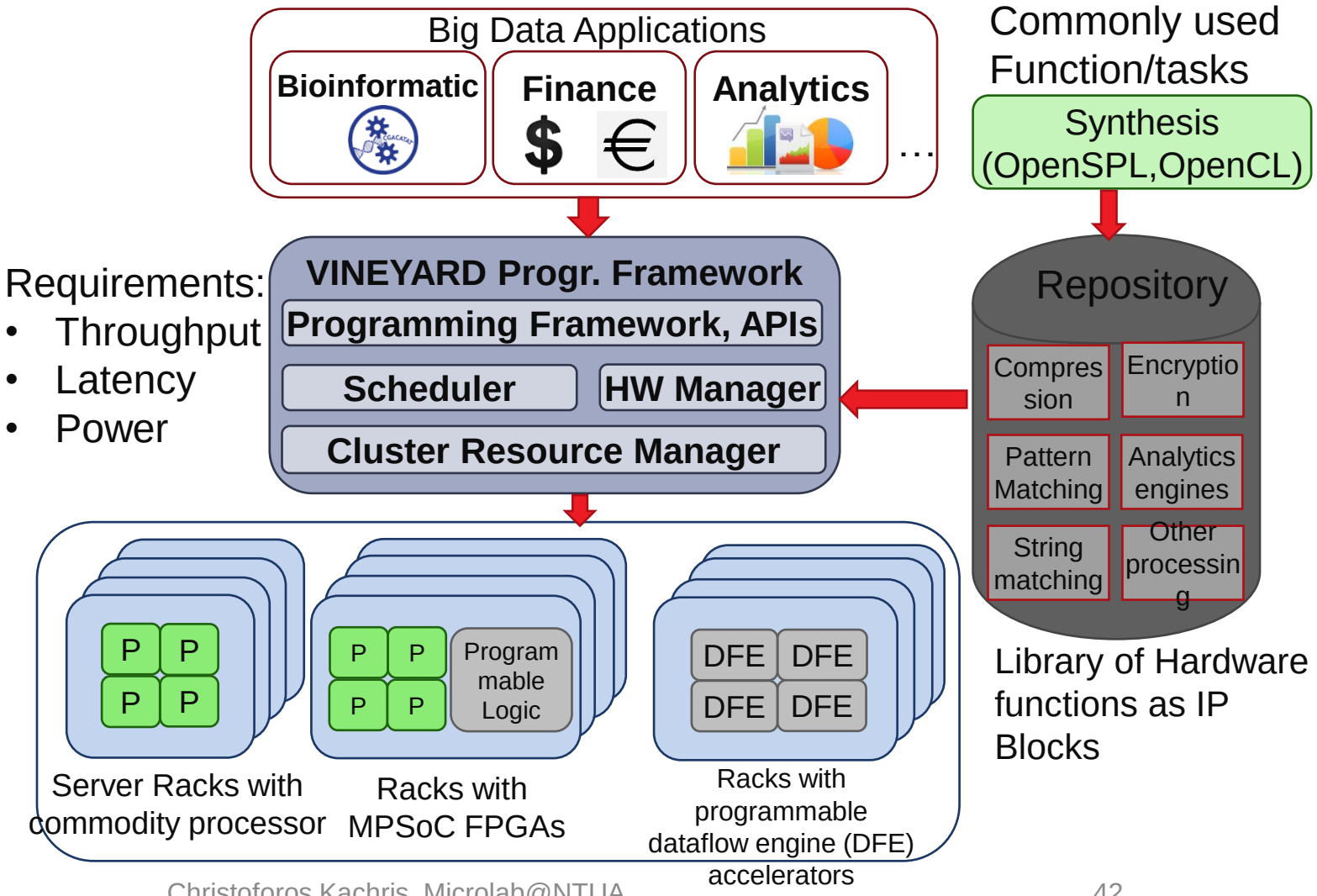
TODAY's DCs

- Higher performance
- Lower power consumption
- Predictable performance

Future Heterogeneous DCs with VINEYARD infrastructure

*"The only way to differentiate server offerings is through **accelerators**, like we saw with cell phones"*, OpenServer Summit 2014 Leendert Van Doorn; AMD

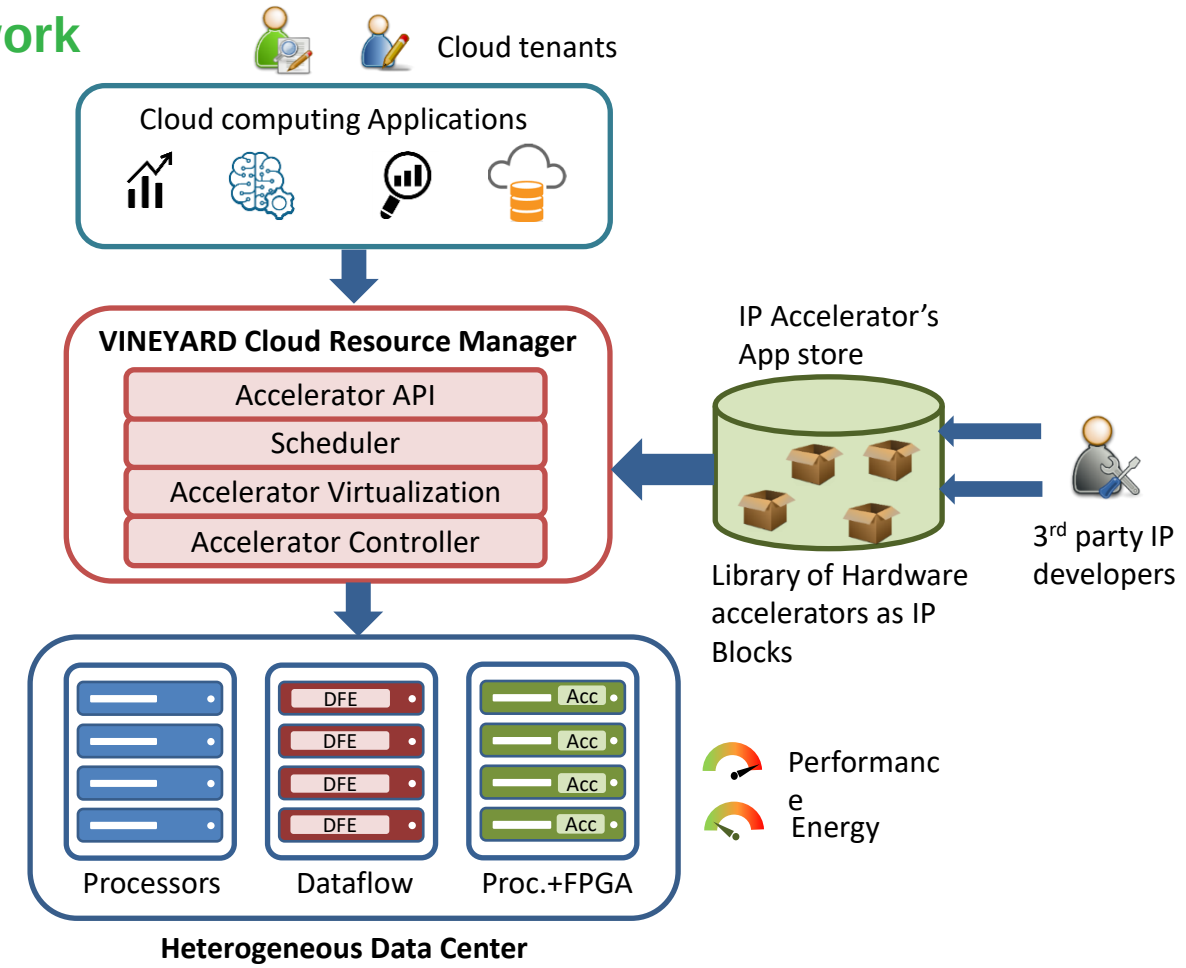
VINEYARD Heterogeneous Accelerators-based Data centre



VINEYARD

VINEYARD Framework

- **Accelerators stored in an AppStore**
- **Cloud users request accelerators based on applications requirements**
- **Decouple Hardware – Software designers**



AWS options

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
a1.medium	1	N/A	2 GiB	EBS Only	$0.0255 per Hour
a1.large	2	N/A	4 GiB	EBS Only	$0.051 per Hour
a1.xlarge	4	N/A	8 GiB	EBS Only	$0.102 per Hour
a1.2xlarge	8	N/A	16 GiB	EBS Only	$0.204 per Hour
a1.4xlarge	16	N/A	32 GiB	EBS Only	$0.408 per Hour
a1.metal	16	N/A	32 GiB	EBS Only	$0.408 per Hour
t3.nano	2	Variable	0.5 GiB	EBS Only	$0.0052 per Hour
t3.micro	2	Variable	1 GiB	EBS Only	$0.0104 per Hour
t3.small	2	Variable	2 GiB	EBS Only	$0.0208 per Hour
t3.medium	2	Variable	4 GiB	EBS Only	$0.0416 per Hour
t3.large	2	Variable	8 GiB	EBS Only	$0.0832 per Hour
t3.xlarge	4	Variable	16 GiB	EBS Only	$0.1664 per Hour
t3.2xlarge	8	Variable	32 GiB	EBS Only	$0.3328 per Hour
t3a.nano	2	Variable	0.5 GiB	EBS Only	$0.0047 per Hour
FPGA Instances - Current Generation					
f1.2xlarge	8	31	122 GiB	1 x 470 NVMe SSD	$1.65 per Hour
f1.4xlarge	16	58	244 GiB	1 x 940 NVMe SSD	$3.30 per Hour
f1.16xlarge	64	201	976 GiB	4 x 940 NVMe SSD	$13.20 per Hour
Machine Learning ASIC Instances					
inf1.xlarge	4	N/A	8 GiB	EBS Only	$0.368 per Hour
inf1.2xlarge	8	N/A	16 GiB	EBS Only	$0.584 per Hour
inf1.6xlarge	24	N/A	48 GiB	EBS Only	$1.904 per Hour
inf1.24xlarge	96	N/A	192 GiB	EBS Only	$7.615 per Hour

GPU Instances - Current Generation					
p3.2xlarge	8	31	61 GiB	EBS Only	$3.06 per Hour
p3.8xlarge	32	97	244 GiB	EBS Only	$12.24 per Hour
p3.16xlarge	64	201	488 GiB	EBS Only	$24.48 per Hour
p3dn.24xlarge	96	337	768 GiB	2 x 900 NVMe SSD	$31.212 per Hour
p2.xlarge	4	16	61 GiB	EBS Only	$0.90 per Hour
p2.8xlarge	32	97	488 GiB	EBS Only	$7.20 per Hour
p2.16xlarge	64	201	732 GiB	EBS Only	$14.40 per Hour
g4dn.xlarge	4	N/A	16 GiB	125 GB NVMe SSD	$0.526 per Hour
g4dn.2xlarge	8	N/A	32 GiB	225 GB NVMe SSD	$0.752 per Hour
g4dn.4xlarge	16	N/A	64 GiB	225 GB NVMe SSD	$1.204 per Hour
g4dn.8xlarge	32	N/A	128 GiB	900 GB NVMe SSD	$2.176 per Hour
g4dn.12xlarge	48	N/A	192 GiB	900 GB NVMe SSD	$3.912 per Hour

Memory Optimized - Current Generation					
x1.16xlarge	64	174.5	976 GiB	1 x 1920 SSD	$6.669 per Hour
x1.32xlarge	128	349	1,952 GiB	2 x 1920 SSD	$13.338 per Hour
x1e.xlarge	4	12	122 GiB	1 x 120 SSD	$0.834 per Hour
x1e.2xlarge	8	23	244 GiB	1 x 240 SSD	$1.668 per Hour
x1e.4xlarge	16	47	488 GiB	1 x 480 SSD	$3.336 per Hour
x1e.8xlarge	32	91	976 GiB	1 x 960 SSD	$6.672 per Hour
x1e.16xlarge	64	179	1,952 GiB	1 x 1920 SSD	$13.344 per Hour
x1e.32xlarge	128	340	3,904 GiB	2 x 1920 SSD	$26.688 per Hour
r5.large	2	10	16 GiB	EBS Only	$0.126 per Hour
r5.xlarge	4	19	32 GiB	EBS Only	$0.252 per Hour
r5.2xlarge	8	37	64 GiB	EBS Only	$0.504 per Hour
r5.4xlarge	16	70	128 GiB	EBS Only	$1.008 per Hour
r5.8xlarge	32	128	256 GiB	EBS Only	$2.016 per Hour

Performance evaluation on Machine Learning

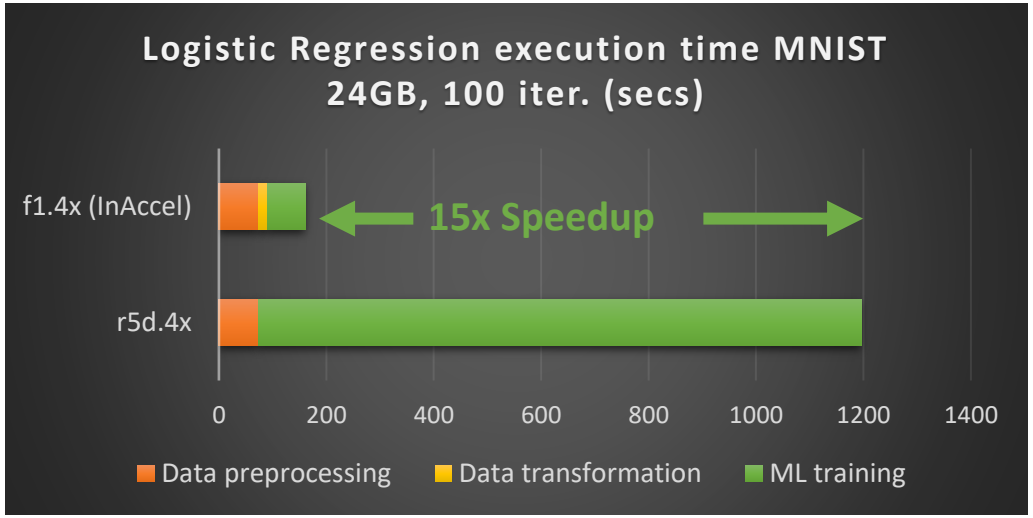
> **Up to 15x speedup for Logistic regression classification**

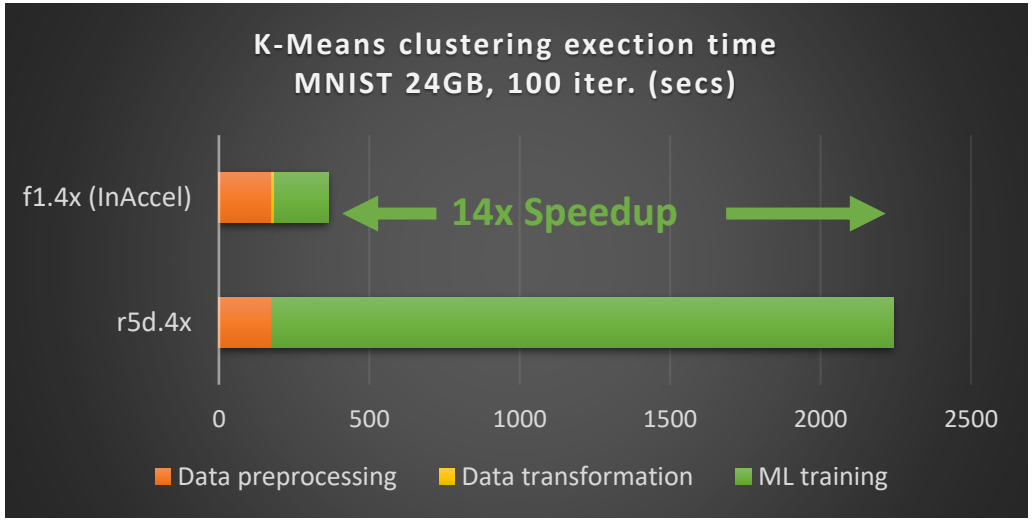
> **Up to 14x speedup for K-means clustering**

1st to offer ML-acceleration on the cloud using FPGAs

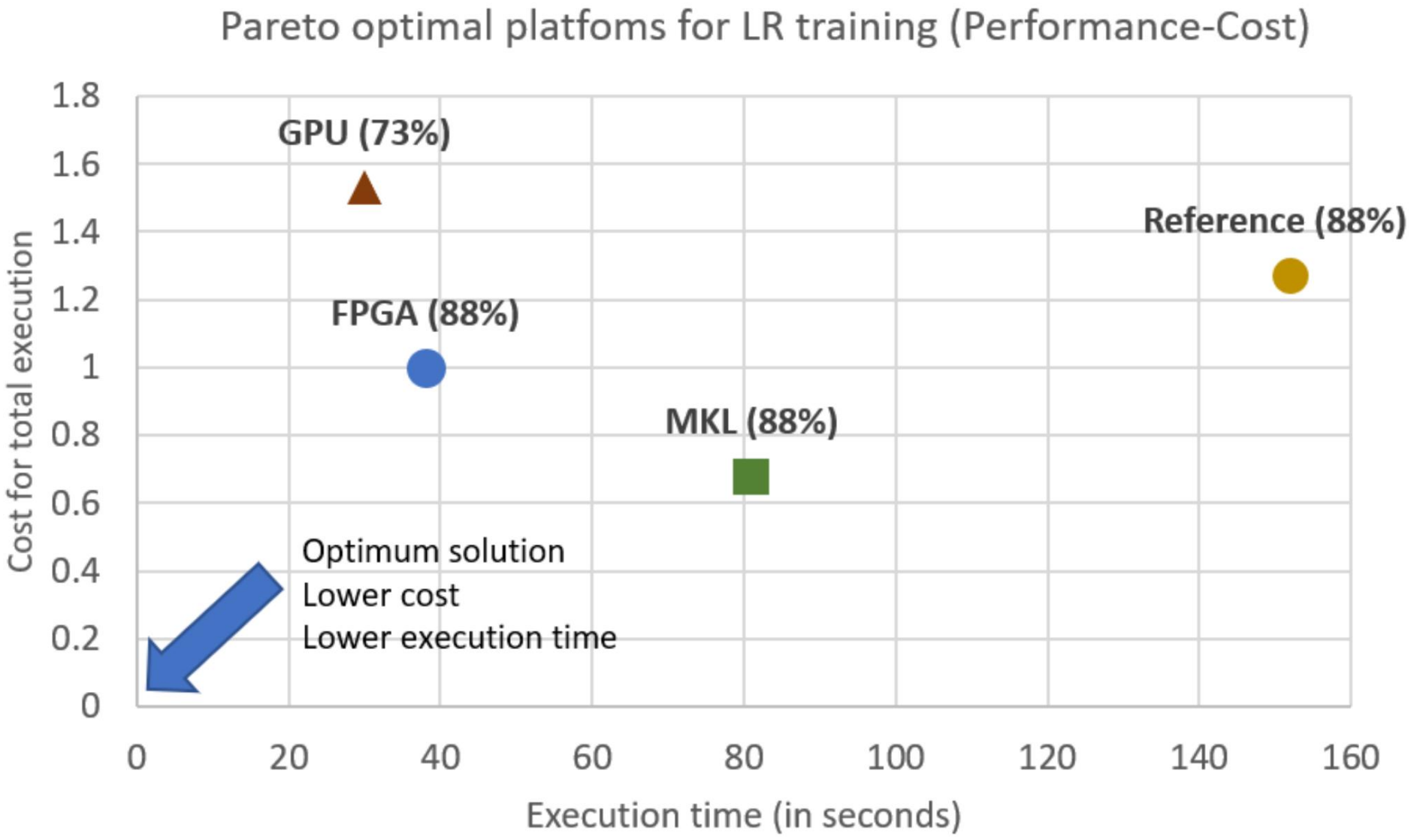
> Spark- GPU* (3.8x – 5.7x)

*[Spark-GPU: An Accelerated In-Memory Data Processing Engine on Clusters]



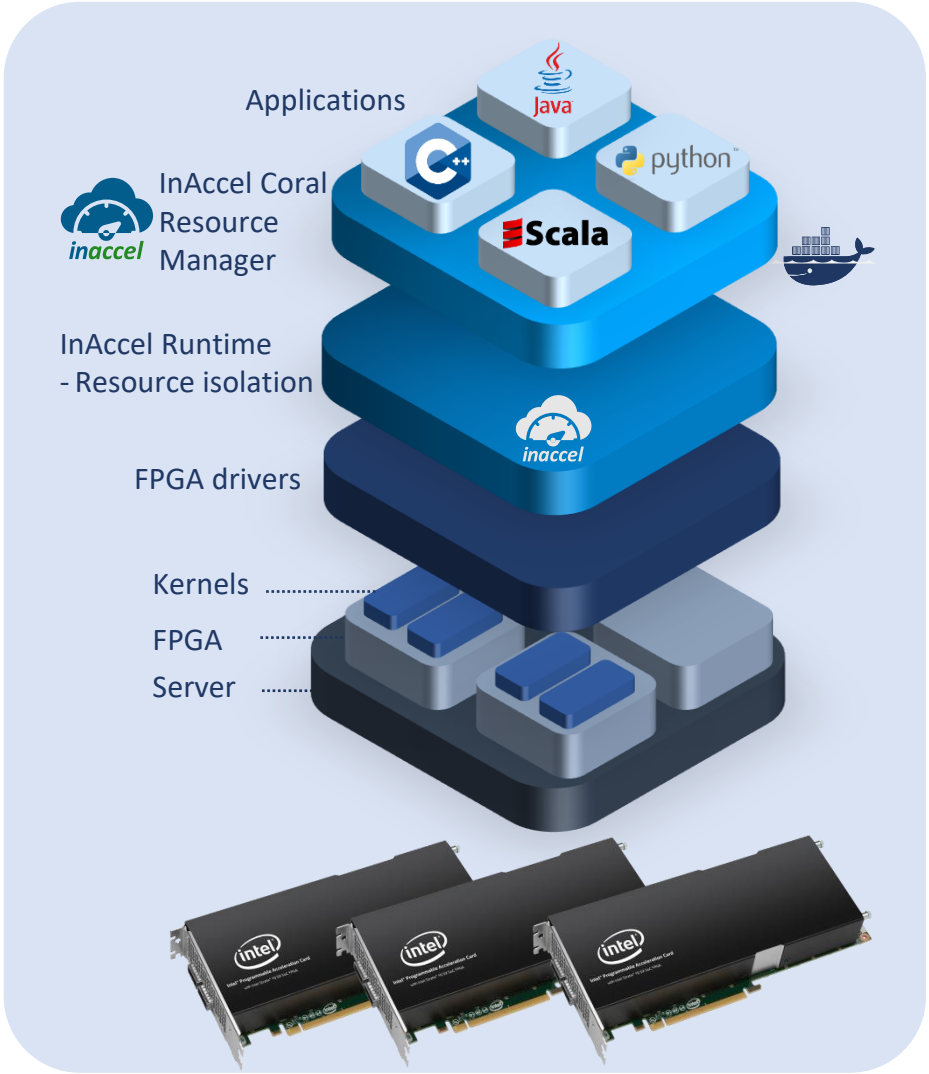


ML training



https://inaccel.com/cpu-gpu-or-fpga-performance-evaluation-of-cloud-computing-platforms-for-machine-learning-training/

Unique FPGA orchestrator by InAccel

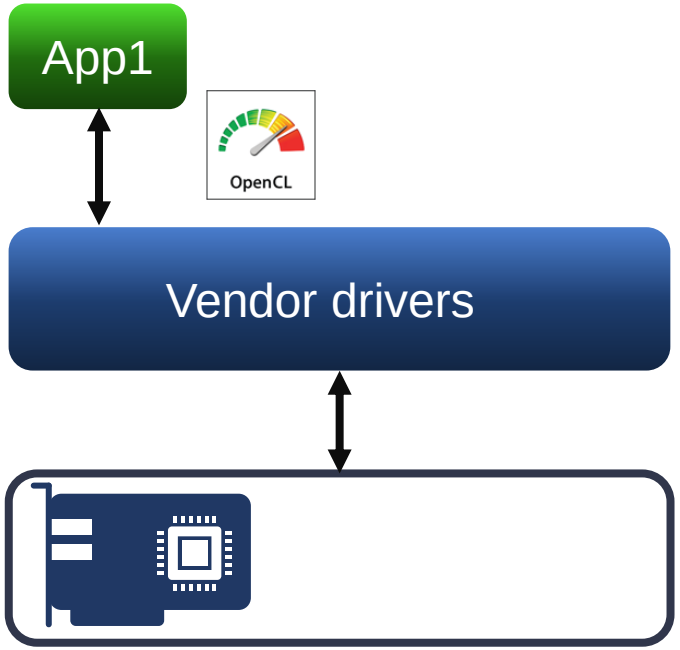


Automating deployment, scaling, and management of FPGA clusters

- **Seamless integration** with C/C++, Python, Java and Scala
- Automatic **virtualization** and scheduling of the applications to the FPGA cluster
- **Fully scalable**: Scale-up (multiple FPGAs per node) and Scale-out (multiple FPGA-based servers over Spark)

Current limitations for FPGA deployment

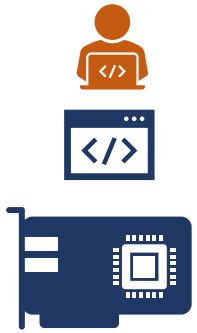
- Currently only **one application** can talk to a single FPGA accelerator through **OpenCL**
- Application can talk to a **single** FPGA.
- Complex device sharing
  - From multiple threads/processes
  - Even from the same thread
- Explicit allocation of the resources (memory/compute units)
- User need to specify which FPGA to use (device ID, etc.)



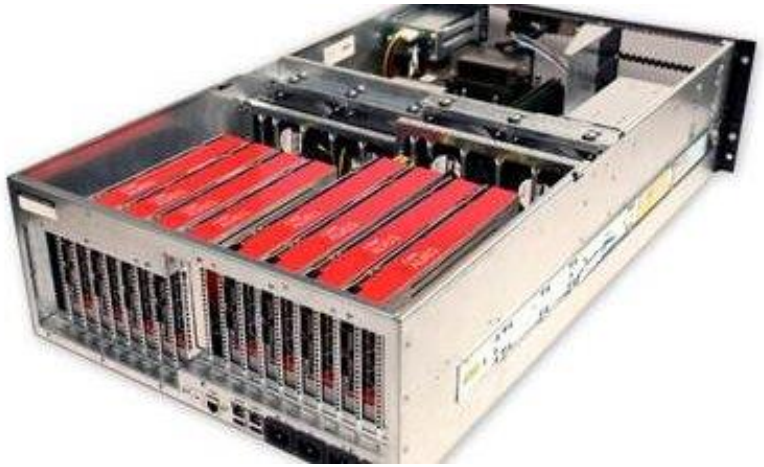
Single FPGA

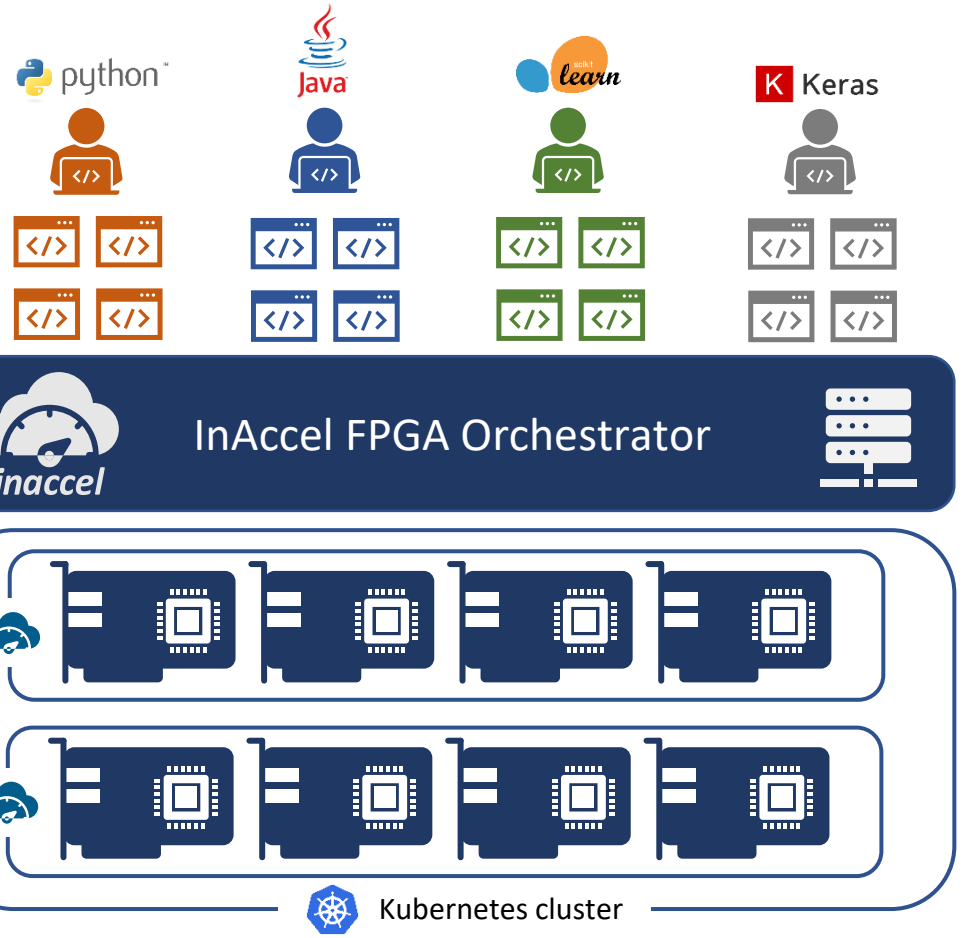
From single instance to data centers

- Easy deployment
- Instant scaling
- Seamless sharing
- Multiple-users
- Multiple applications
- Isolation
- Privacy



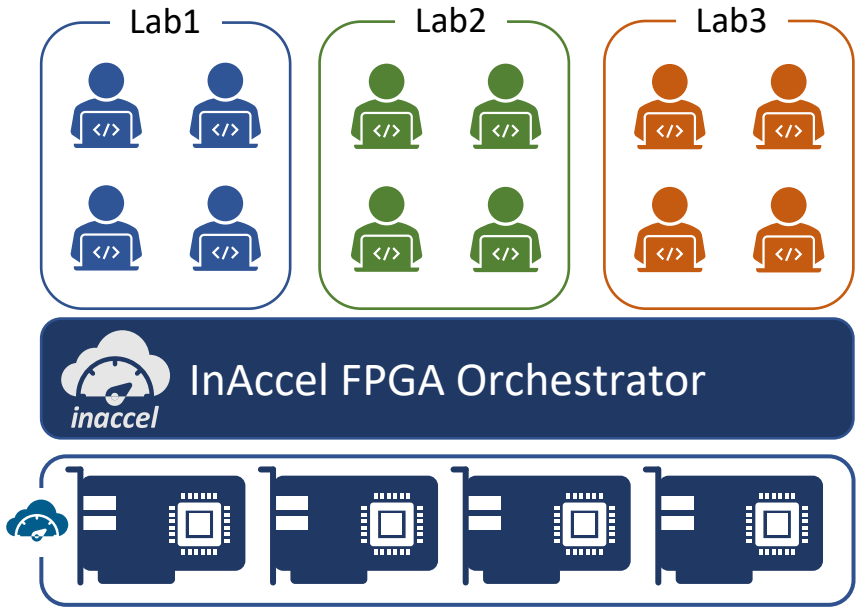
inaccel





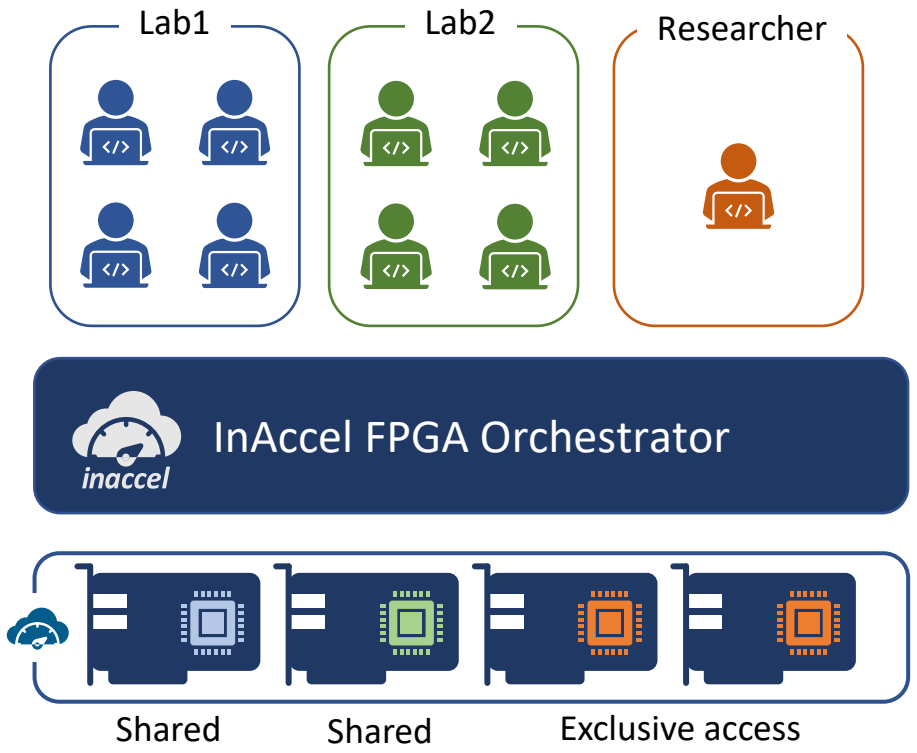
Universities

- **How do you allow multiple students to share the available FPGAs?**
- Many universities have limited number of FPGA cards that want to share with multiple students.
- InAccel FPGA orchestrator allows multiple students to share one or more FPGAs seamlessly.
- It allows students to just invoke the function that want to accelerate and InAccel FPGA manager performs the serialization and the scheduling of the functions to the available FPGA resources.



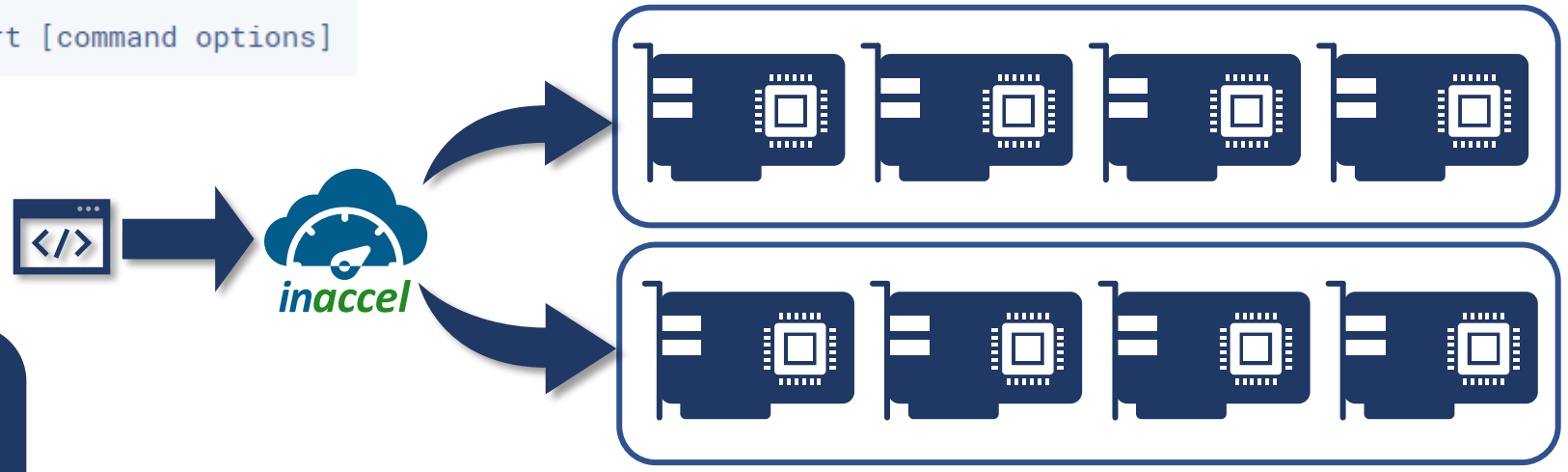
Universities

- **But the researchers want exclusive access**
- InAccel orchestrator allows to select which FPGA cards will be available for multiple students and which FPGAs can be allocated exclusively to researchers and Ph.D. students (so they can get accurate measurements for their papers).
- The FPGAs that are shared with multiple students will perform on a best-effort approach (InAccel manager performs the serialization of the requested access) while the researchers have exclusive access to the FPGAs with zero overhead.

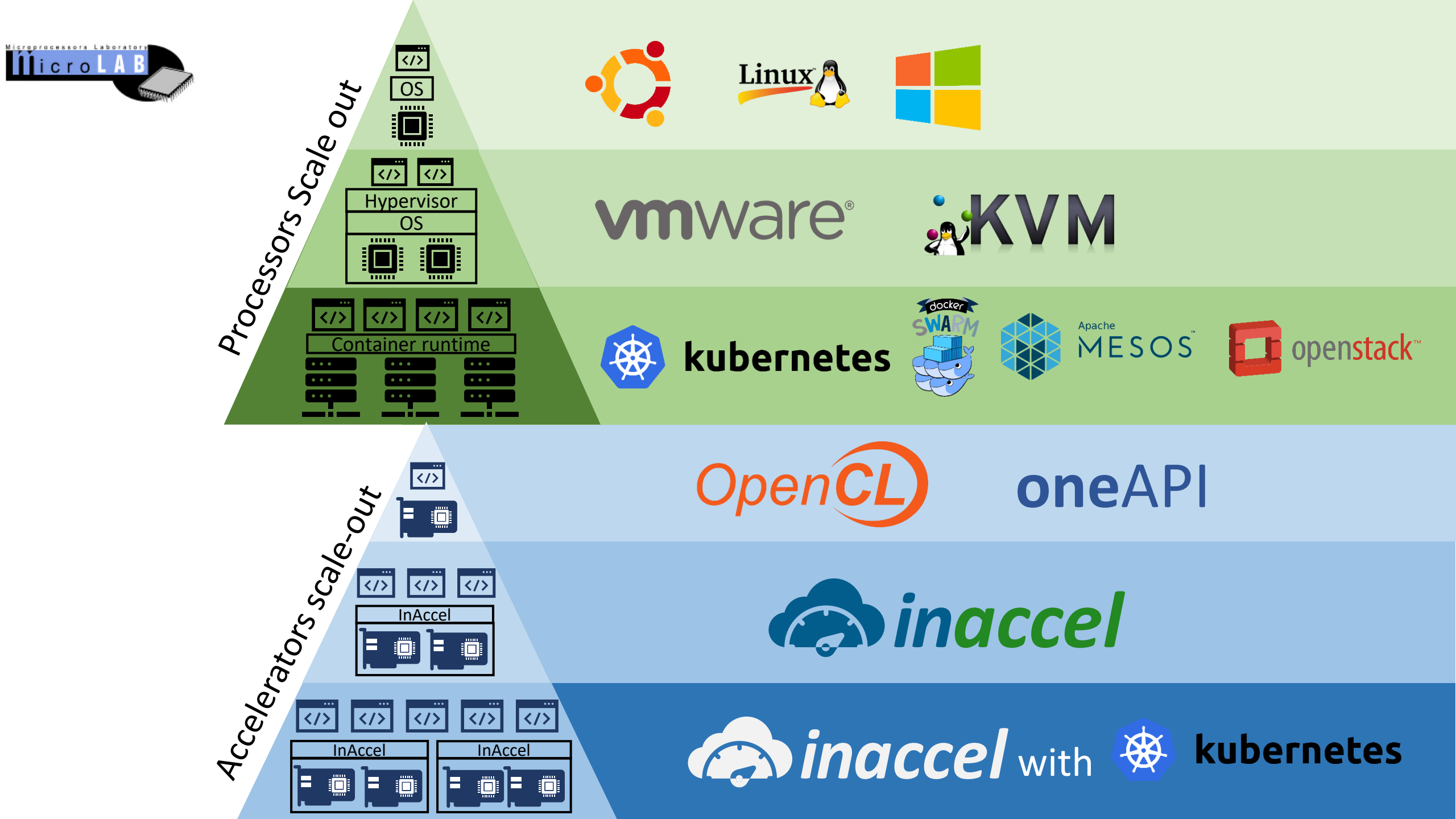


Instant Scalability

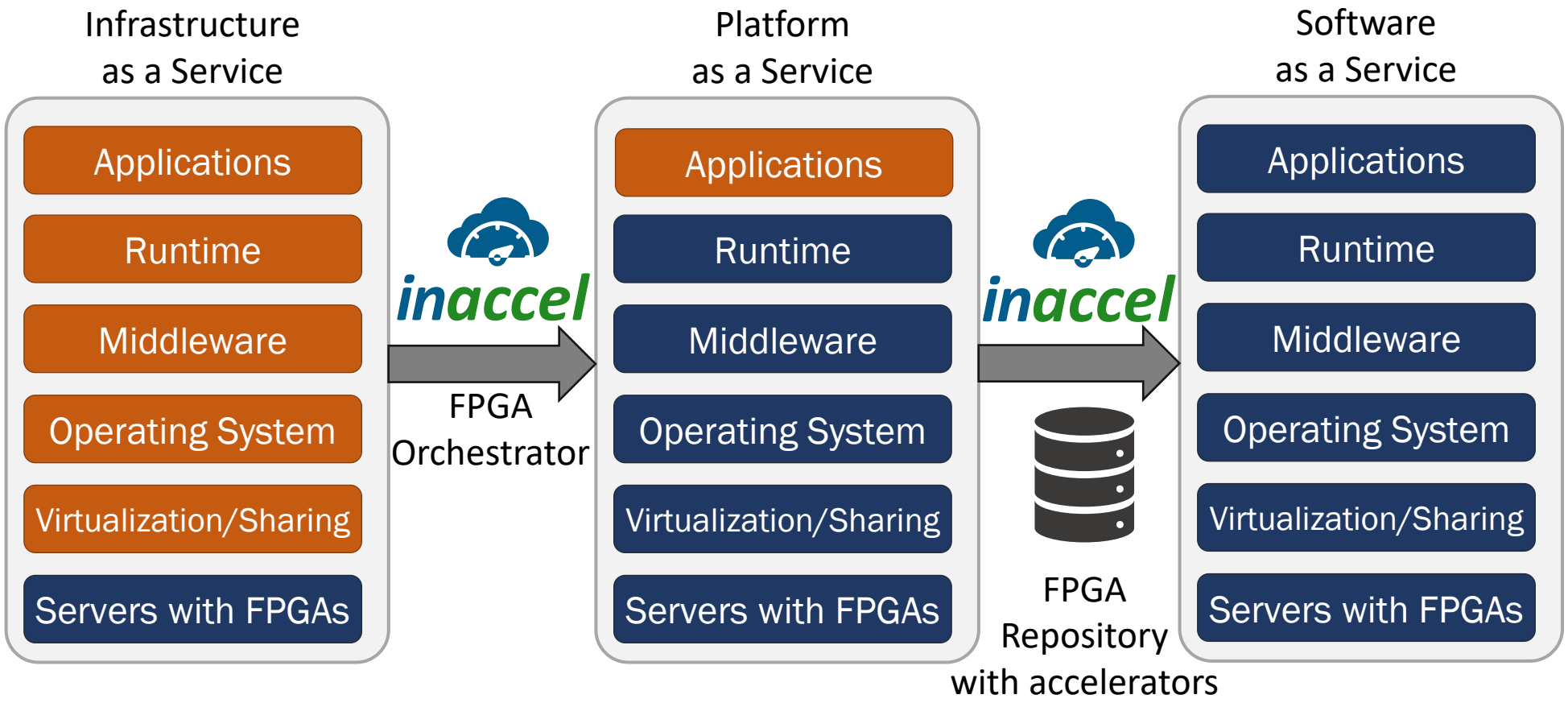
```
inaccel coral start [command options]
```



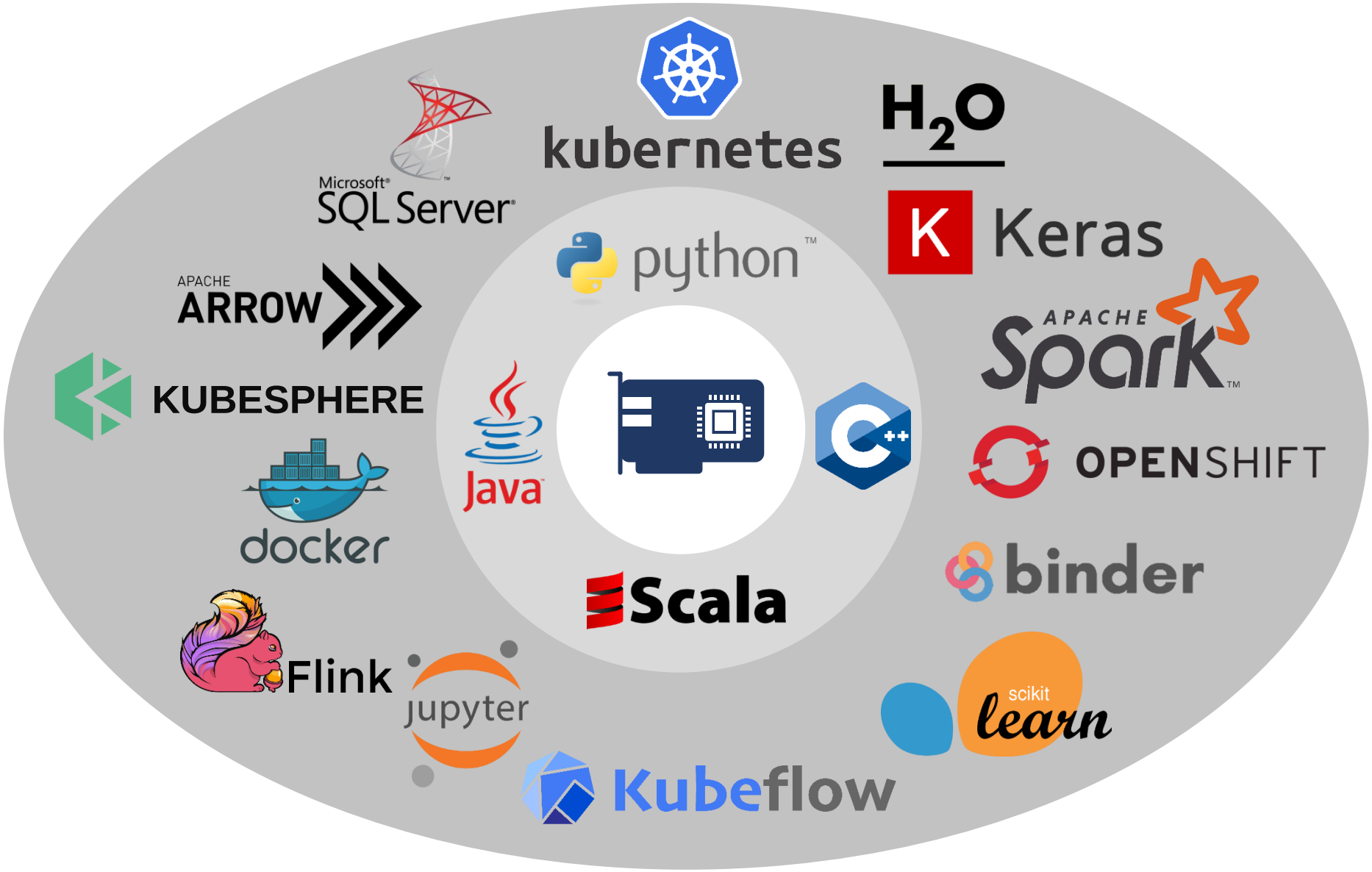
Distribution of multi-thread applications to multiple clusters With a single command


Microprocessors Laboratory
microLAB
Processors Scale out
OS
Hypervisor
OS
Container runtime
Linux
vmware
KVM
kubernetes
docker
SWARM
Apache
MESOS
openstack
Accelerators scale-out
InAccel
InAccel
InAccel
OpenCL
oneAPI
inaccel
inaccel with
kubernetes

From IaaS to PaaS and SaaS for FPGAs



Seamless Integration with any framework



Lab Exercise

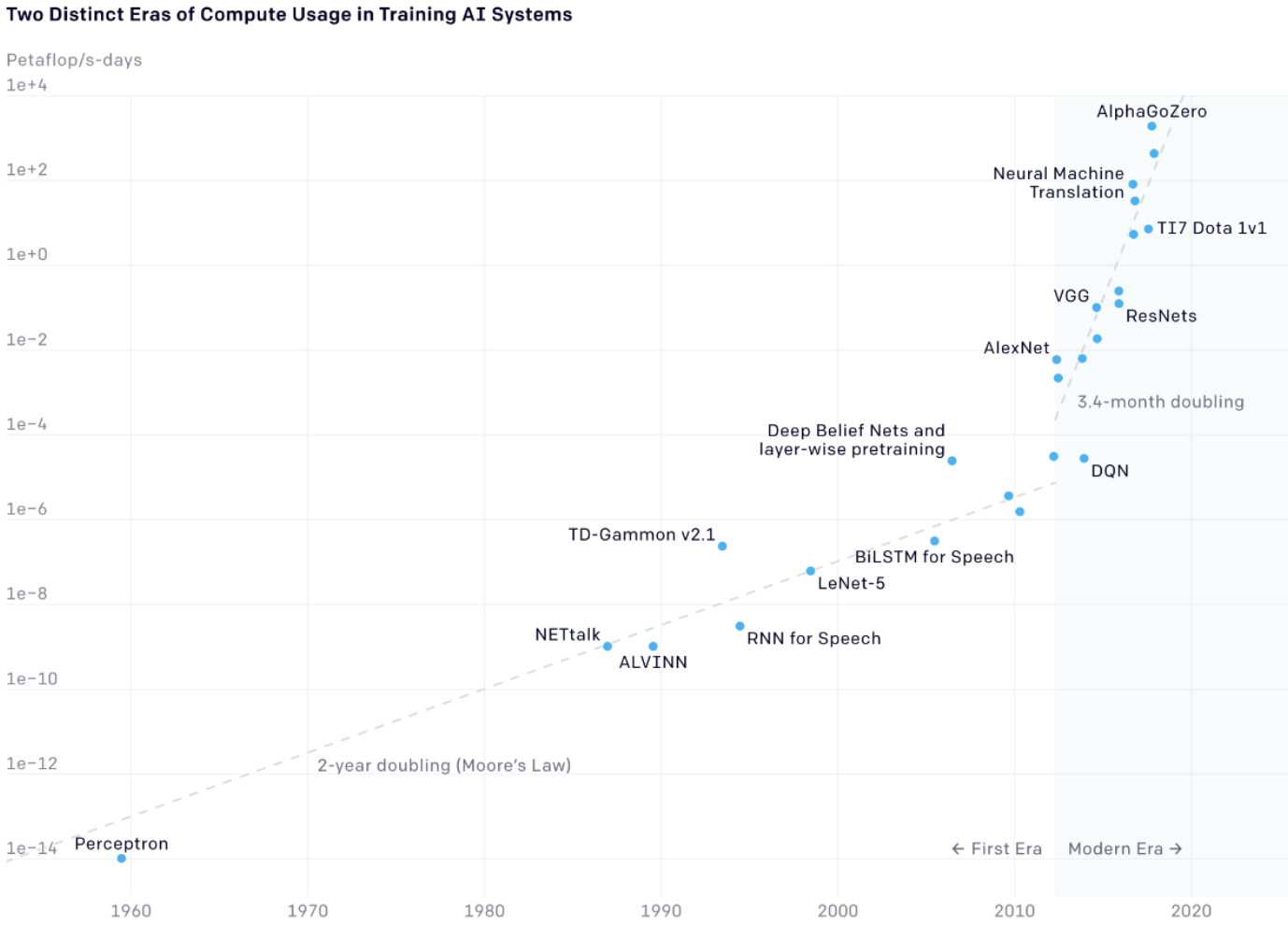
- In this lab you are going to create your first accelerated application
- Use scikit learn to find out the speedup you get upon running Naive Bayes algorithm using the original (CPU) and FPGA implementation.

Conclusions

- **Future Data Center will have to sustain huge amount of network traffic**
- **However the power consumption will have to remain almost the same**
- **FPGA acceleration as a promising solution for Machine Learning providing**
  - **high throughput**,
  - **low latency** and
  - **energy efficient processing**

Domain Specific Accelerators

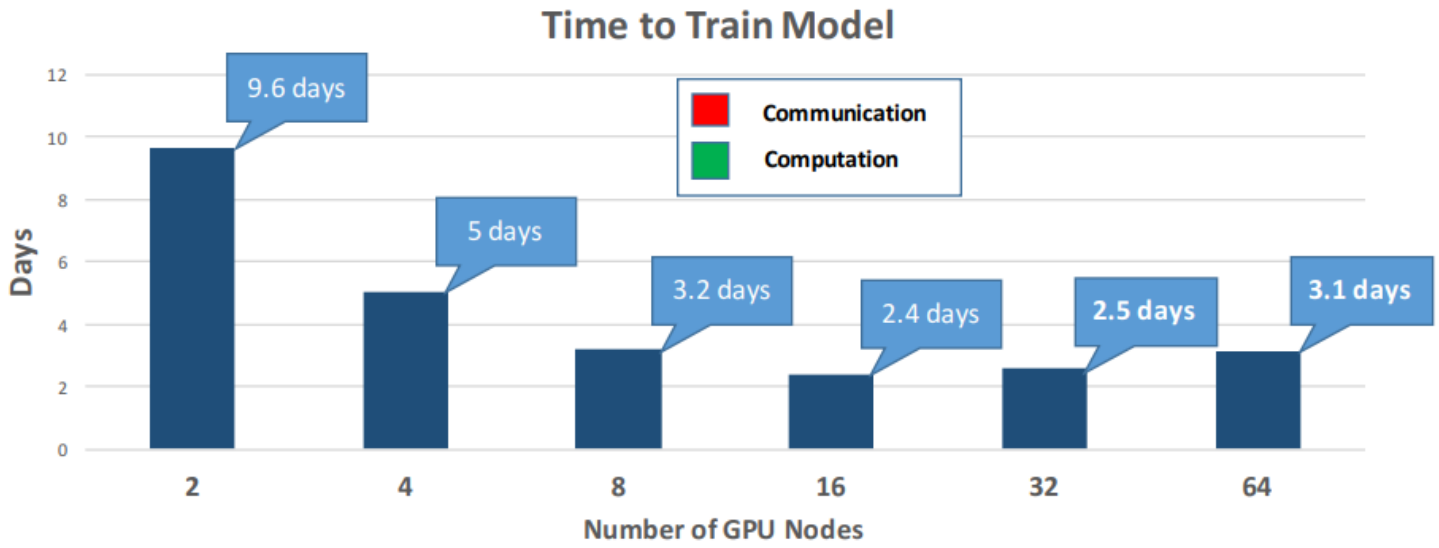
The amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time (by comparison, Moore's Law had a 2-year doubling period)



Distributed ML

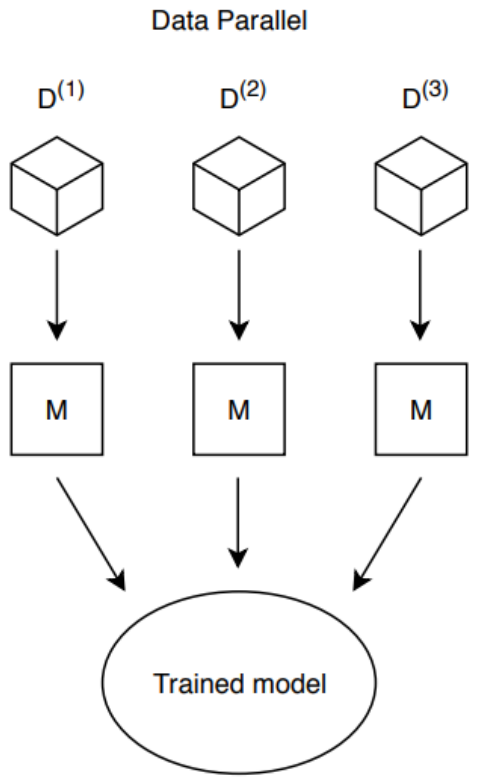
- **CSCS: Europe's Top Supercomputer (World 3rd) • 4500+ GPU Nodes, state-of-the-art interconnect Task:**

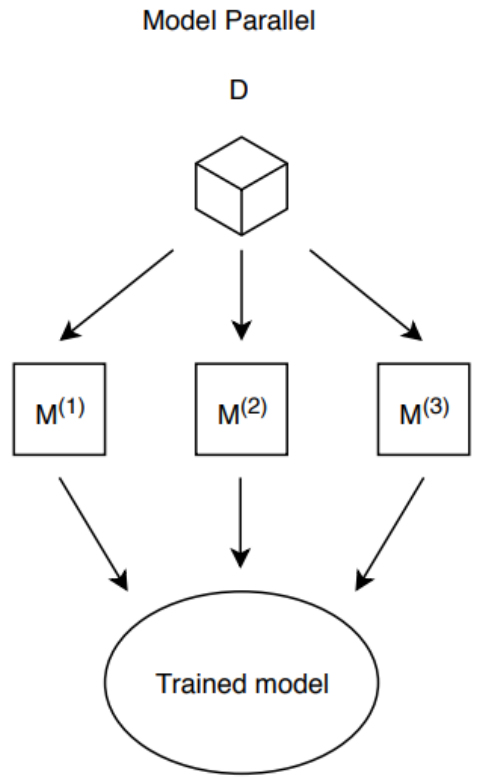
- **Image Classification (ResNet-152 on ImageNet)**
  - Single Node time (TensorFlow): 19 days
  - 1024 Nodes: 25 minutes (in theory)



Distributed ML

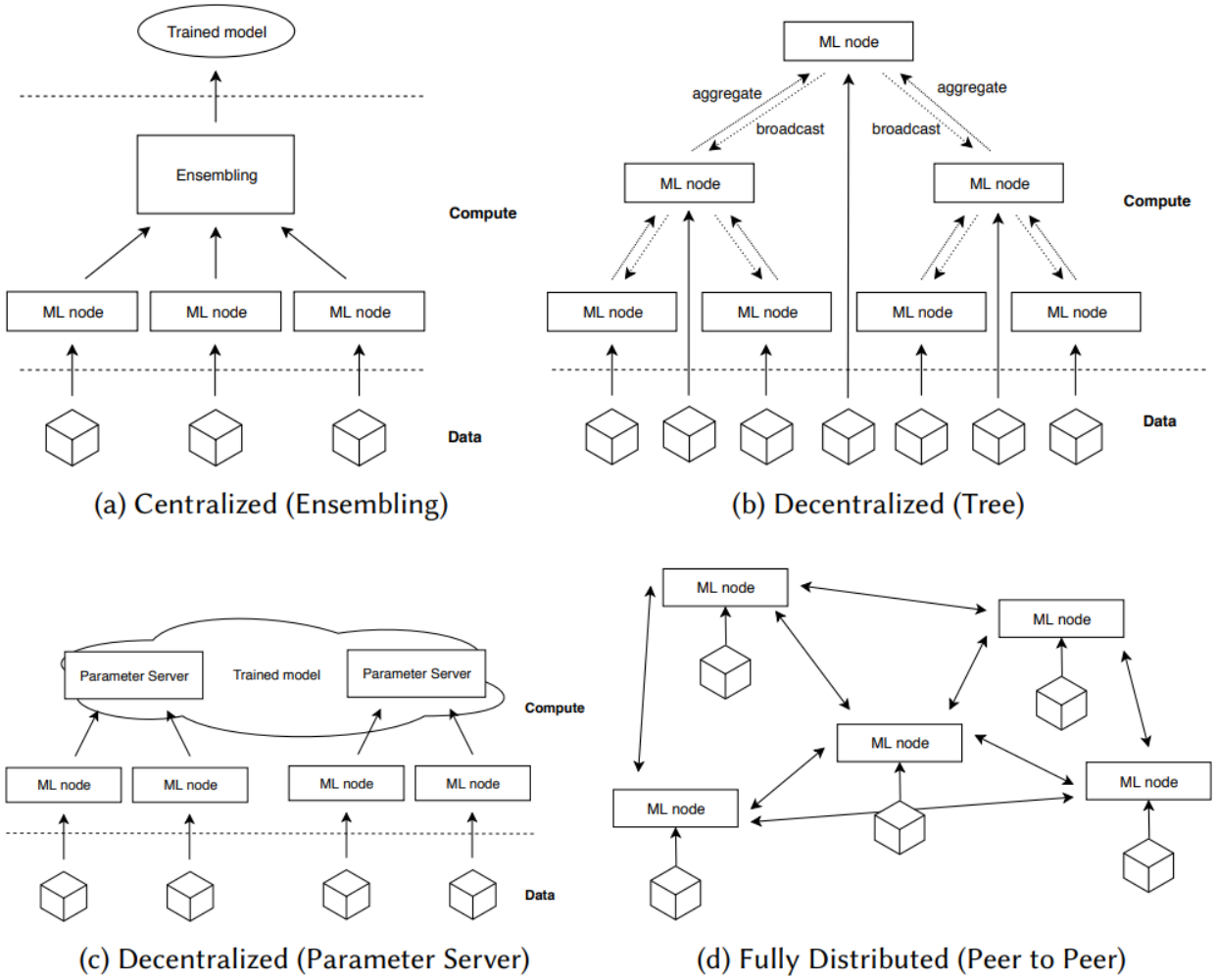
- **Parallelism in Distributed Machine Learning.**
- **Data parallelism trains multiple instances of the same model on different subsets of the training dataset,**
- **model parallelism distributes parallel paths of a single model to multiple nodes**





A Survey on Distributed Machine Learning: https://arxiv.org/ftp/arxiv/papers/1912/1912.09789.pdf

- Centralized systems (Figure 3a) employ a strictly hierarchical approach to aggregation, which happens in a single central location.
- Decentralized systems allow for intermediate aggregation, either with a replicated model that is consistently updated when the aggregate is broadcast to all nodes such as in tree topologies (Figure 3b) or with a partitioned model that is shared over multiple parameter servers (Figure 3c).
- Fully distributed systems (Figure 3d) consists of a network of independent nodes that ensemble the solution together and where no speciffic roles are assigned to certain nodes



Distributed ML ecosystem

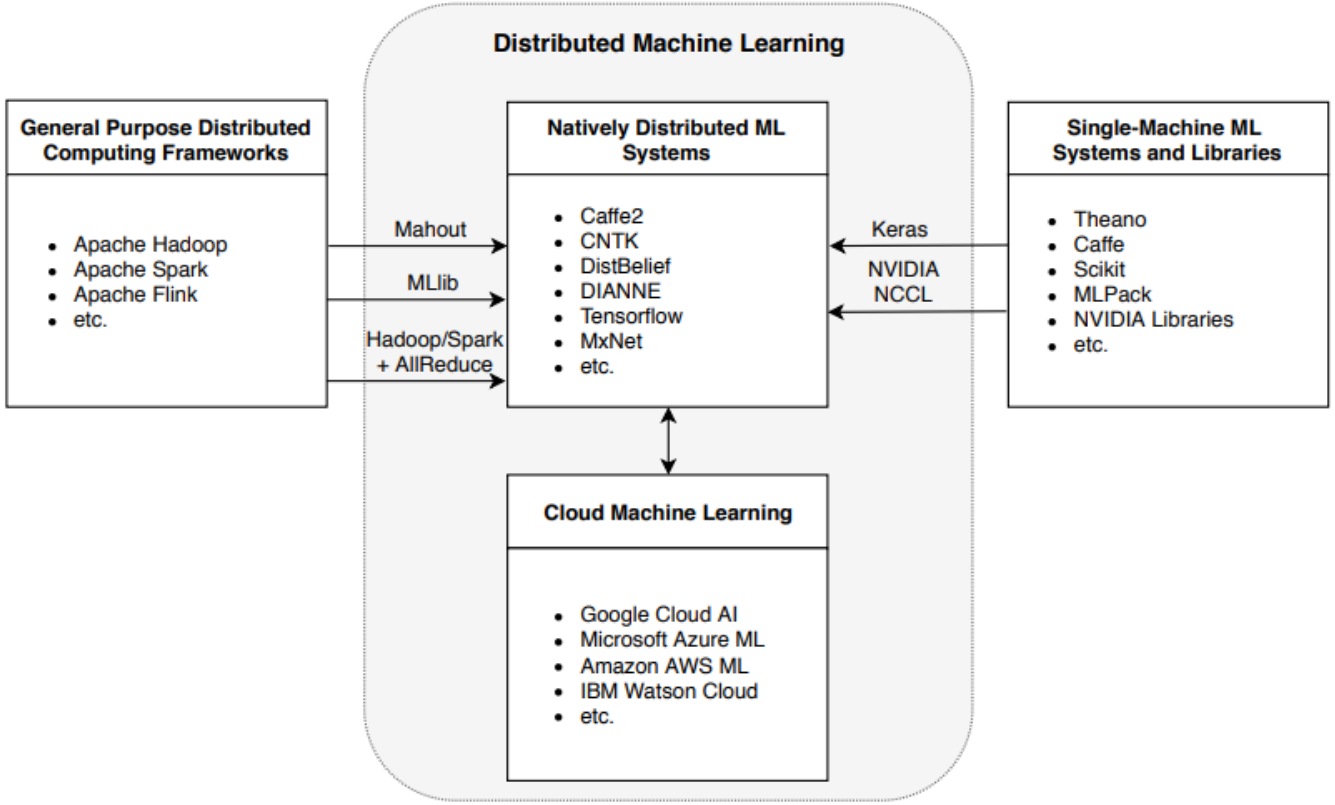


Fig. 4. Distributed Machine Learning Ecosystem. Both general purpose distributed frameworks and single-machine ML systems and libraries are converging towards Distributed Machine Learning. Cloud emerges as a new delivery model for ML.

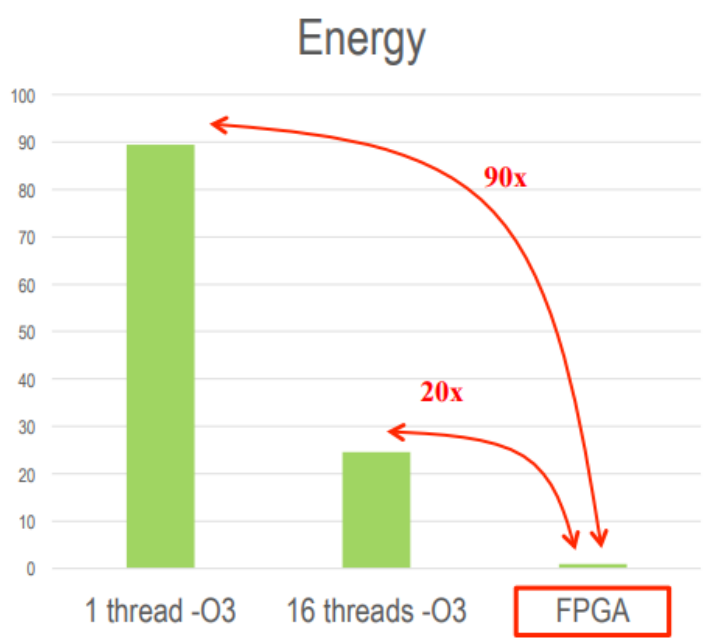
Data Science and ML platforms

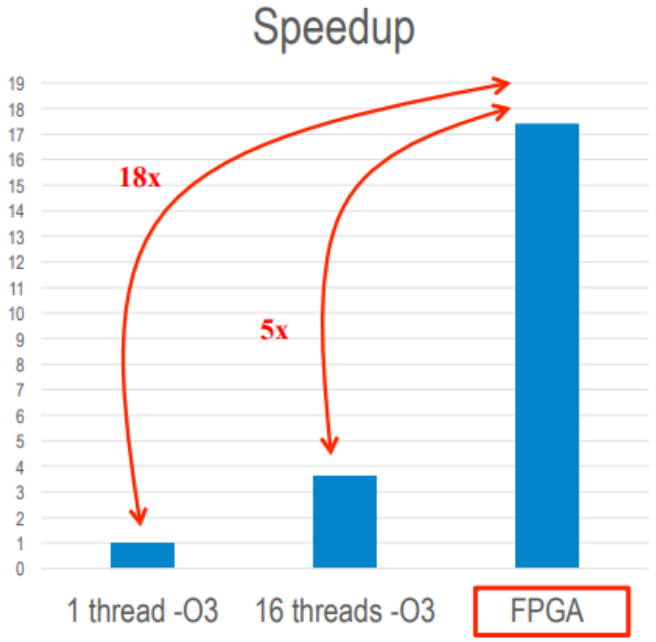
FPGA for ML

- **In many applications, neural network is trained in back-end CPU or GPU clusters ◆ FPGA:**
- **very suitable for latency-sensitive real-time inference job**
  - Unmanned vehicle
  - Speech Recognition
  - Audio Surveillance
  - Multi-media

CPU vs FPGAs

Experimental Results: vs. CPU





CPU	Xeon E5-2430 (32nm)	16 cores	2.2 GHz	gcc 4.7.2 –O3 OpenMP 3.0
FPGA	Virtex7-485t (28nm)	448 PEs	100MHz	Vivado 2015.2 Vivado HLS 2015.2

18

http://cadlab.cs.ucla.edu/~cong/slides/HALO15_keynote.pdf

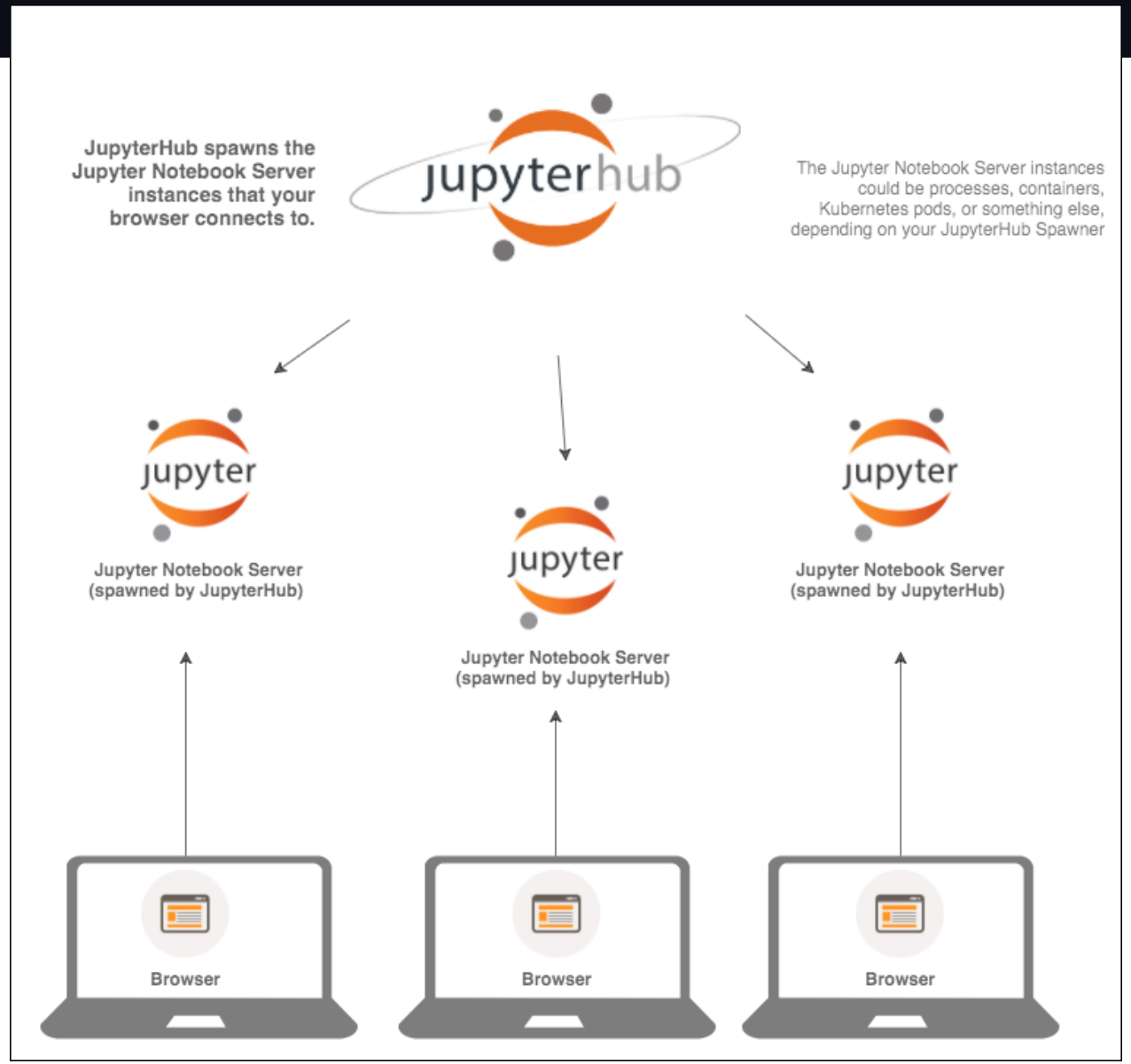
Machine Learning on FPGAs

- **Classification**
  - Naïve Bayes
- **Training**
  - Logistic regression
- **DNN**
  - Resnet50

Jupyter - JupyterHub

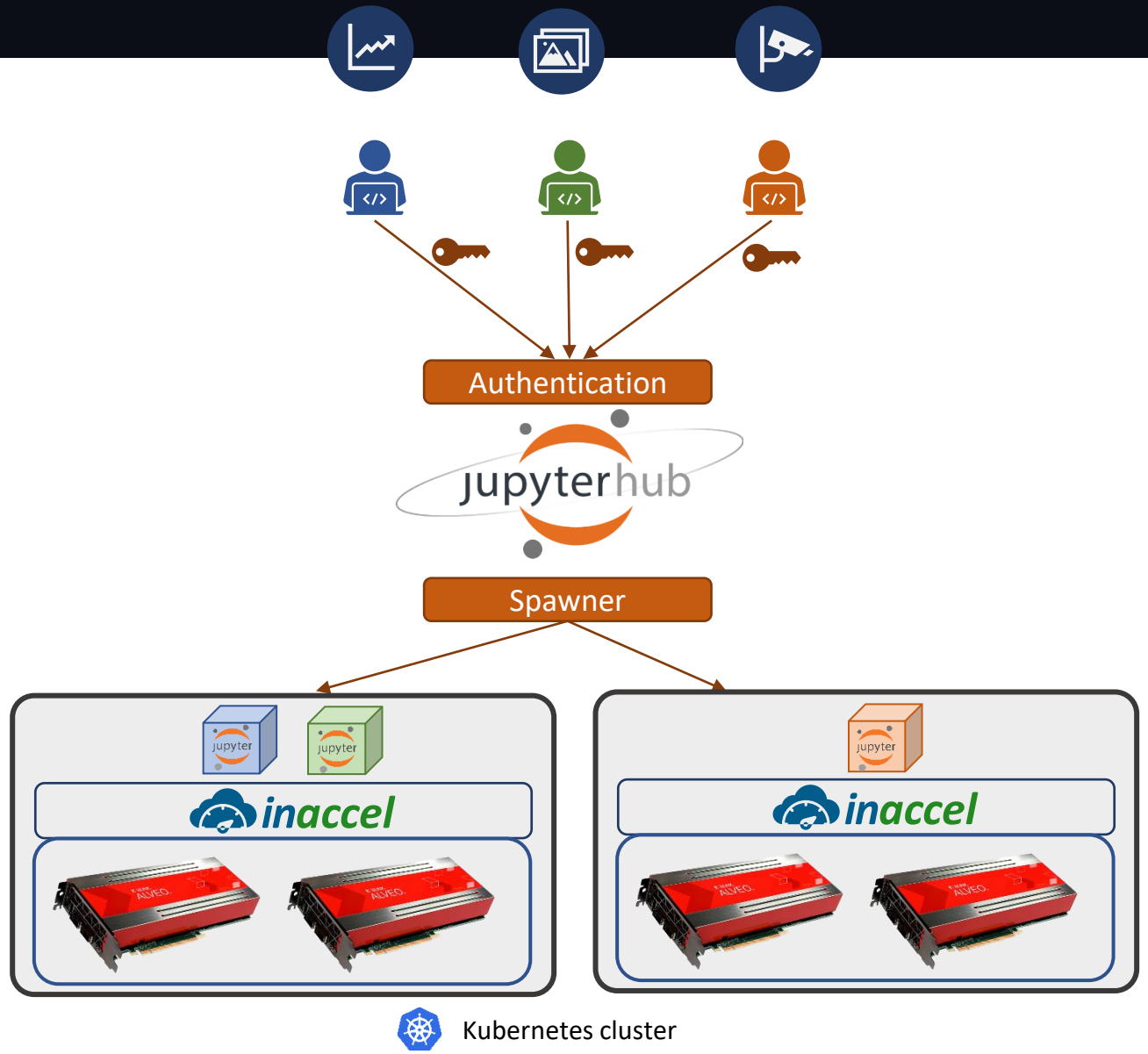
- Deploy and run your FPGA-accelerated applications using Jupyter Notebooks

- InAccel manager allows the instant deployment of FPGAs through HupyterHub

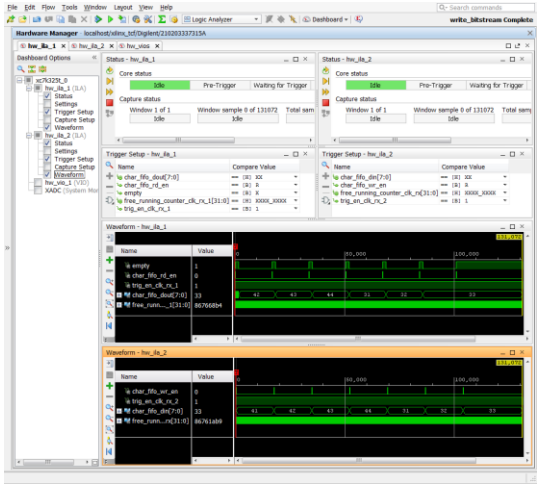


JupyterHub on FPGAs

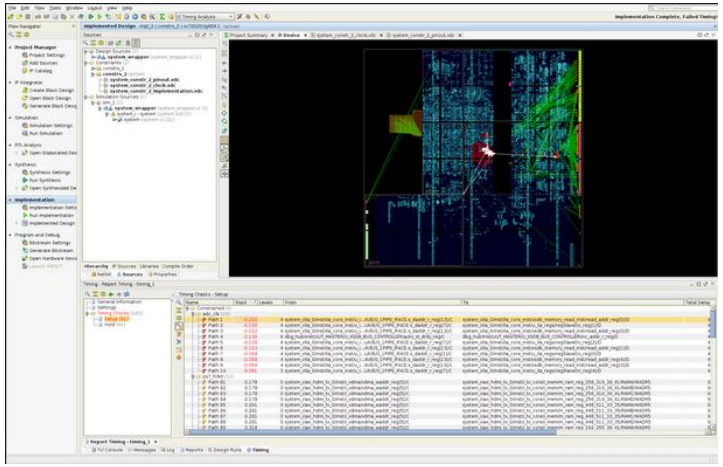
- Instant acceleration of Jupyter Notebooks with zero code-changes
- Offload the most computational intensive tasks on FPGA-based servers



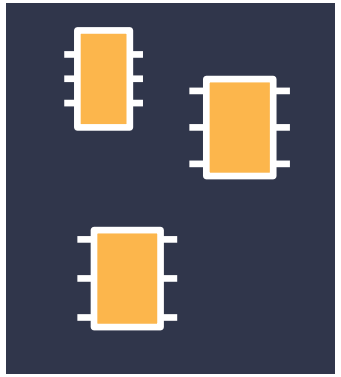
FPGA flow



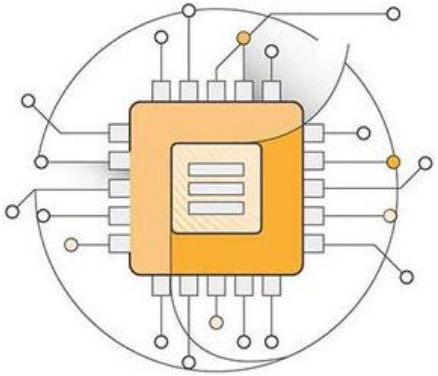
FPGA Logic Design using Xilinx Vivado on C4 or M4 instance



FPGA Place-and-Route using Xilinx Vivado on C4 or M4 instance



Generate an bitstream

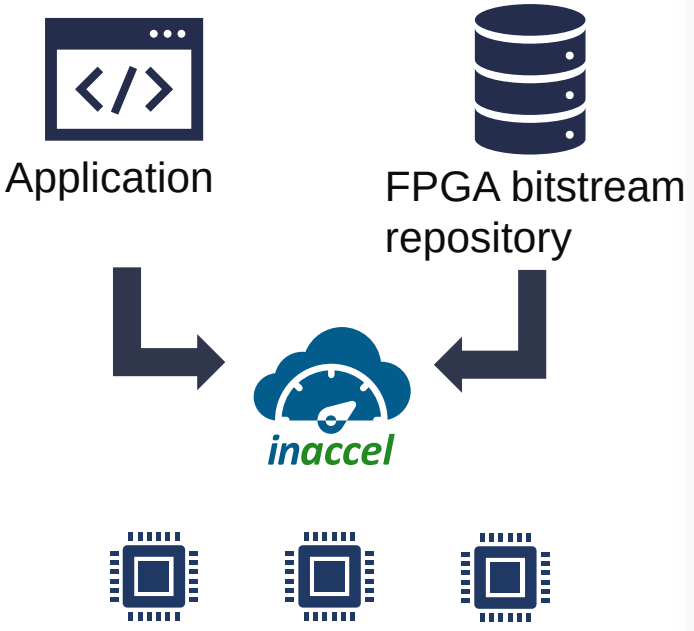


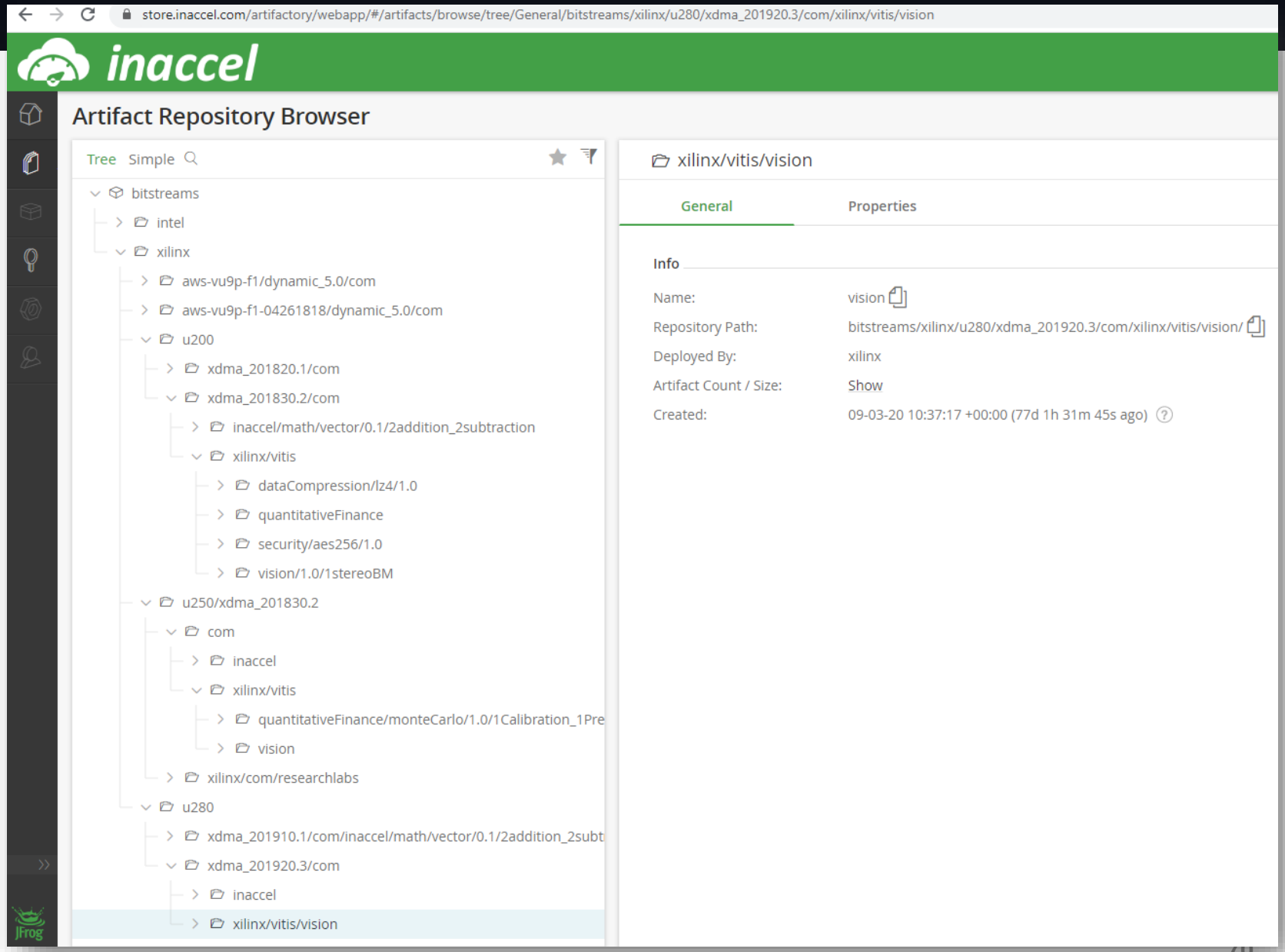
Program the FPGA

Bitstream repository

- **FPGA Resource Manager is integrated with a bitstream repository that is used to store FPGA bitstreams**

https://store.inaccel.com





Lab Exercise

- Introduction
- Creating a Bitstream Artifact
- Running the first FPGA accelerated application
- Scikit-Learn on FPGAs

- Naive Bayes Example

- Logistic Regression Example

https://edu.inaccel.com/

Useful links

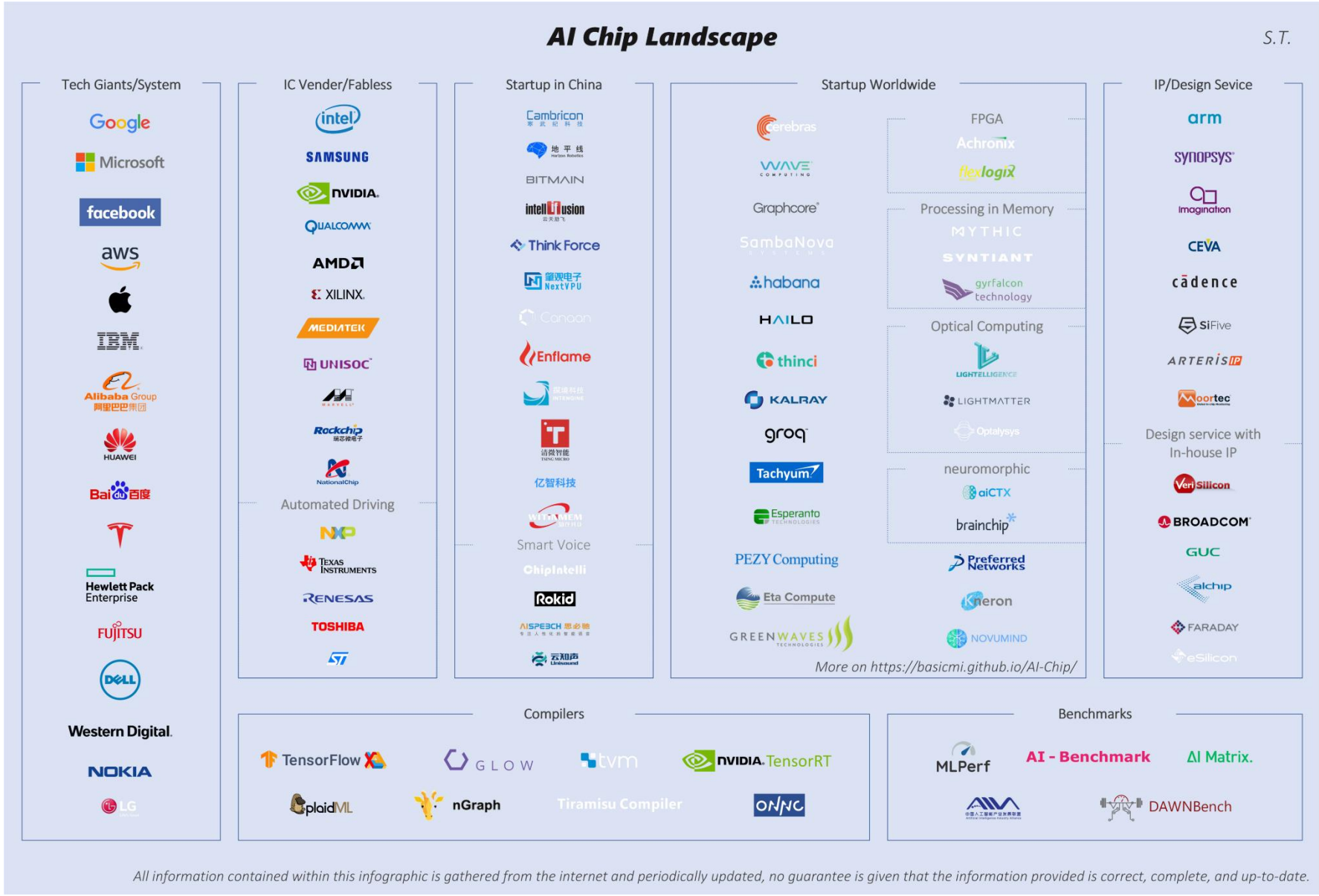
- MIT: Tutorial on Hardware Accelerators for Deep Neural Networks
  - http://eyeriss.mit.edu/tutorial.html

- **Intel**
  - https://software.intel.com/content/www/us/en/develop/training/course-deep-learning-inference-fpga.html

- **UCLA: Machine Learning on FPGAs**
  - http://cadlab.cs.ucla.edu/~cong/slides/HALO15_keynote.pdf

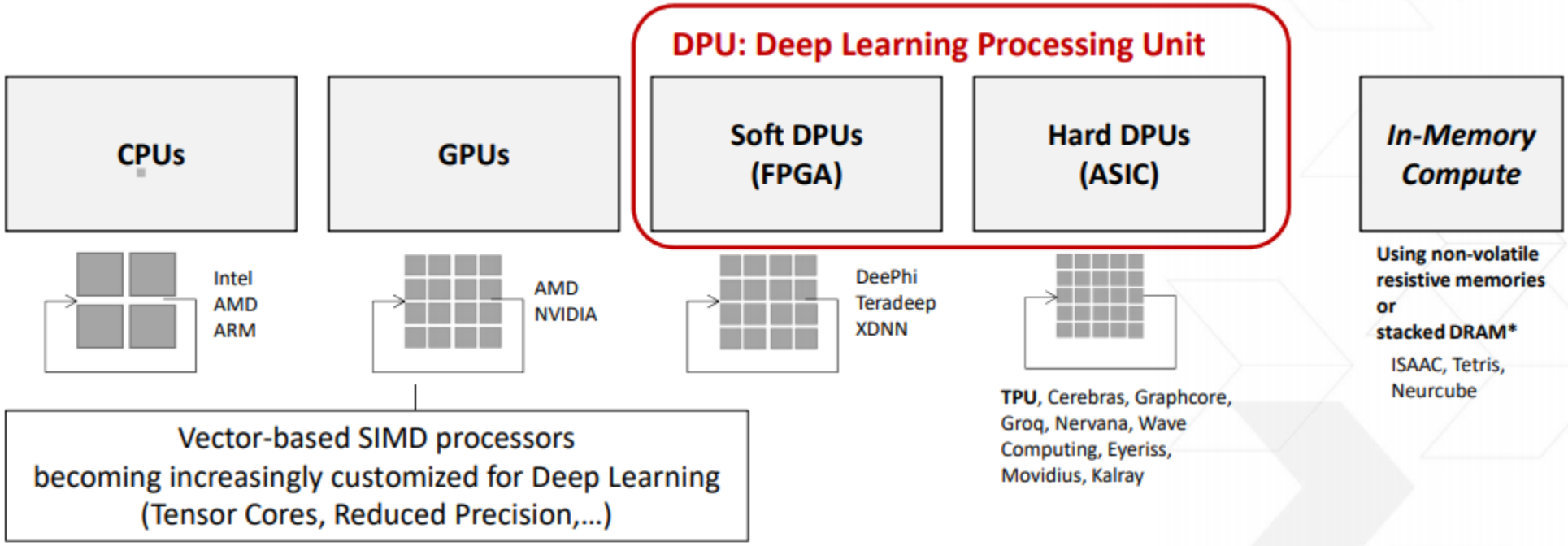
- **Distributed ML**
  - https://www.podc.org/data/podc2018/podc2018-tutorial-alistarh.pdf

AI chip Landscape



https://basicmi.github.io/AI-Chip/

Spectrum of new architectures for DNN



*Shafiee, A., Nag, A., Muralimanohar, N., Balasubramonian, R., Strachan, J.P., Hu, M., Williams, R.S. and Srikumar, V., 2016. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. ACM SIGARCH

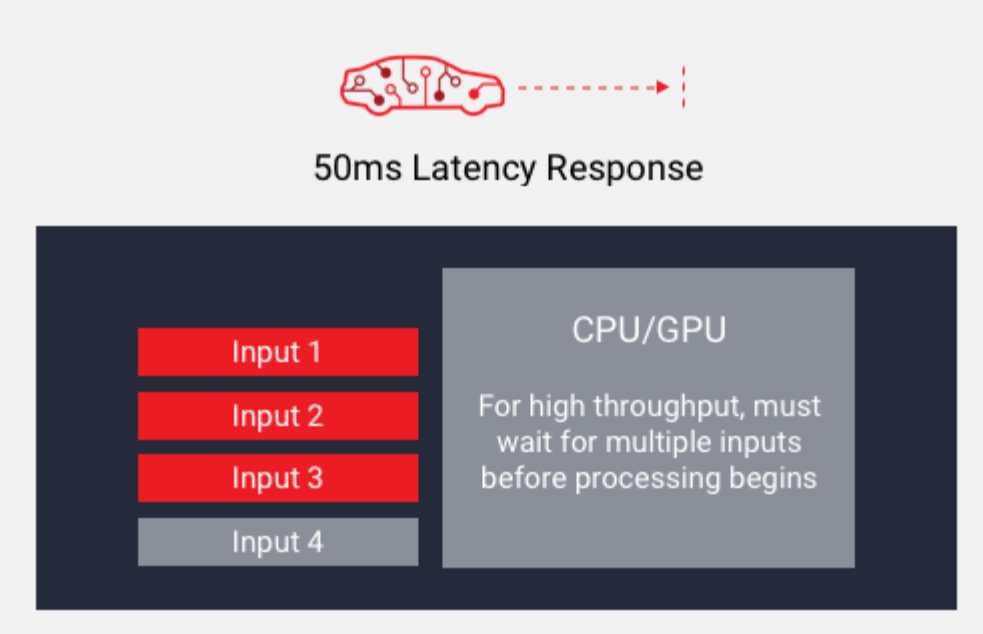
Chi, P., Li, S., Xu, C., Zhang, T., Zhao, J., Liu, Y., Wang, Y. and Xie, Y., 2016, June. Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. In ACM SIGARCH

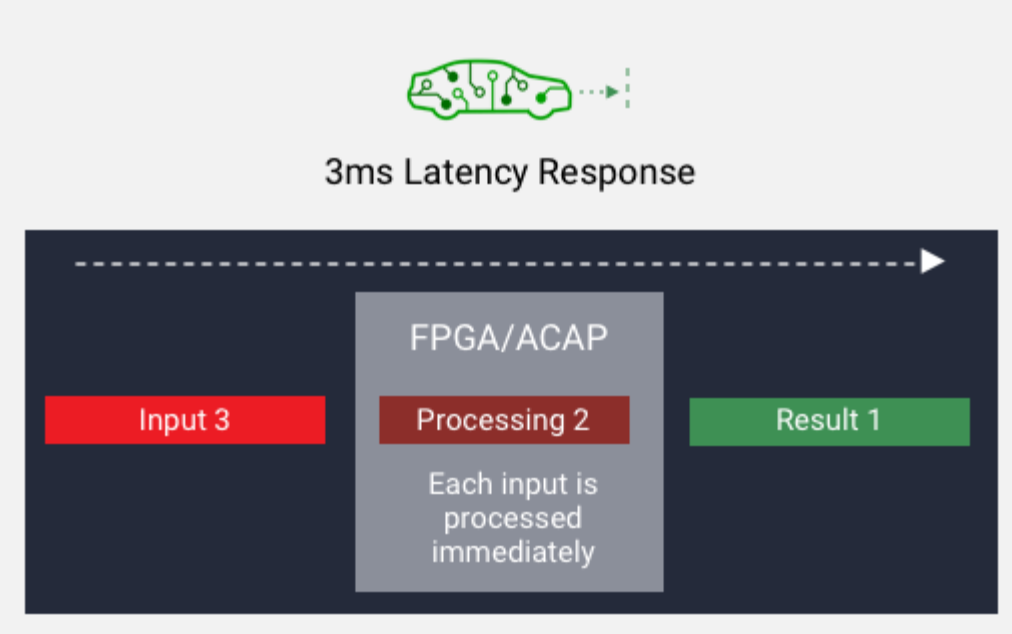
Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., Li, L., Chen, T., Xu, Z., Sun, N. and Temam, O., 2014, December. Dadiannao: A machine-learning supercomputer. In Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (pp. 609-622). IEEE Computer Society.

XILINX.

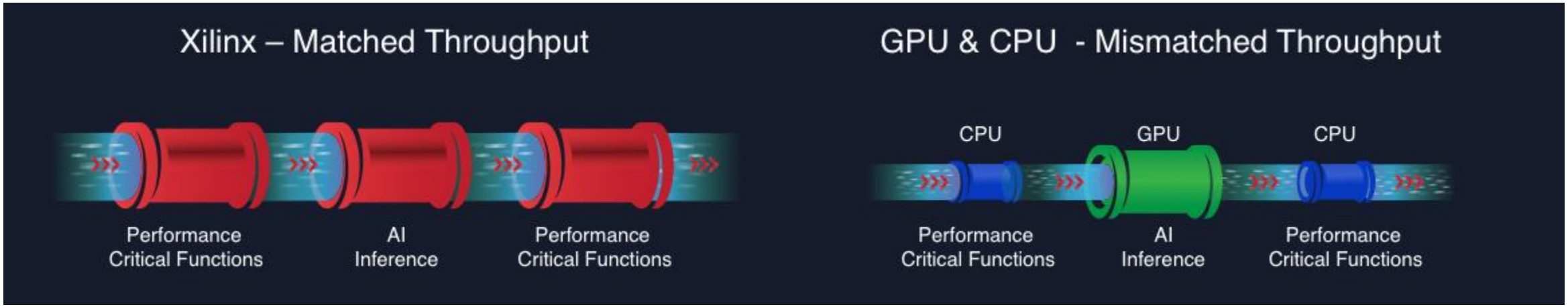
DNN requirements

- Throughput
- Latency
- Energy
- Power
- Cost

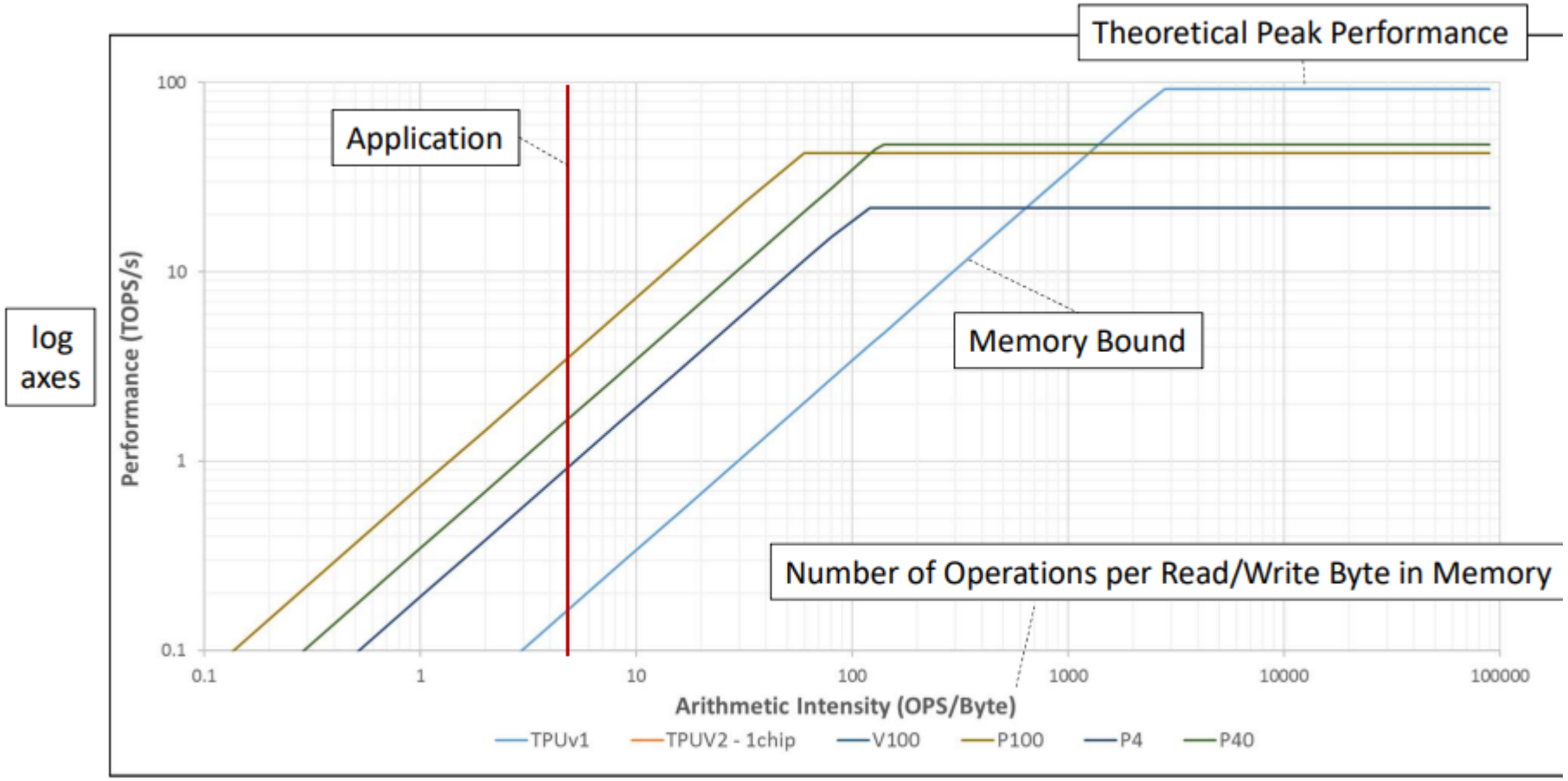




- Optimized hardware acceleration of both AI inference and other performance-critical functions by tightly coupling custom accelerators into a dynamic architecture silicon device.
- This delivers end-to-end application performance that is significantly greater than a fixed-architecture AI accelerator like a GPU;



Roofline

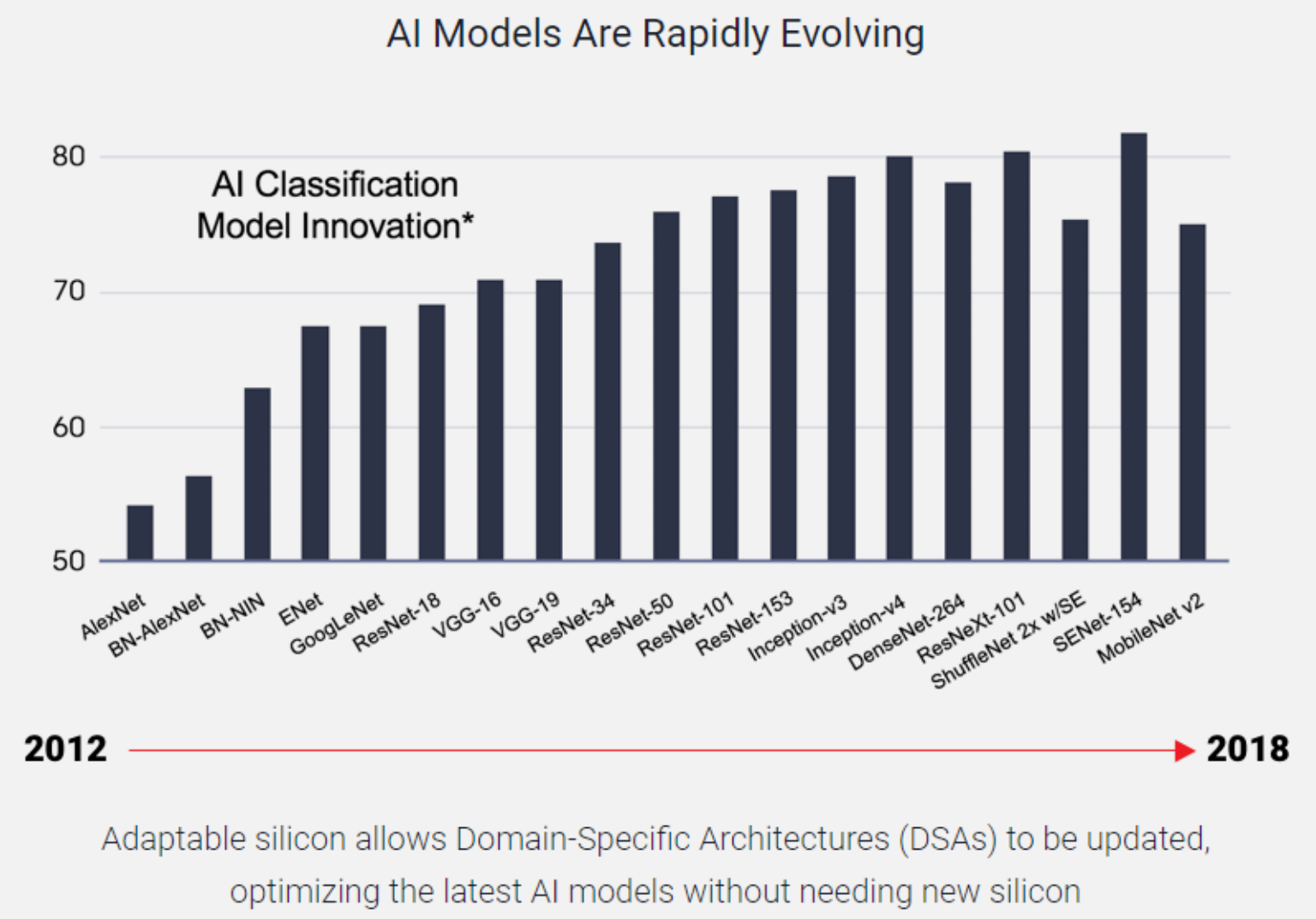


log axes

>> 27

*Williams, S., Waterman, A. and Patterson, D., 2009.
Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM

Adaptive to new models



FPGAs for DNN

- **The xDNN processing engine has dedicated execution paths for each type of command (download, conv, pooling, element-wise, and upload). This allows for convolution commands to be run in parallel with other commands if the network graph allows it**

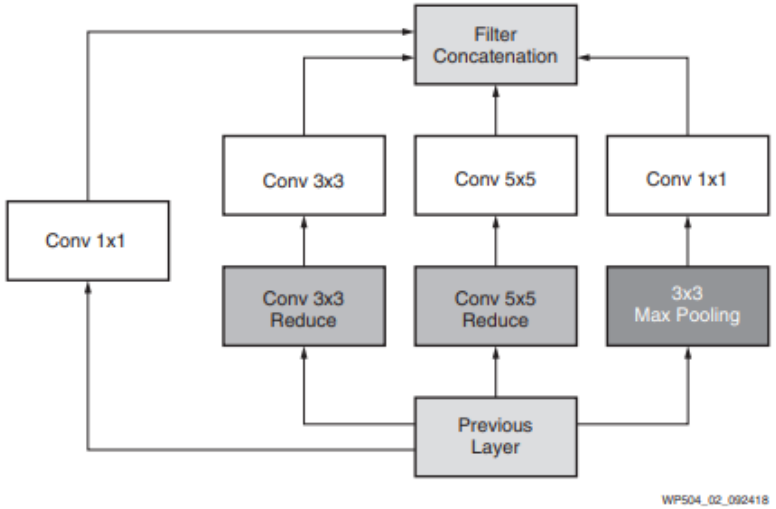
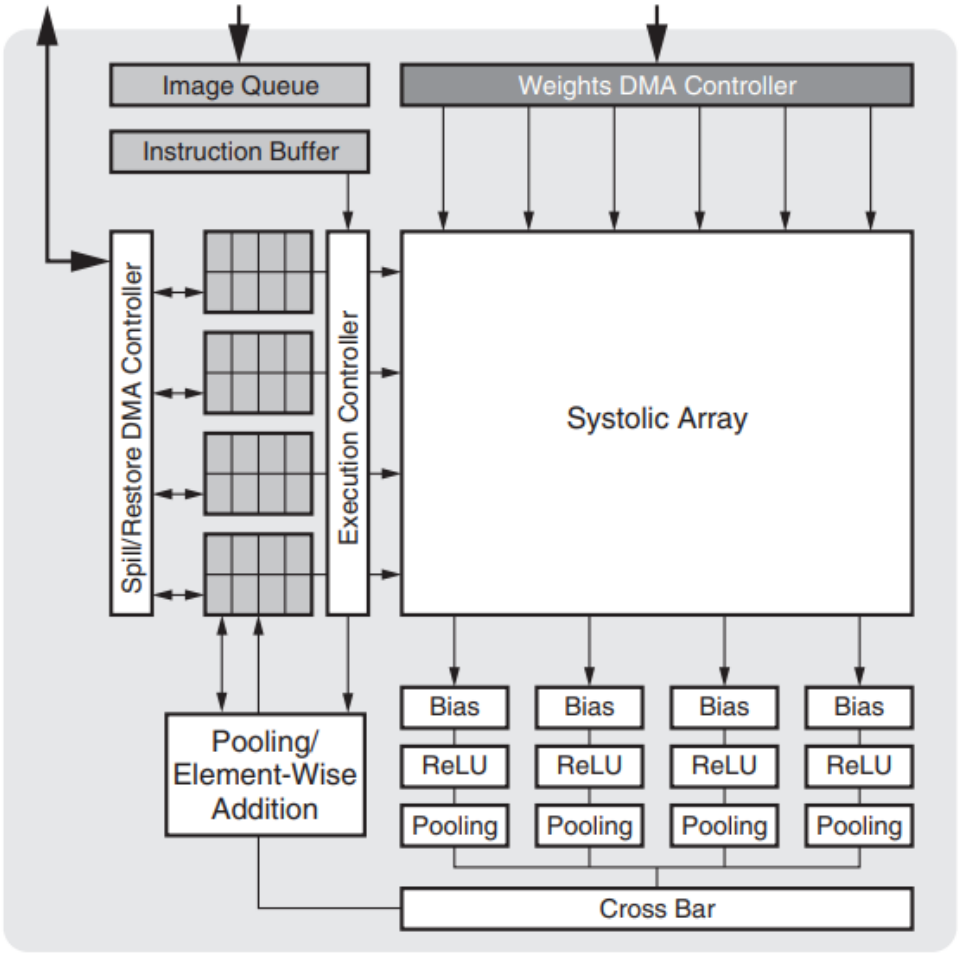
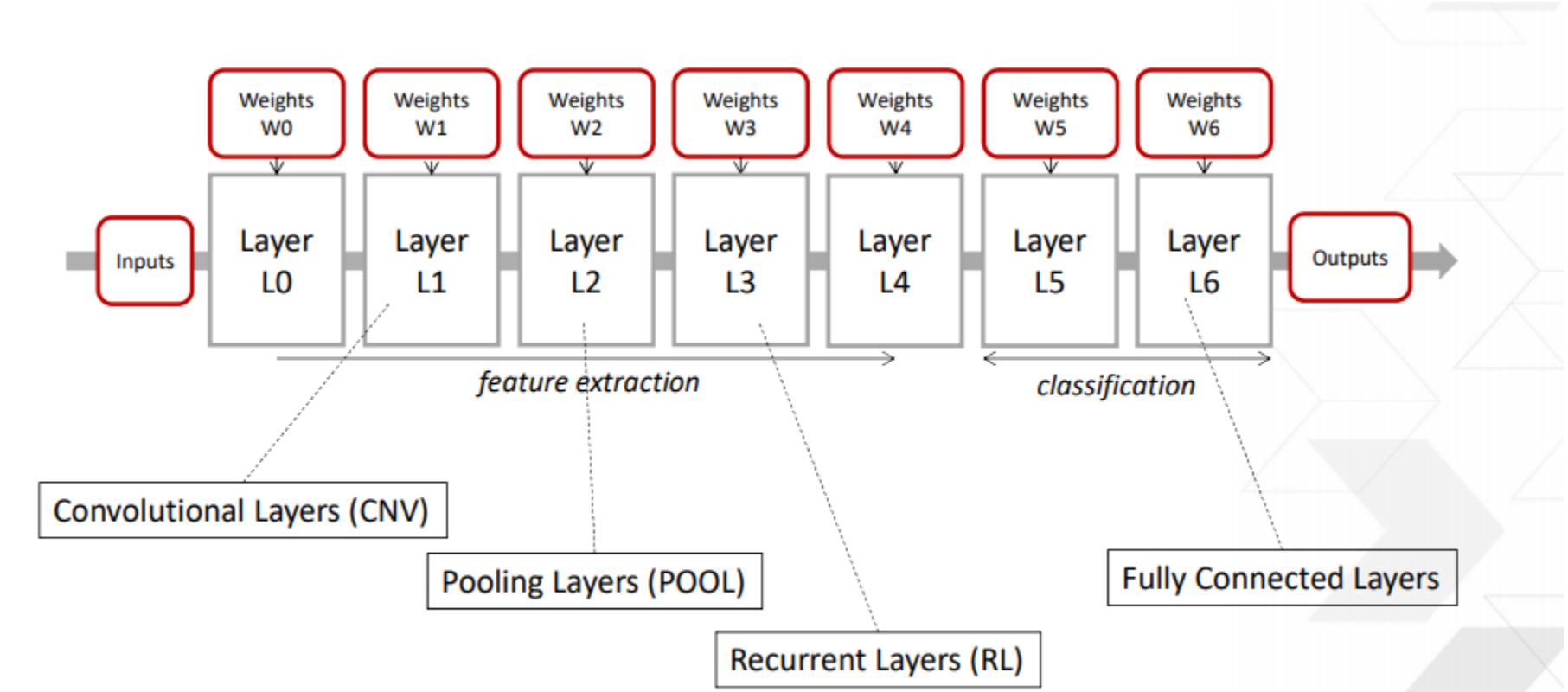


Figure 2: **Inception Layer in GoogLeNet v1**



DNN layers



https://www.xilinx.com/publications/events/machine-learning-live/colorado/HotChipsOverview.pdf

CPU-FPGA

- **Even though the xDNN processing engine supports a wide range of CNN operations, new custom networks are constantly being developed—and sometimes, select layers/instructions might not be supported by the engine in the FPGA. Layers of networks that are not supported in the xDNN processing engine are identified by the xfDNN compiler and can be executed on the CPU. These unsupported layers can be in any part of the network—beginning, middle, end, or in a branch.**

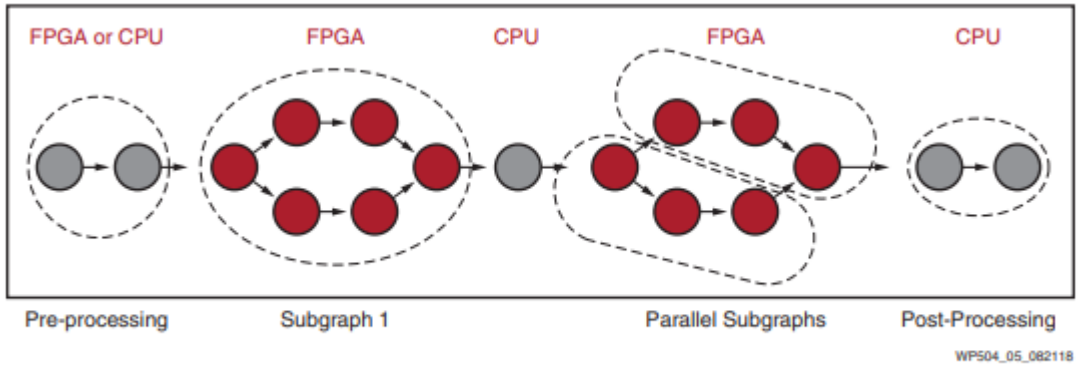


Figure 5: **Processing Partitioned by the Compiler**

CPU-FPGA

- networks and models are prepared for deployment on xDNN through Caffe, TensorFlow, or MxNet.
- FPGA supports layers for xDNN while running unsupported layers on the CPU.

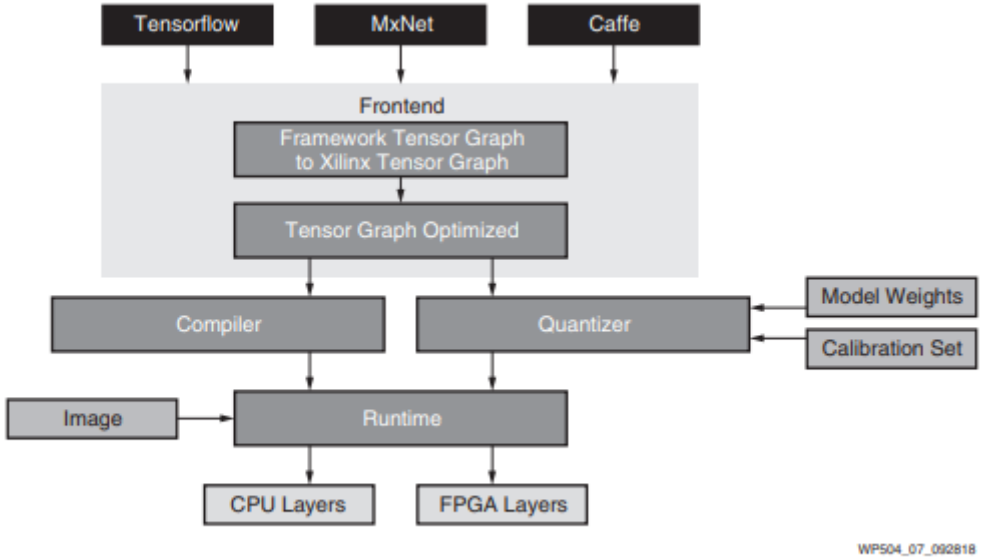


Figure 7: **xfDNN Flow Diagram**

Optimized architecture

- **Network optimization by fusing layers, optimizing memory dependencies in the network, and pre-scheduling the entire network. This removes CPU host control bottlenecks.**

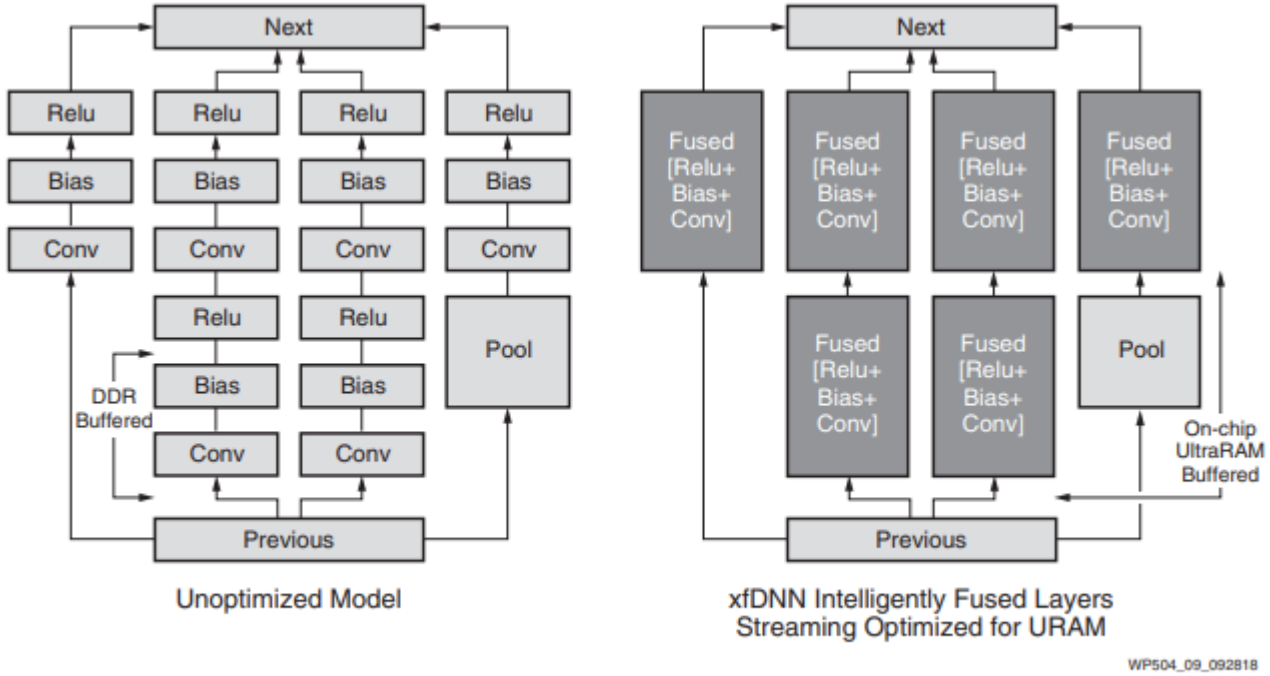
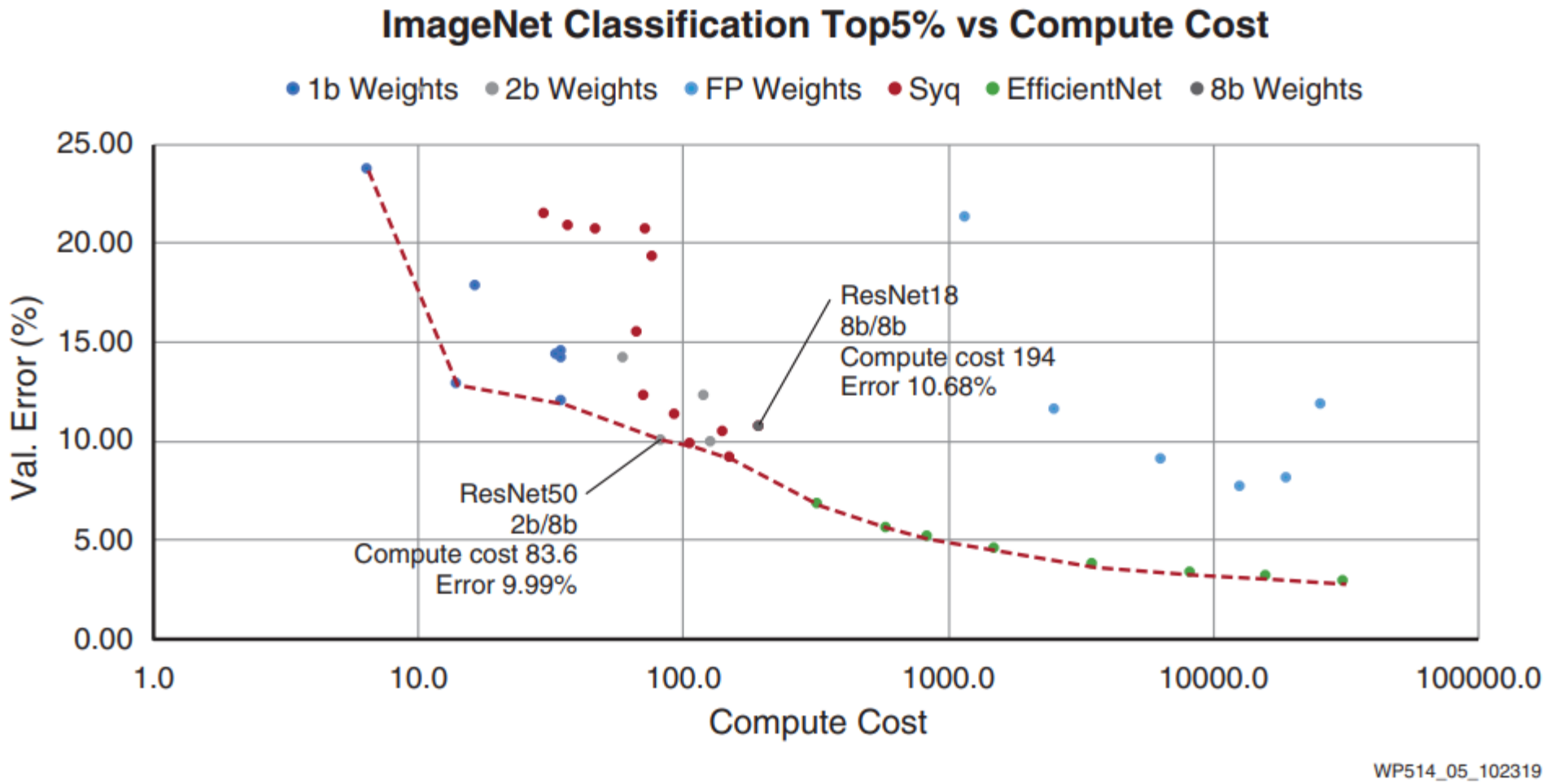


Figure 9: **xfDNN Compiler Optimizations**

DNN tradeoffs

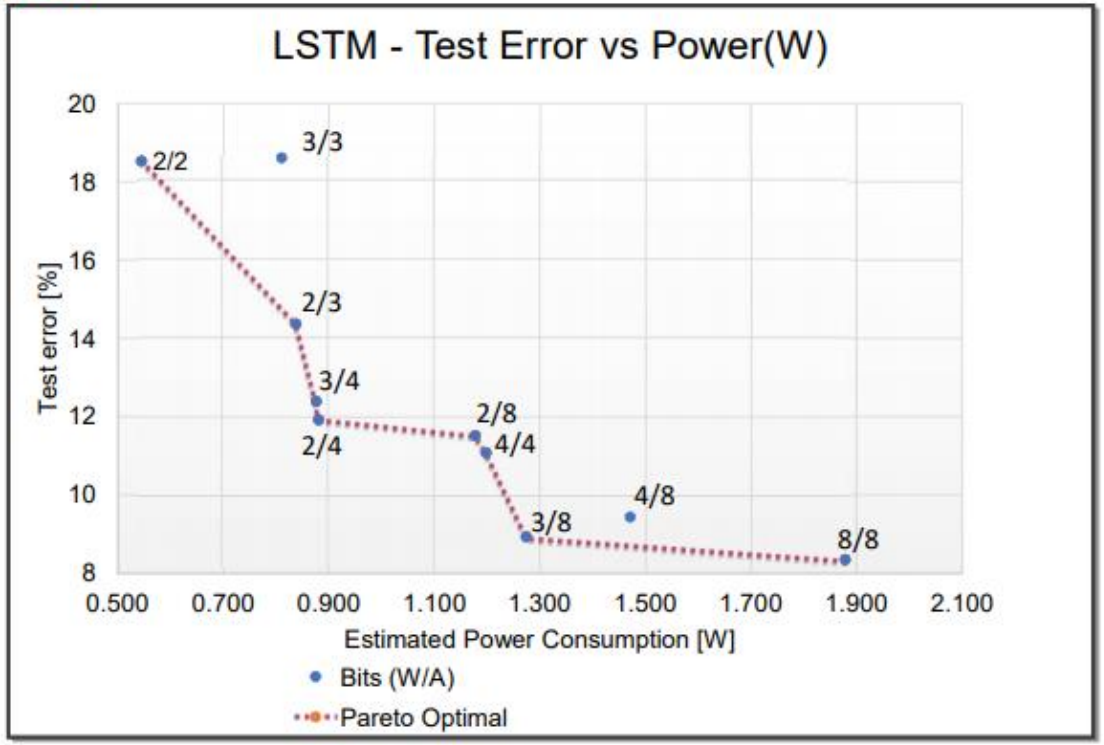


https://www.xilinx.com/support/documentation/white_papers/wp514-emerging-dnn.pdf

Precision vs Performance vs power

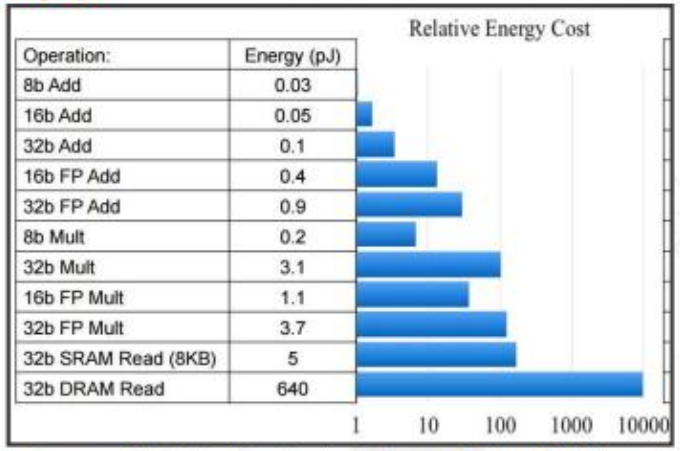
Reducing Precision Inherently Saves Power

FPGA:



Target Device ZU7EV • Ambient temperature: 25 °C • 12.5% of toggle rate • 0.5 of Static Probability • Power reported for PL accelerated block only

ASIC:



Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

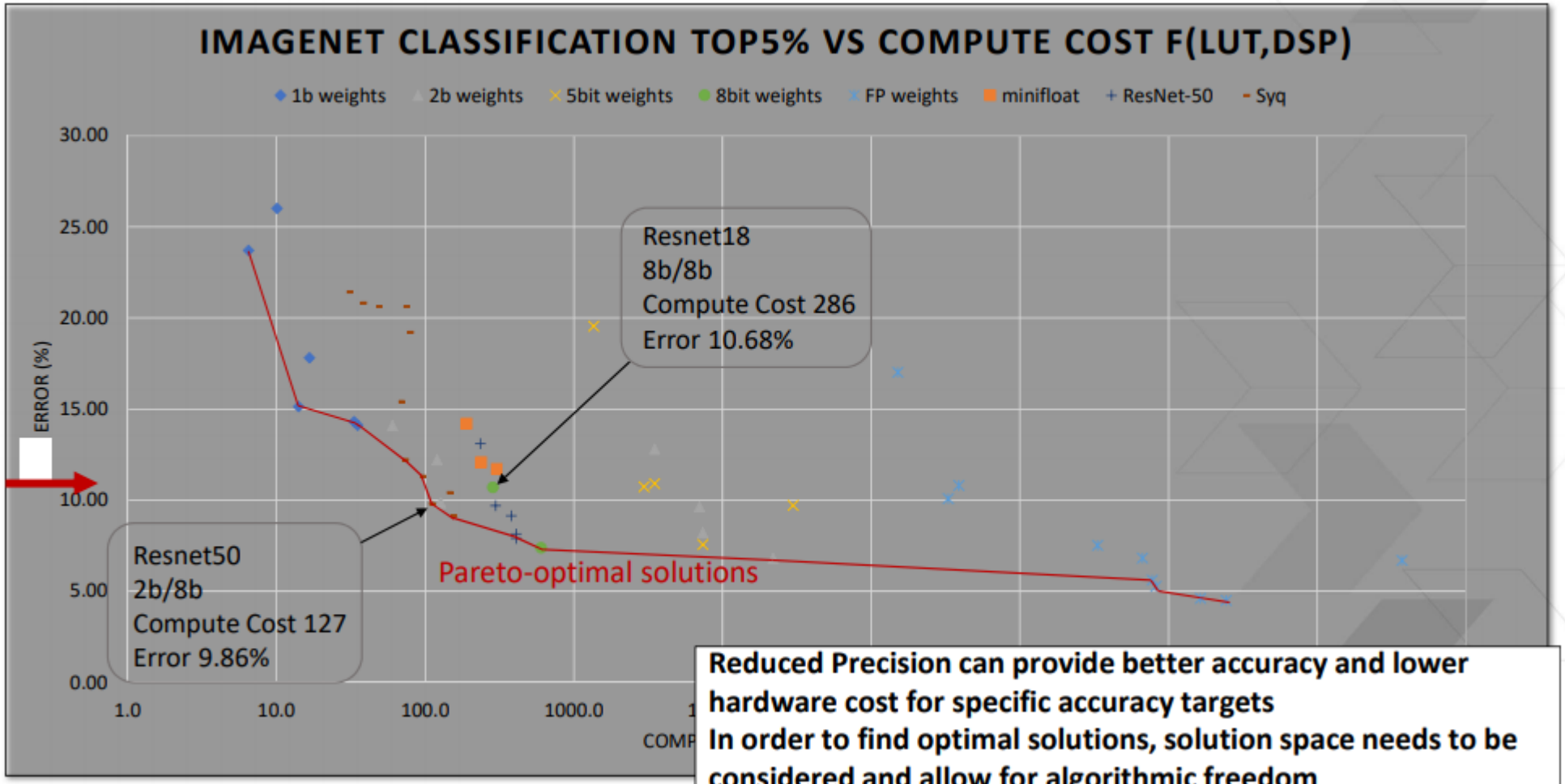
*Source: Bill Dally (Stanford), Cadence **Embedded Neural Network Summit**, February 1, 2017*

>> 35

Rybalkin, V., Pappalardo, A., Ghaffar, M.M., Gambardella, G., Wehn, N. and Blott, M. "FINN-L: Library Extensions and Design Trade-off Analysis for Variable Precision LSTM Networks on FPGAs."

XILINX.

Design Space trade offs



Reduced Precision can provide better accuracy and lower hardware cost for specific accuracy targets
In order to find optimal solutions, solution space needs to be considered and allow for algorithmic freedom

Table I: FPGA and GPU comparison breal

	Runtime (s)			perf/W
Kernel	FPGA	GPU	ratio	ratio
Hotspot	88,593	12,097	0.14	0.59
GICOV	148	438	2.97	7.76
Dilate	234	347	1.48	4.51
MGVF	89,715	11,816	0.13	0.50
SRAD	1,950	1,790	0.92	4.52
BP-1	536	371	0.69	3.10
BP-2	1,995	358	0.18	0.58
StepFactor	4,004	607	0.15	0.58
Flux	145	11	0.08	0.35
LUD	181,055	9,042	0.05	0.17
Kmeans	16,975	3,211	0.19	0.62
KNN	2,538	258	0.10	0.32
SC	15,464	1,187	0.08	0.35
NW	48	362	7.54	19.29
PF	28,750	24,680	0.86	2.85

J. Cong et al., Understanding Performance Differences of FPGAs and GPUs

Winners

Feature	Analysis	Winner
DNN Training	GPU floating point capabilities are greater	GPU
DNN Inference	FPGA can be customized, and has lower latency	FPGA
Large data analysis	CPUs support largest memory and storage capacities. FPGAs are good for inline processing.	CPU/FPGA
Timing latency	Algorithms implemented on FPGAs provide deterministic timing, can be an order of magnitude faster than GPUs	FPGA
Processing/Watt	Customized designs can be optimal	FPGA
Processing/$$	GPUs win because of large processing capabilities. FPGA configurability enables use in a broader acceleration space.	GPU/FPGA
Interfaces	FPGA can implement many different interfaces	FPGA
Backward compatibility	CPUs have more stable architecture than GPUs. Migrating RTL to new FPGAs requires some work.	CPU
Ease of change	CPUs and GPUs provide an easier path to changes to application functionality.	GPU/CPU
Customization	FPGAs provide broader flexibility	FPGA
Size	CPU and FPGA's lower power consumptions leads to smaller volume solutions	CPU/FPGA
Development	CPUs are easier to program than GPUs, both easier than FPGA	CPU

Figure 3 Summary of CPU, GPU, and FPGA comparison

https://www.semanticscholar.org/paper/Unified-Deep-Learning-with-CPU%2C-GPU%2C-and-FPGA-Rush-Sirasao/64c8428e93546479d44a5a3e44cb3d2553eab284#extracted

Links, more info

FPGA-based Accelerators of Deep Learning Networks for Learning and Classification: A Review

AHMAD SHAWAHNA[1], SADIQ M. SAIT[1,2], (Senior Member, IEEE), AND AIMAN EL-MALEH[1], (Member, IEEE)