MLconf Online™

Accelerated ML, instantly

Chris Kachris

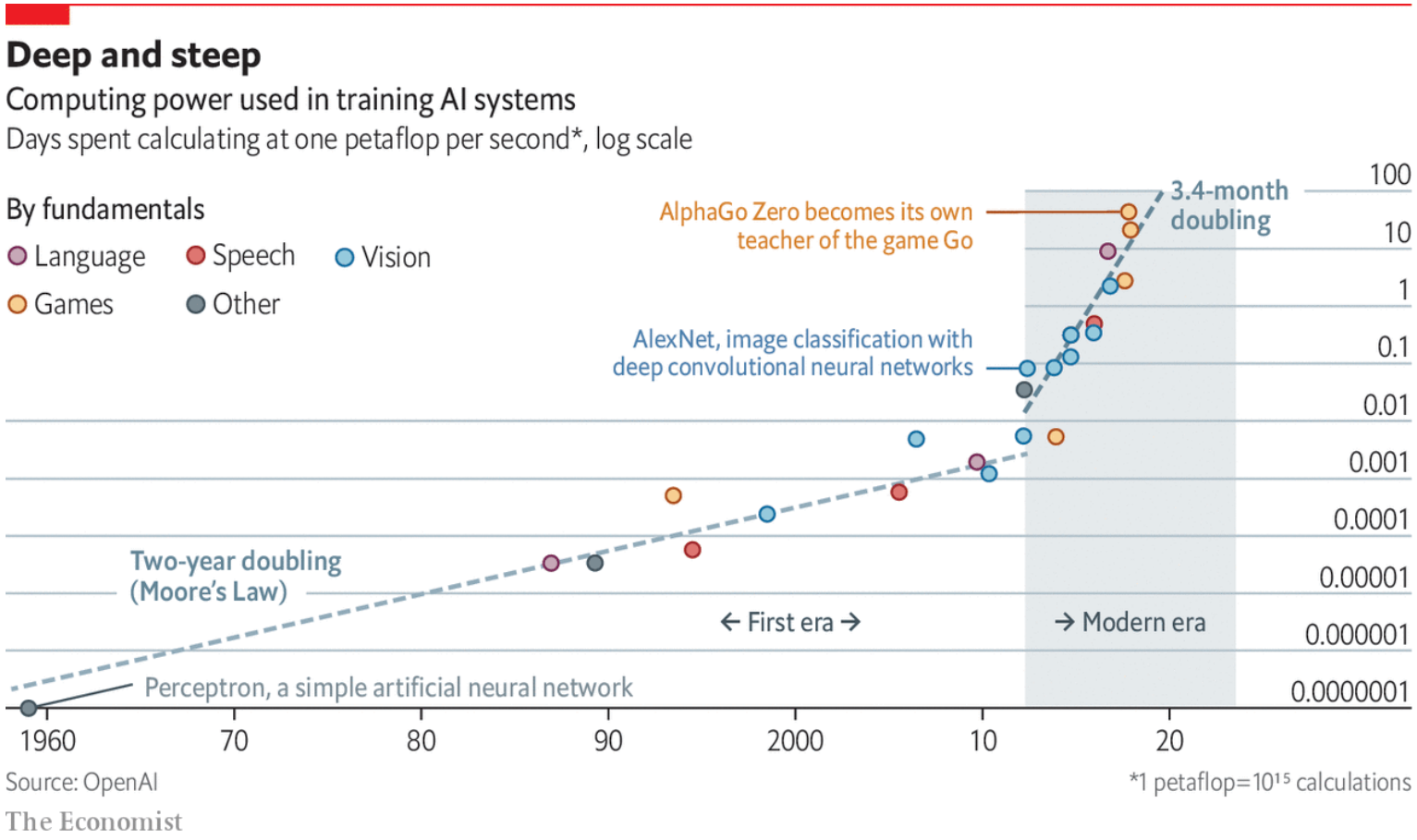
CEO, Co-founder

Need for Computing power

"**computing power** needed to carry out machine learning neural networks is **doubling every 3.5 months**."

Cliff Young, Google

New ML models need more powerful platform



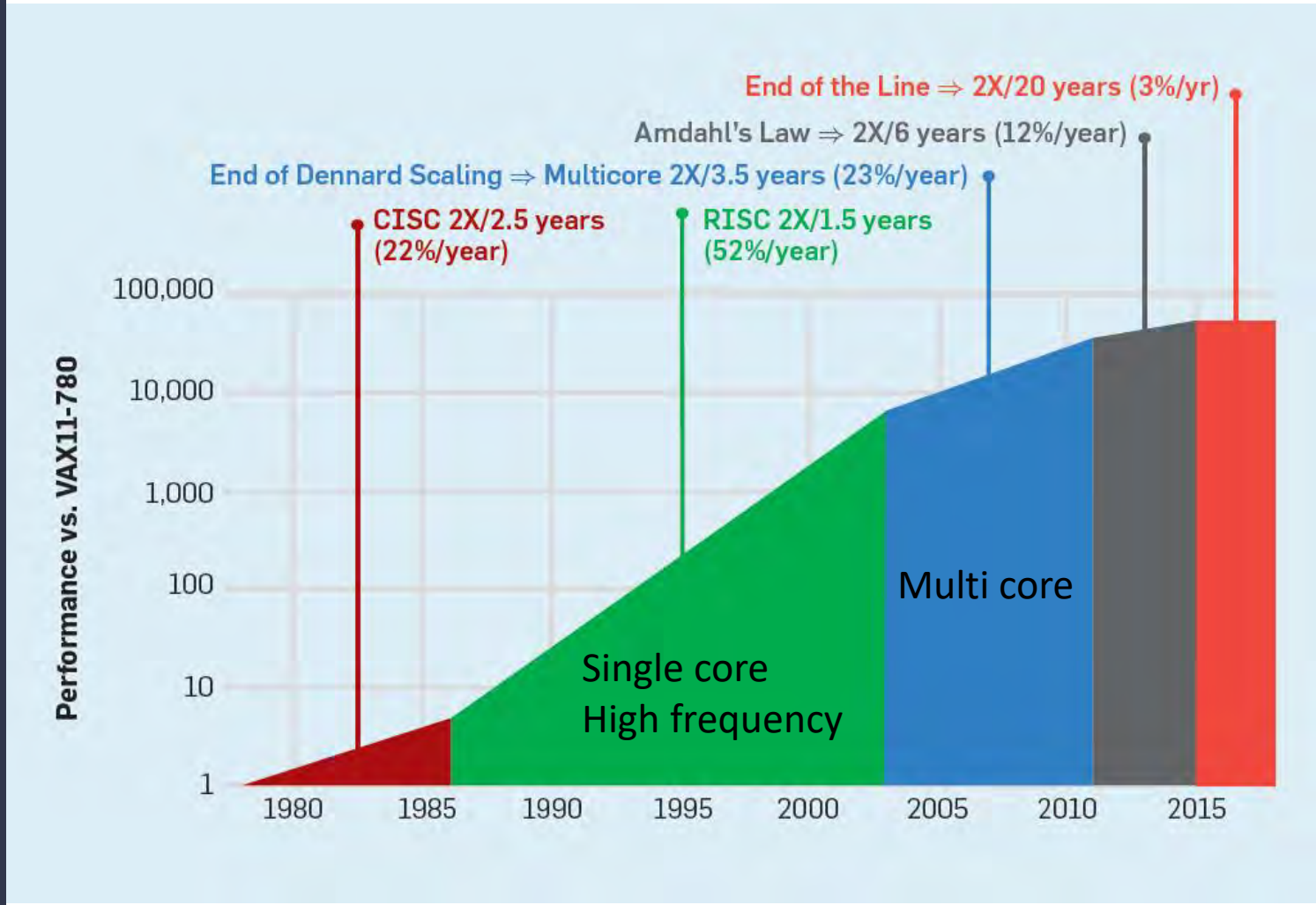
https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem

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

What's left for faster computing?



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

CPUs cannot keep increasing the performance as in the past

What's Left?

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

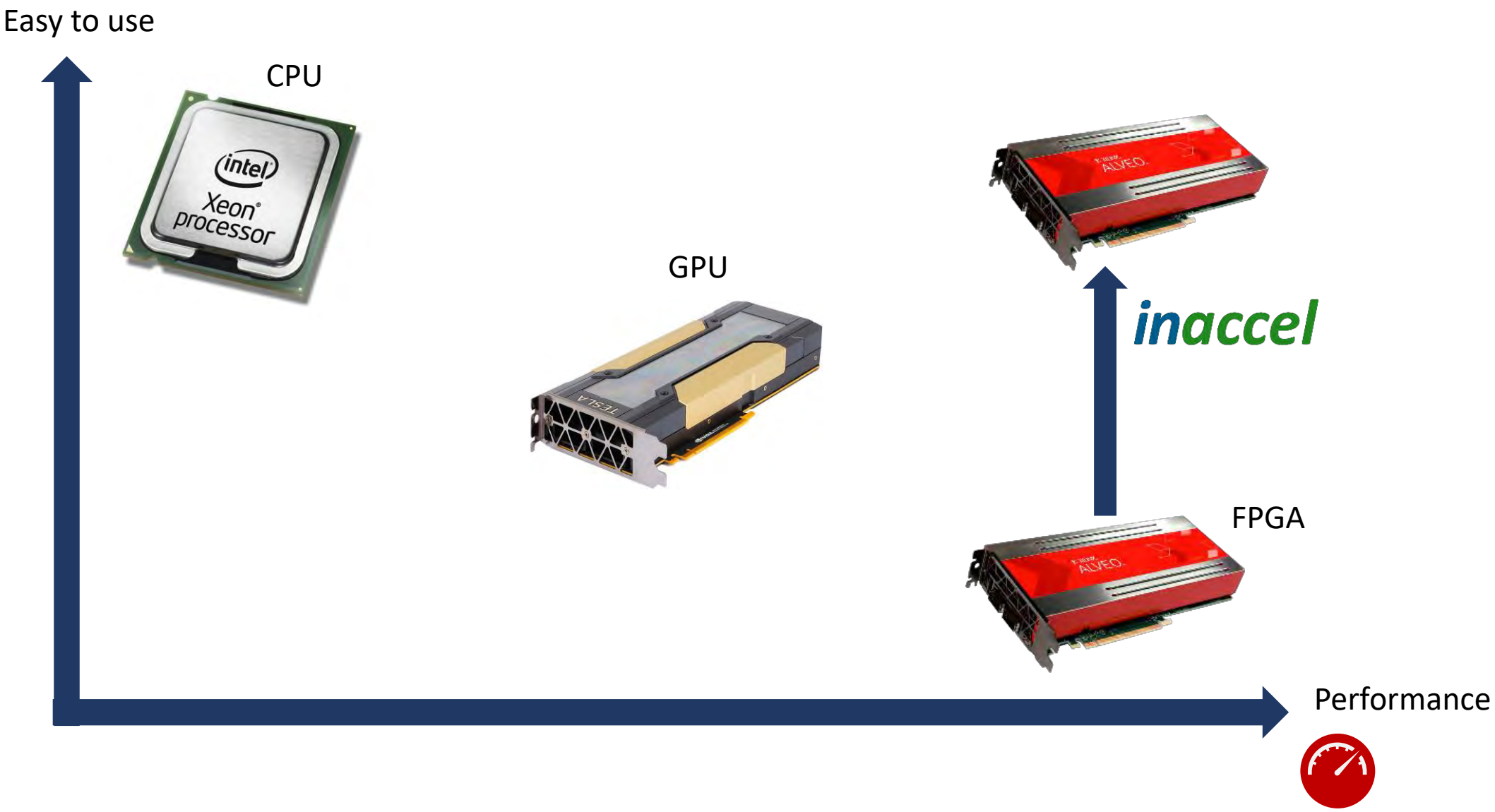
- Just do a few tasks, but extremely well

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

David Patterson, 2019

Next **WAVE** of **COMPUTING: Accelerators**

Accelerators



Accelerated Machine Learning

https://inaccel.com/accelerated-data-science/

What we do: Accelerated Machine Learning

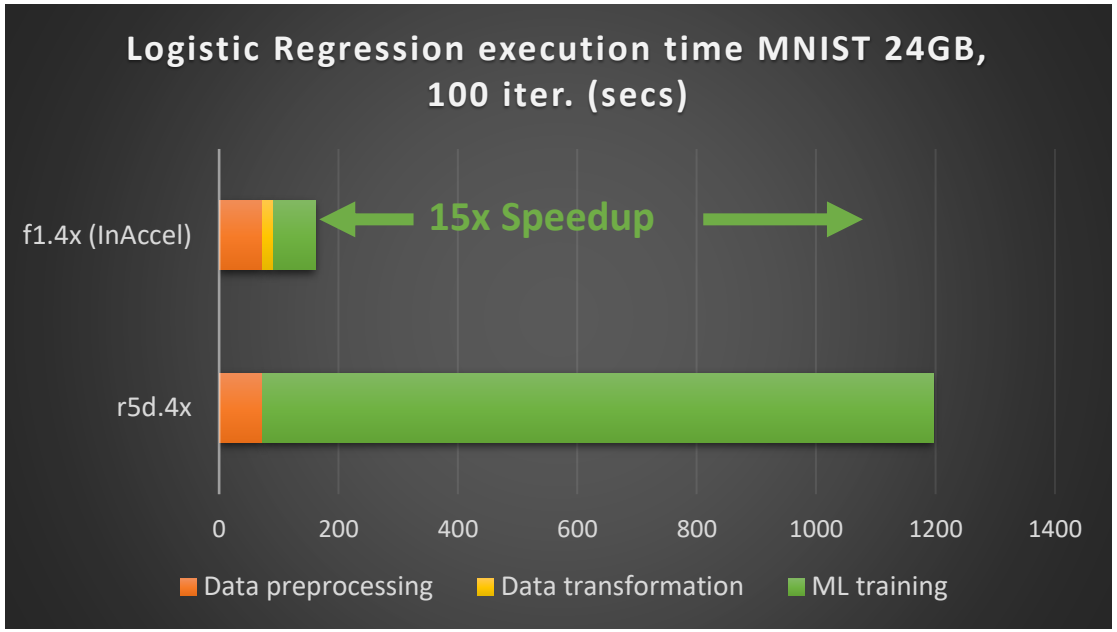
Help companies **speedup** their ML applications by using **accelerators** (FPGAs) seamlessly (**ML as a Service**):

10x – 20x Faster

2x Lower cost

Zero code changes

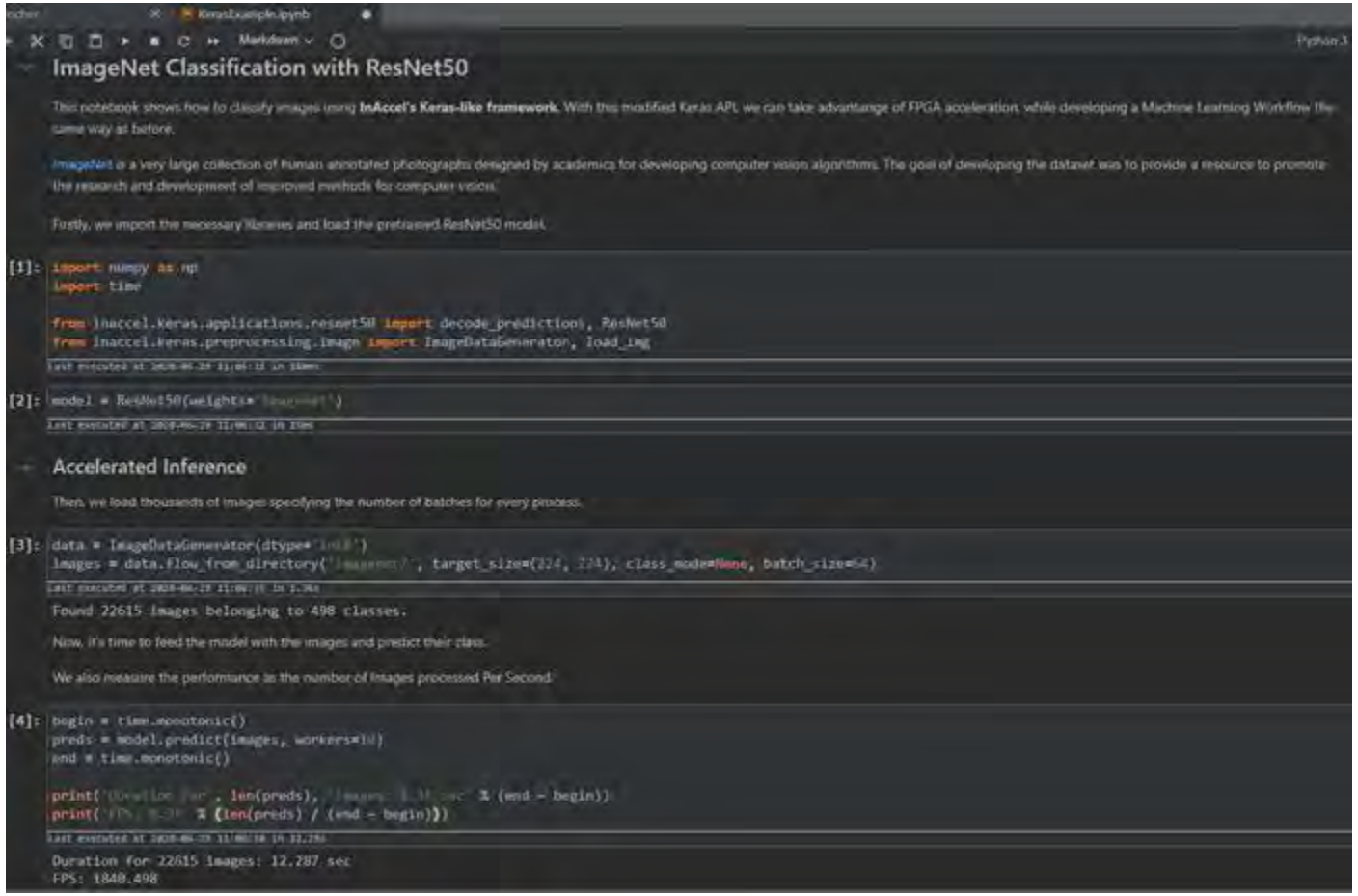
Integrated Accelerated Machine Learning/DNN Platform



https://www.youtube.com/watch?v=hDLAYNsF39s

Same tools (Python), Faster results

InAccel Studio is a unique portal for Accelerated Machine learning



https://inaccel.com/accelerated-machine-learning/

- **Familiar tools**
- **Faster results**
  - 10x-20x faster
  - Zero code changes

Instant Acceleration

Just add **inaccel** and enjoy **15x faster** execution

File Edit View Run Kernel Tabs Settings Help

InAccel Cloud light

/ shared / ml /

Name	Last Modified
data	12 days ago
img	19 days ago
LogisticRegression.ipynb	a month ago
MLlibPipelines.ipynb	12 days ago
NaiveBayes.ipynb	a month ago
XGBoost.ipynb	19 days ago

LogisticRegression.ipynb

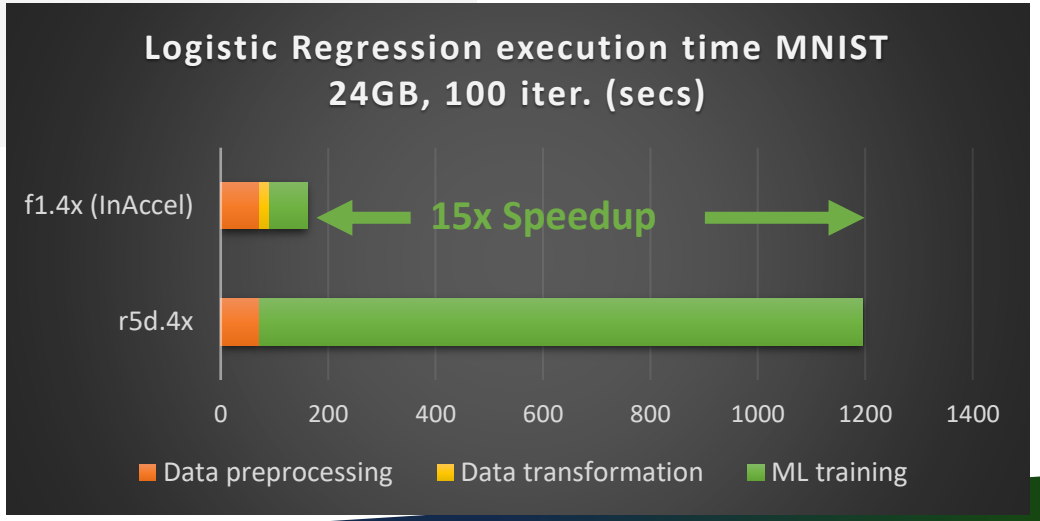
Markdown Pythor

Logistic Regression Hyperparameter Tuning using FPGAs

This notebook shows how to train and apply many accelerated sklearn models, with a k-fold cross validation and hyperparameter tuning step.

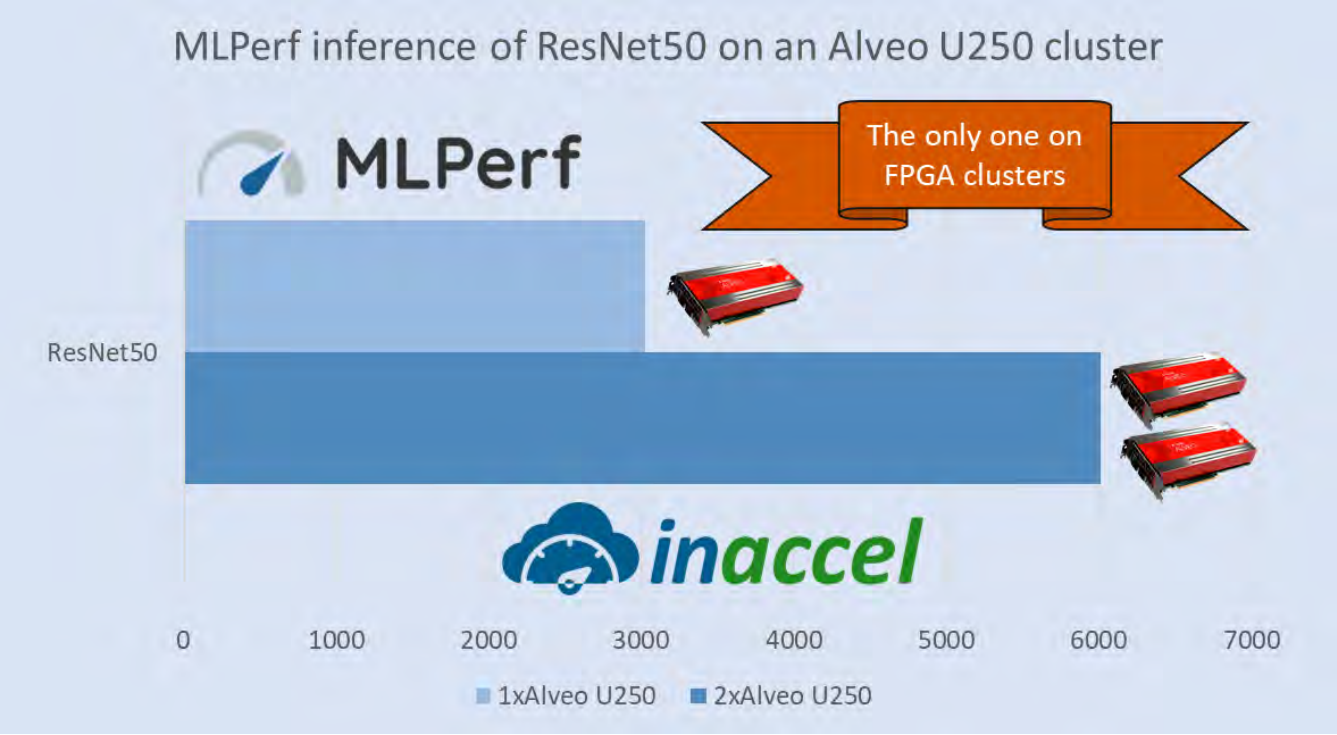
[ ]:
```
from inaccel.sklearn.linear_model import LogisticRegression
from sklearn.datasets import fetch_openml
from sklearn.linear_model import LogisticRegression as LogisticRegressionCPU
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
```

15x faster ML training



MLPerf

ID	Submitter	System	Processor	#	Accelerator	#	Software	Model Used	Results				
									Task	Image classification			
									Data	ResNet			
									Accuracy (%FP32 ref				99.00
									Scenario	Single Stream	Multiple Stream	Server	Offline
									Units	latency in ms	streams	query/s	samples/s
CATEGORY: Available													
0.7-158	Deci	n2-highcpu-16	Intel(R) Xeon(R) Cascade Lake CPU	1			openvino_2020.4.287	resnet		2.13			1,093
0.7-159	Deci	n2-highcpu-2	Intel(R) Xeon(R) Cascade Lake CPU	1			openvino_2020.4.287	resnet		6.77			148
0.7-160	dividiti	AWS g4dn.4xlarge	Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz	1			OpenVINO 2020	resnet50		2.67			
0.7-161	dividiti	AWS g4dn.4xlarge	Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz	1			OpenVINO 2020	ssd-resnet34					
0.7-162	InAccel	alveo2.ethz.ch	Intel(R) Xeon(R) Gold 6234	2	Xilinx Alveo U250	1	InAccel Keras 2.3.1.2	resnet		7.00	58		3,026
0.7-163	InAccel	alveo2.ethz.ch	Intel(R) Xeon(R) Gold 6234	2	Xilinx Alveo U250	2	InAccel Keras 2.3.1.2	resnet					6,008



Instant Acceleration by more than 10x

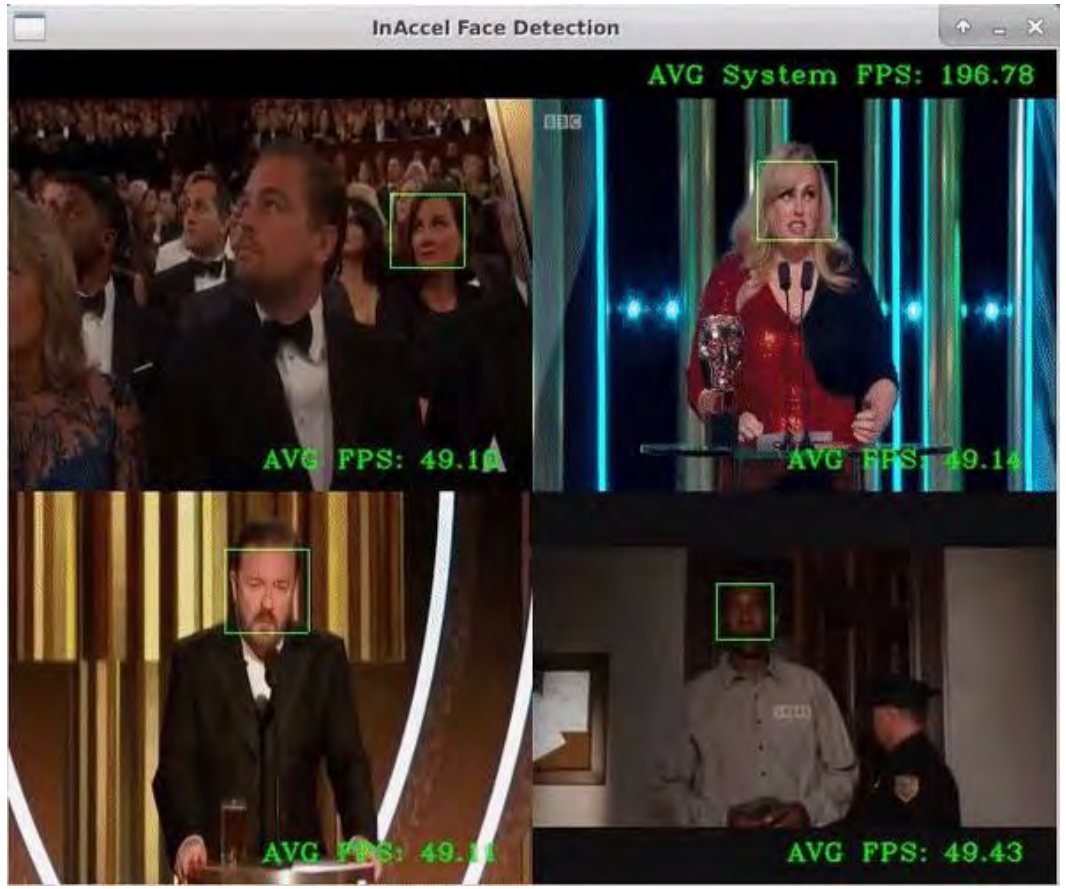
Now we're ready to build a pipeline and fit it. This puts the data through all of the feature processing, model tuning & training we described in a single call.

```
In [*]: from pyspark.ml import Pipeline

        pipeline = Pipeline(stages=[labelIndexer, featuresScaler, cv, indexToLabel])

        %time model = pipeline.fit(train)
```

The next cell converts the test set from LibSVM to **Parquet** memory format, in order to serve as our streaming source.

15x faster

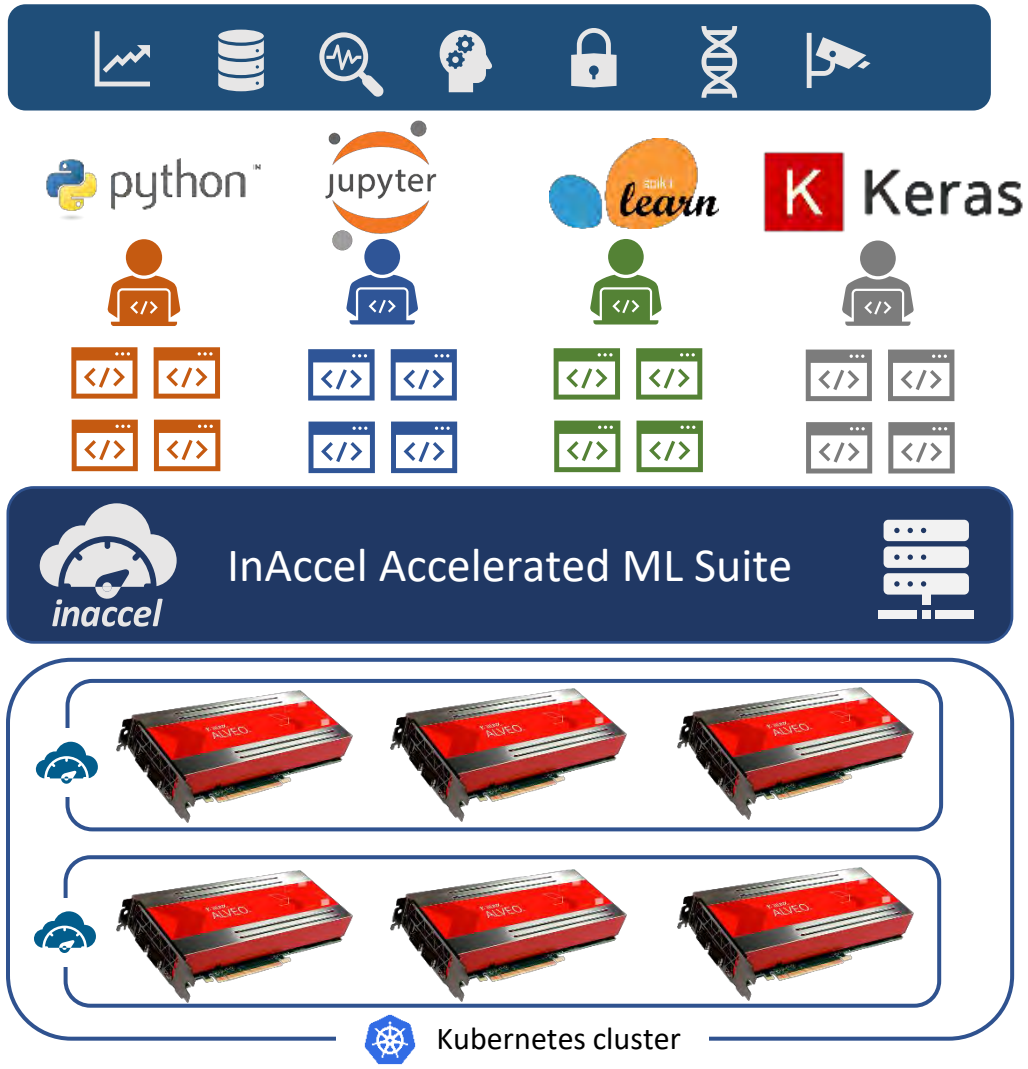
Video analytics: Face detection



1700fps per server

How we do it

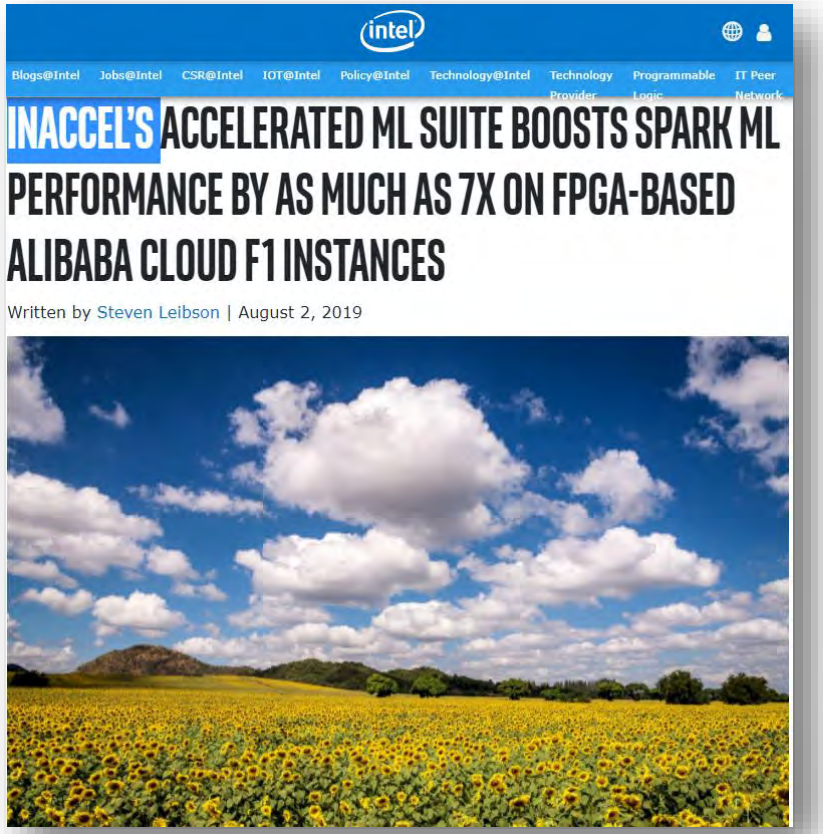
- We leverage the **power of programmable accelerators (FPGA)**
- A unique platform for easy deployment, scaling and resource management of FPGA
- **Pricing model**: Pay-as-you-go (Subscription)



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

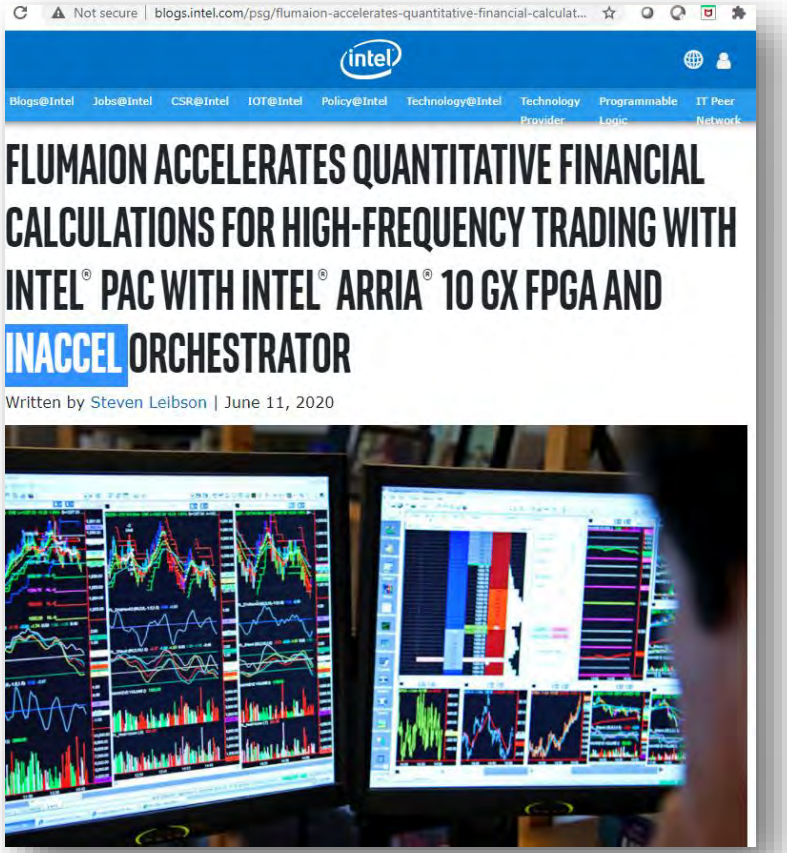
Use cases

Machine Learning



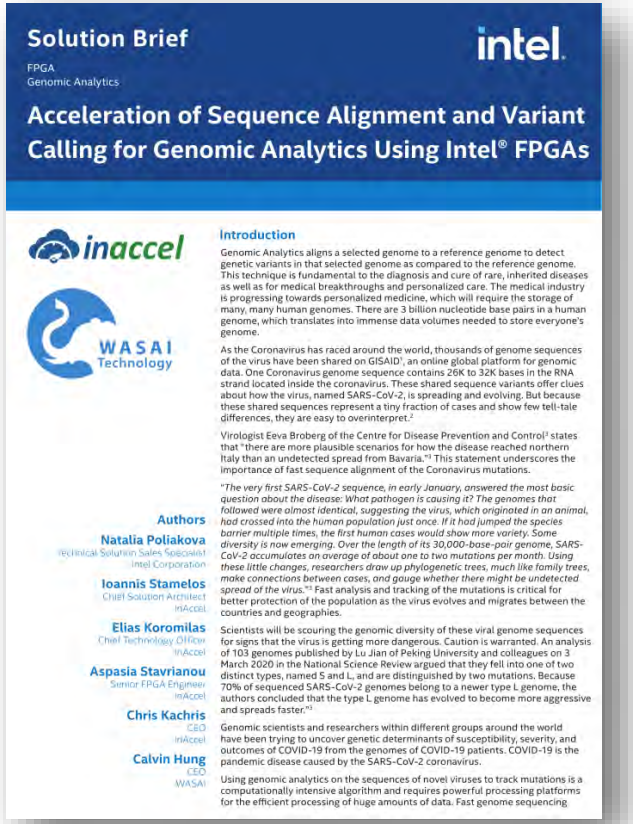
https://blogs.intel.com/psg/inaccels-accelerated-ml-suite-boosts-spark-ml-performance-by-as-much-as-7x-on-fpga-based-alibaba-cloud-f1-instances/

Quantitative Finance



https://blogs.intel.com/psg/flumaion-accelerates-quantitative-financial-calculations/

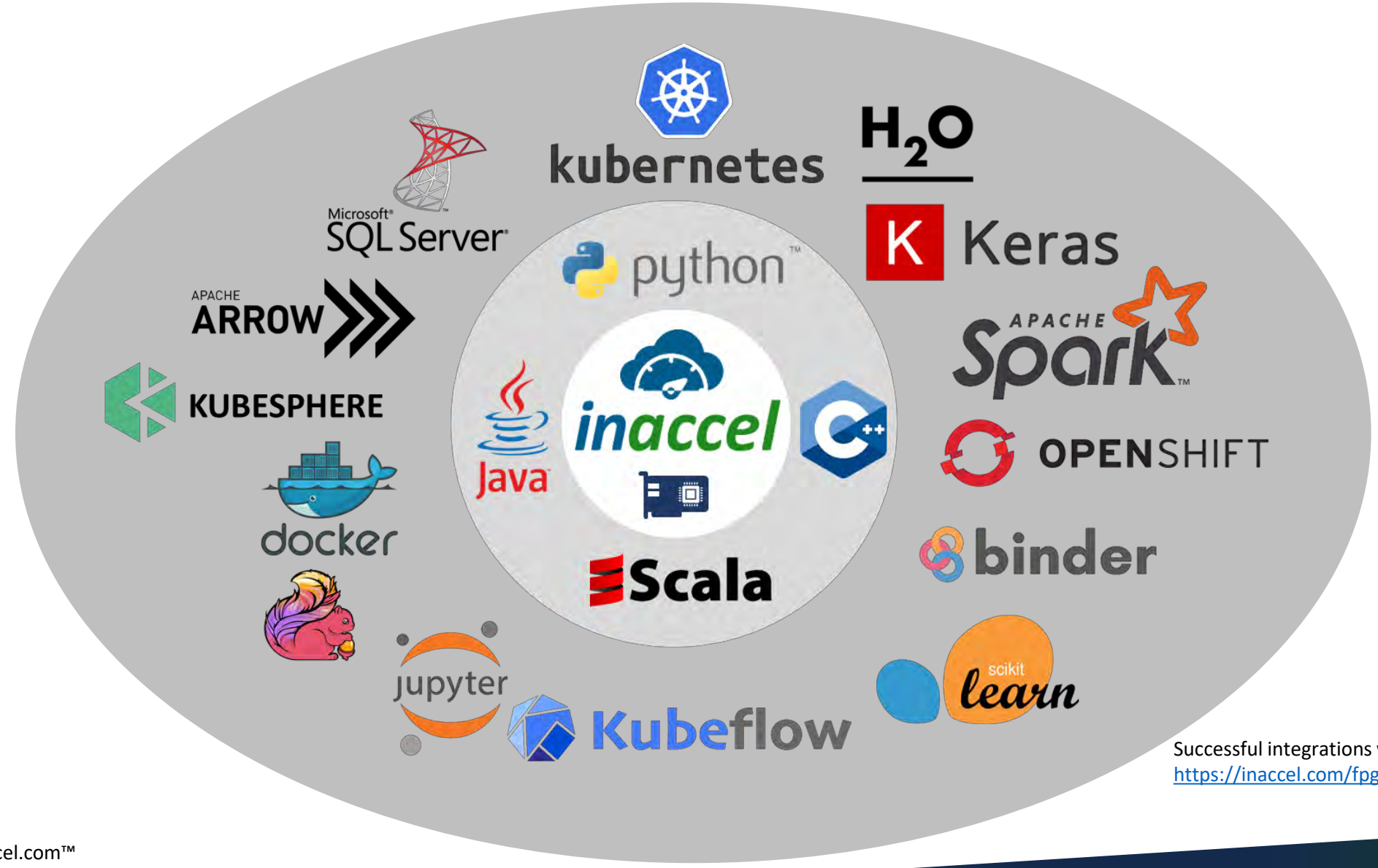
Genomics



On line free Data Science platforms

	GPU	FPGA
inaccel		 10x
Azure Notebooks	NVIDIA	
Amazon SageMaker	NVIDIA	
colab	NVIDIA	
DataCamp	NVIDIA	
kaggle	NVIDIA	

Seamless Integration with any framework



Successful integrations with Multiple Frameworks
https://inaccel.com/fpga-integration-easier-than-ever/

Market size – FPGAs in Data Centers

InAccel is targeting the Market size for **FPGA Data Center and HPC** Accelerators for **application Acceleration**

- Machine learning
- Genomics
- Quantitative Financial
- Analytics
- Databases
- Security
- Vision

2023

TAM: $500 Million

(Compute acceleration not including smartNIC and storage)

SAM: $200 Million

(Compute acceleration not including smartNIC and storage)

Users

inaccel

inspur

Flumaion

XILINX

SUNG KYUN KWAN UNIVERSITY

National Technical University of Athens

Excelpoint®

ETH Zürich

INTI

NAVER

Alnylam® PHARMACEUTICALS

VAULTMINER

PLC2

NVXL

DINOPLUS

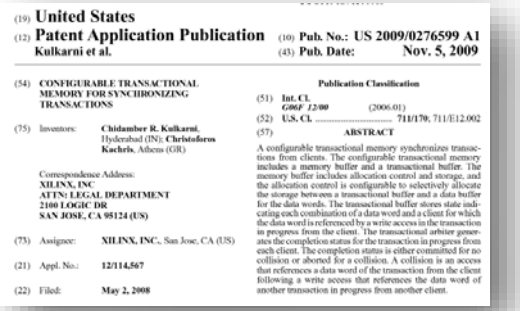
VVDN® TECHNOLOGIES

University of Sarajevo

金仕达 Kingstar

Co-founders Team

- Dr. Chris Kachris
  - Xilinx Research labs (2007)
  - Book: Hardware Accelerators in Data Centers
  - First paper on FPGAs for MapReduce
  - 80+ publications on FPGAs, 1800+ citations
- Elias Koromilas
  - Winner of the Xilinx Open Hardware Contest, Spark on PYNQ 2017
- Ioannis Stamelos
  - Winner of the Xilinx Open Hardware Contest, Spark on PYNQ 2017



(19) United States
(12) Patent Application Publication
Kulkarni et al.
(10) Pub. No.: US 2009/0276599 A1
(43) Pub. Date: Nov. 5, 2009

(54) CONFIGURABLE TRANSACTIONAL MEMORY FOR SYNCHRONIZING TRANSACTIONS

(75) Inventors: Chidamber R. Kulkarni, Hyderabad (IN); Christoforos Kachris, Athens (GR)

Correspondence Address:
XILINX, INC
ATTN: LEGAL DEPARTMENT
2100 LOGIC DR
SAN JOSE, CA 95124 (US)

(73) Assignee: XILINX, INC., San Jose, CA (US)

(21) Appl. No.: 12/114,567

(22) Filed: May 2, 2008

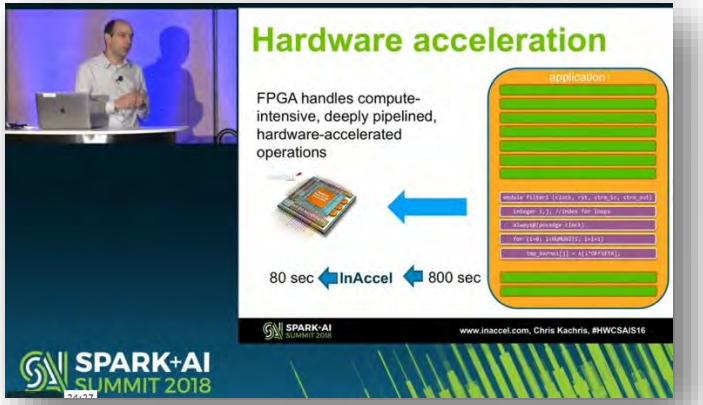
Publication Classification

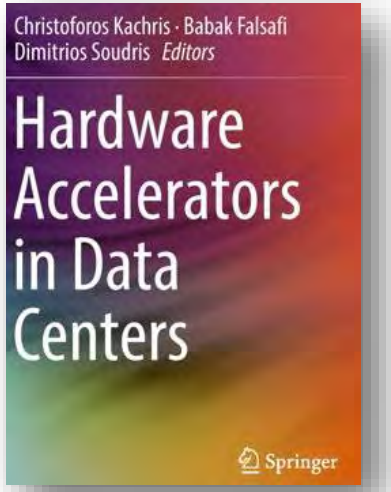
(51) Int. Cl.
G06F 12/00 (2006.01)

(52) U.S. Cl. 711/170; 711/E12.002

(57) ABSTRACT

A configurable transactional memory synchronizes transactions from clients. The configurable transactional memory includes a memory buffer and a transactional buffer. The memory buffer includes allocation control and storage, and the allocation control is configurable to selectively allocate the storage between a transactional buffer and a data buffer for the data words. The transactional buffer stores state indicating each combination of a data word and a client for which the data word is referenced by a write access in the transaction in progress from the client. The transactional arbiter generates the completion status for the transaction in progress from each client. The completion status is either committed for no collision or aborted for a collision. A collision is an access that references a data word of the transaction from the client following a write access that references the data word of another transaction in progress from another client.





InAccel Team

Advisory board

Christos Makiyama
Founder and President
at Silicon Planet Corporation

Genelle Heim
Managing Director at Grayson Hayden Group

(Ex-Vice President of Marketing at Xilinx)

Chris Kachris
CEO, co-founder

Elias Koromilas
CTO, co-founder

Ioannis Stamelos
COO, co-founder

Vasilis Amourgianos
FPGA Engineer

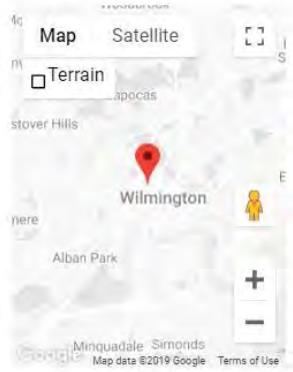
Vangelis Gkiastas
ML Engineer

Aspasia Stavrianou
DevOps engineer

InAccel, Inc. Corporate overview

- Founded in January 2018 (Seed fund: $600 USD in June’18)
- Registered in Delaware, USA
- Membership:

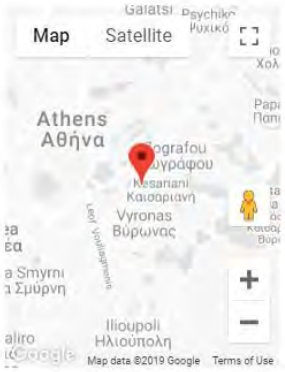
Proudly supported by:



Headquarters

500 Delaware Ave STE 1, #1960
Wilmington, DE 19801
USA

(+1) 408 260 5724



Design Center

Formionos 47
Kesariani 116 33
Athens, Greece

(+30) 211 1825 436

QUESTIONS?

MLconf Online™

THANK YOU!

@inaccel

/in/kachris