



Application Acceleration made simple

www.inaccel.com

info@inaccel.com

What software developers want



% OF RESPONDENTS WHO CONSIDERED THE FEATURE

VERY IMPORTANT

More than one feature could be selected.



91%
PERFORMANCE



69%
EASE OF
DEPLOYMENT



76%
EASE OF
PROGRAMMING



INTEL® FPGA SDK FOR OPENCL™

Source: Databricks, Apache Spark Survey 2016, Report

What software developers want

% OF RESPONDENTS WHO CONSIDERED THE FEATURE

VERY IMPORTANT

More than one feature could be selected.



91%
PERFORMANCE



69%
EASE OF
DEPLOYMENT



76%
EASE OF
PROGRAMMING

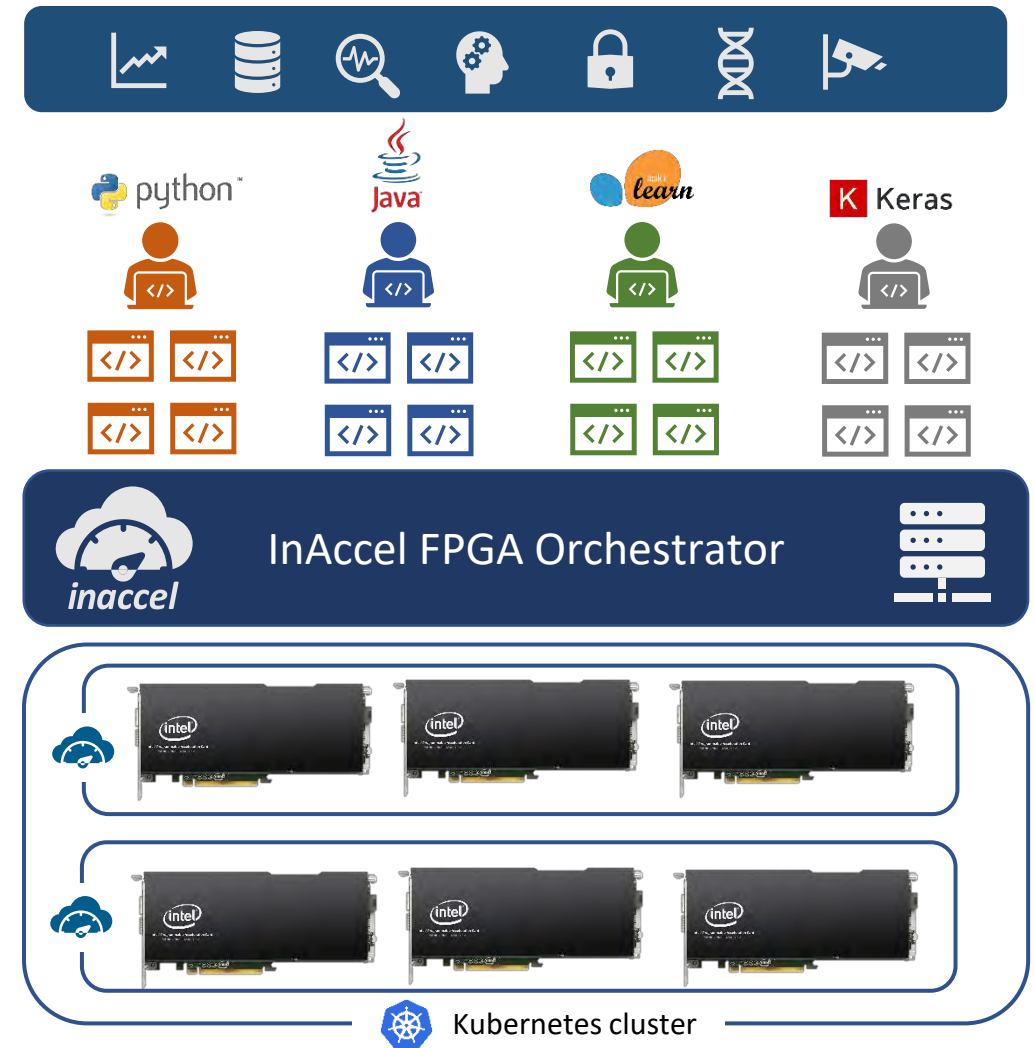
INTEL® FPGA SDK FOR OPENCL™



Source: Databricks, Apache Spark Survey 2016, Report

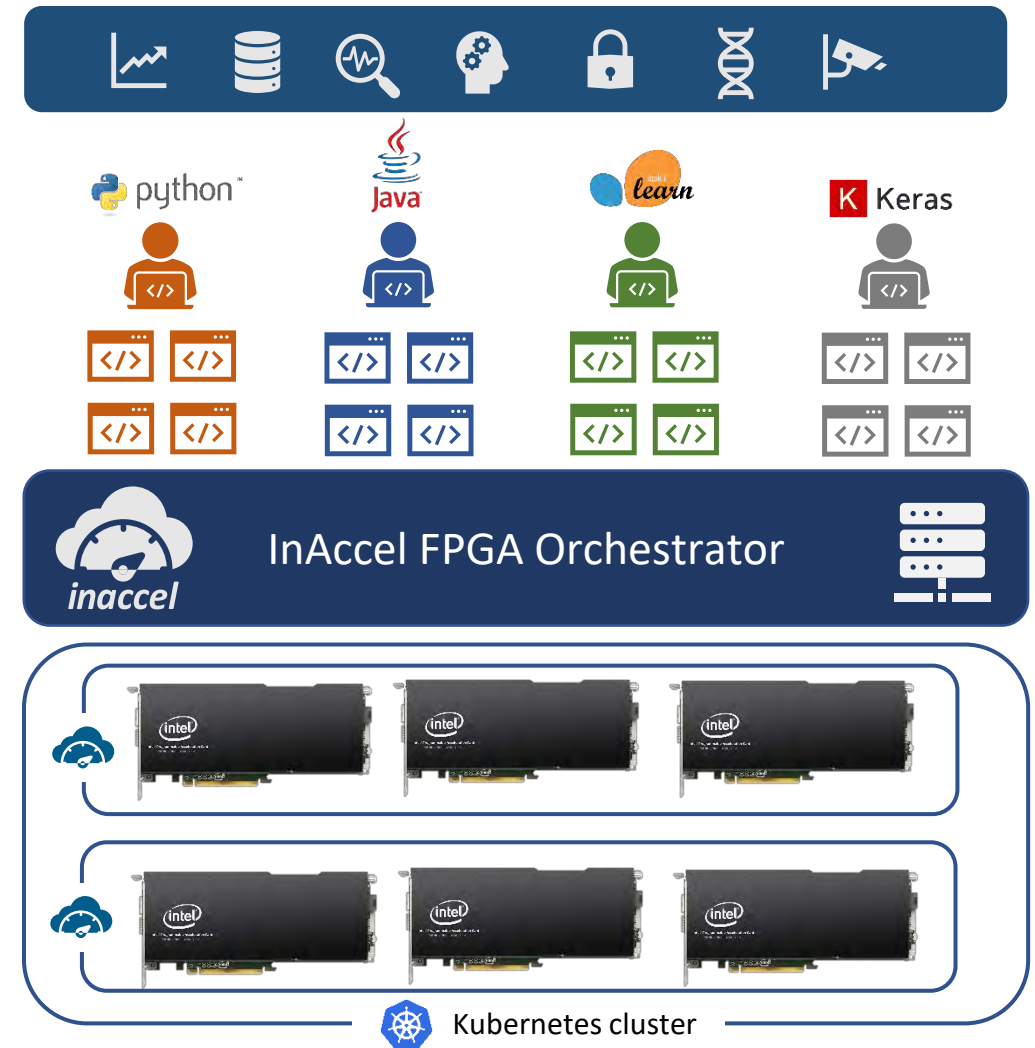
Easiest way to utilize the power of FPGAs

-  Integration to C/C++, Python, Java
-  Instant scaling to multiple FPGA
-  Multi-tenant deployment
-  Automatic configuration & deployment



Easiest way to utilize the power of FPGAs

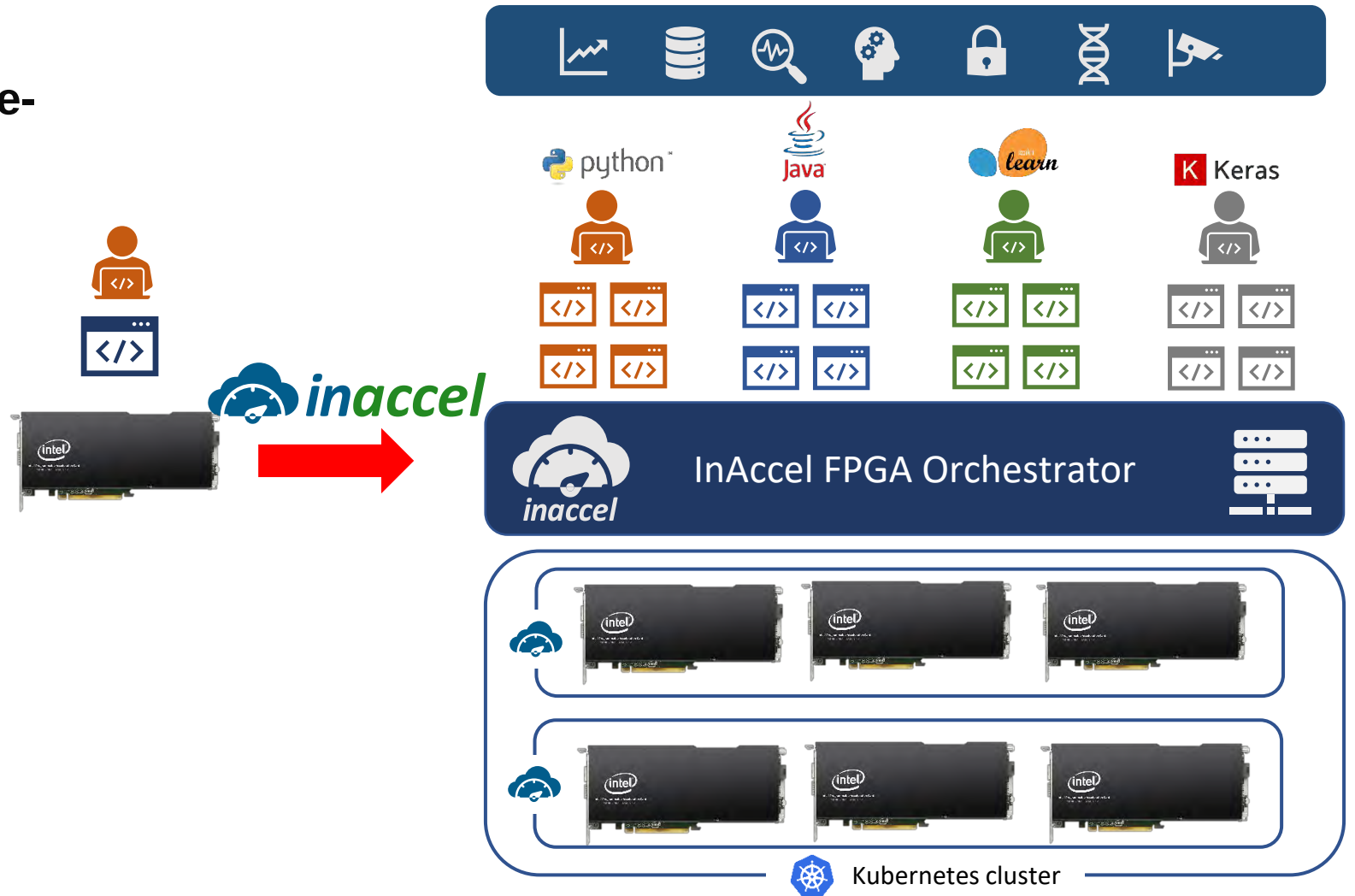
-  Integration to C/C++, Python, Java
-  Instant scaling to multiple FPGA
-  Multi-tenant deployment
-  Automatic configuration & deployment



From single instance to data centers



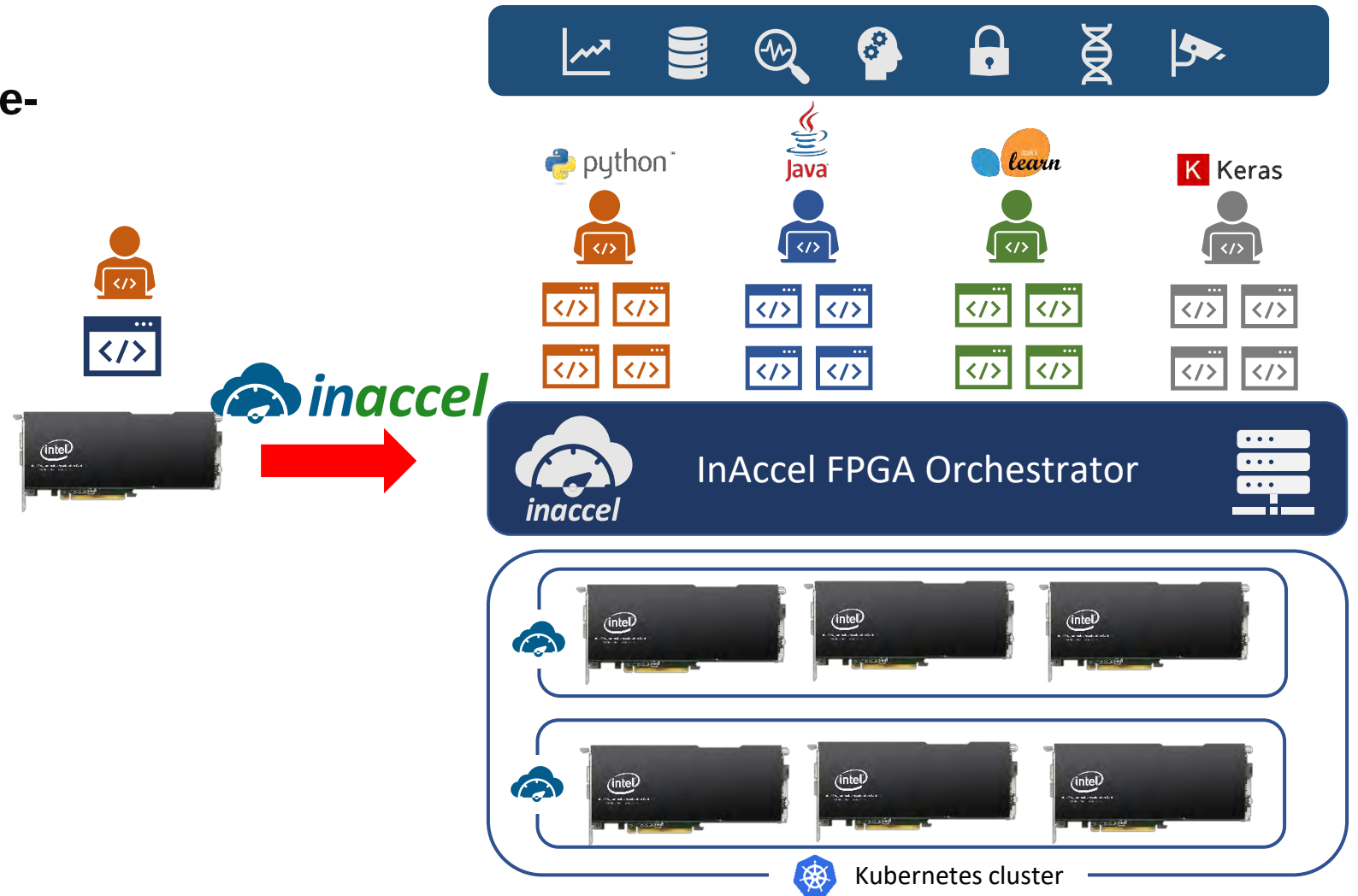
- > Easy deployment (software-alike function invoking)
- > Instant scaling
- > Seamless sharing
- > Multi-tenant deployment
- > Multiple applications
- > Isolation
- > Privacy
- > Fault-tolerant



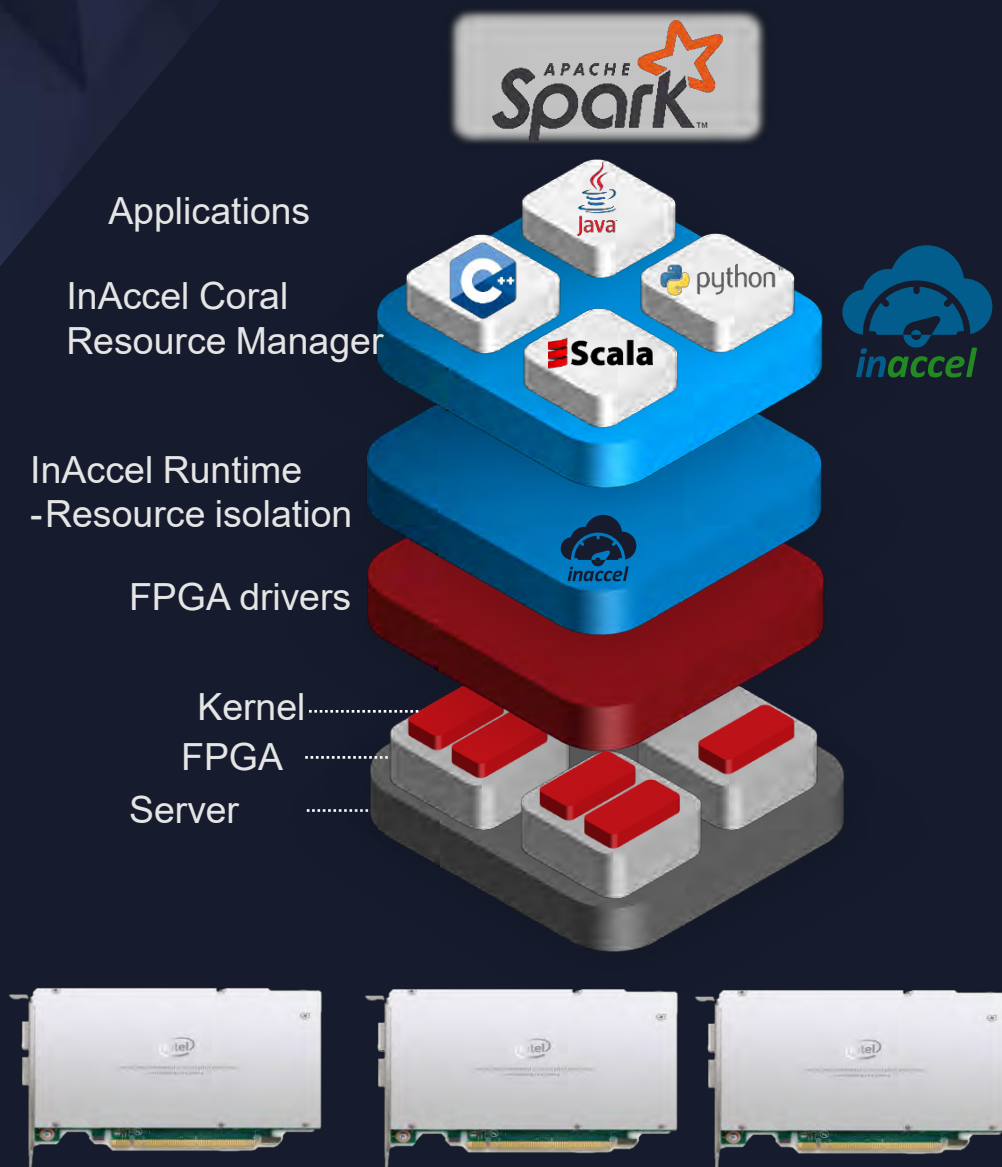
From single instance to data centers



- > Easy deployment (software-alike function invoking)
- > Instant scaling
- > Seamless sharing
- > Multi-tenant deployment
- > Multiple applications
- > Isolation
- > Privacy
- > Fault-tolerant



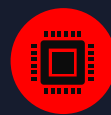
Scalable Orchestrator for FPGA clusters



Automated Deployment, Scaling and Management of FPGA clusters



Seamless invoking from C/C++, Python, Java and Scala. No need for OpenCL



Automatic configuration and management of the FPGA **bitstreams** and memory

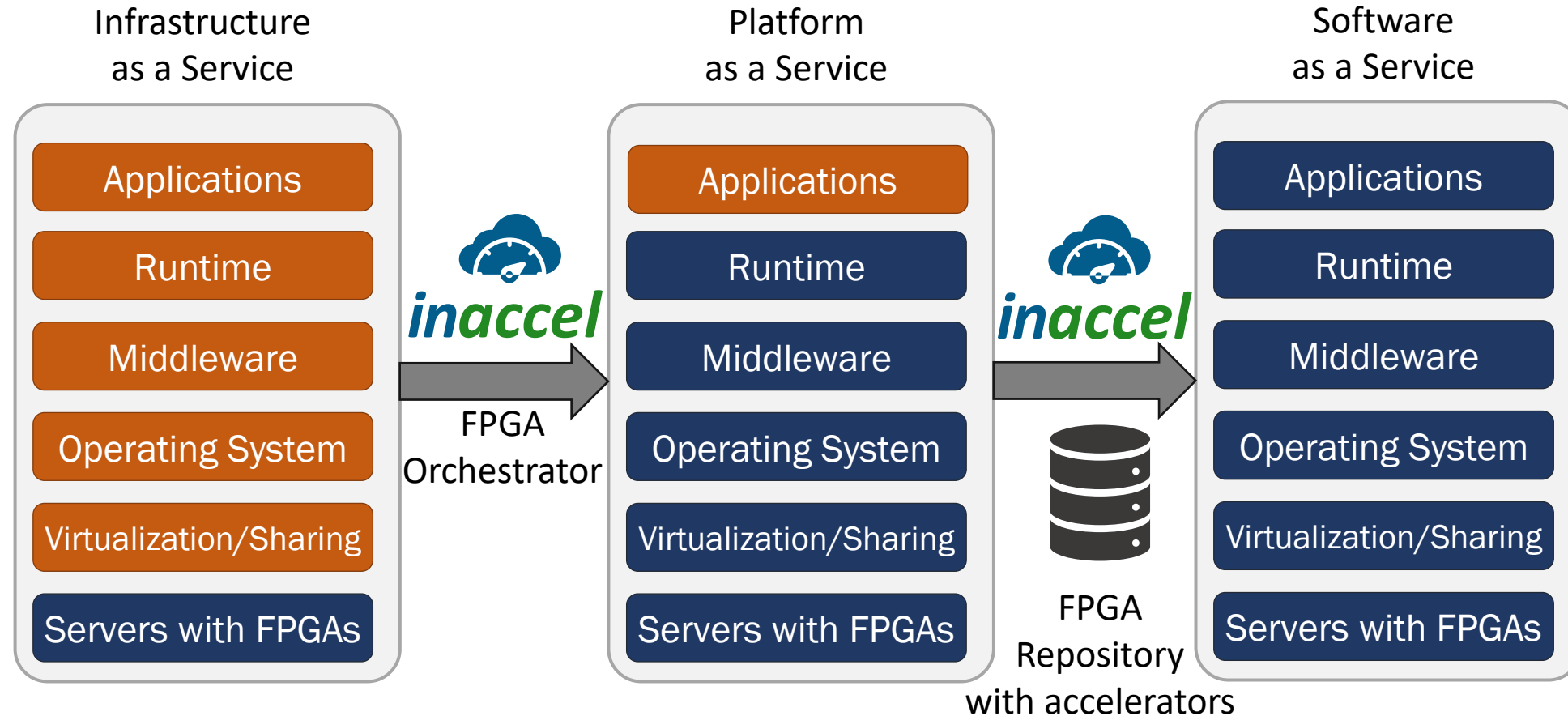


Seamless **resource management** of the FPGA cluster from multiple threads/processes/applications/users



Fully **scalable**: Scale-up (multiple FPGAs per node) and Scale-out (multiple FPGA-based servers over Spark)

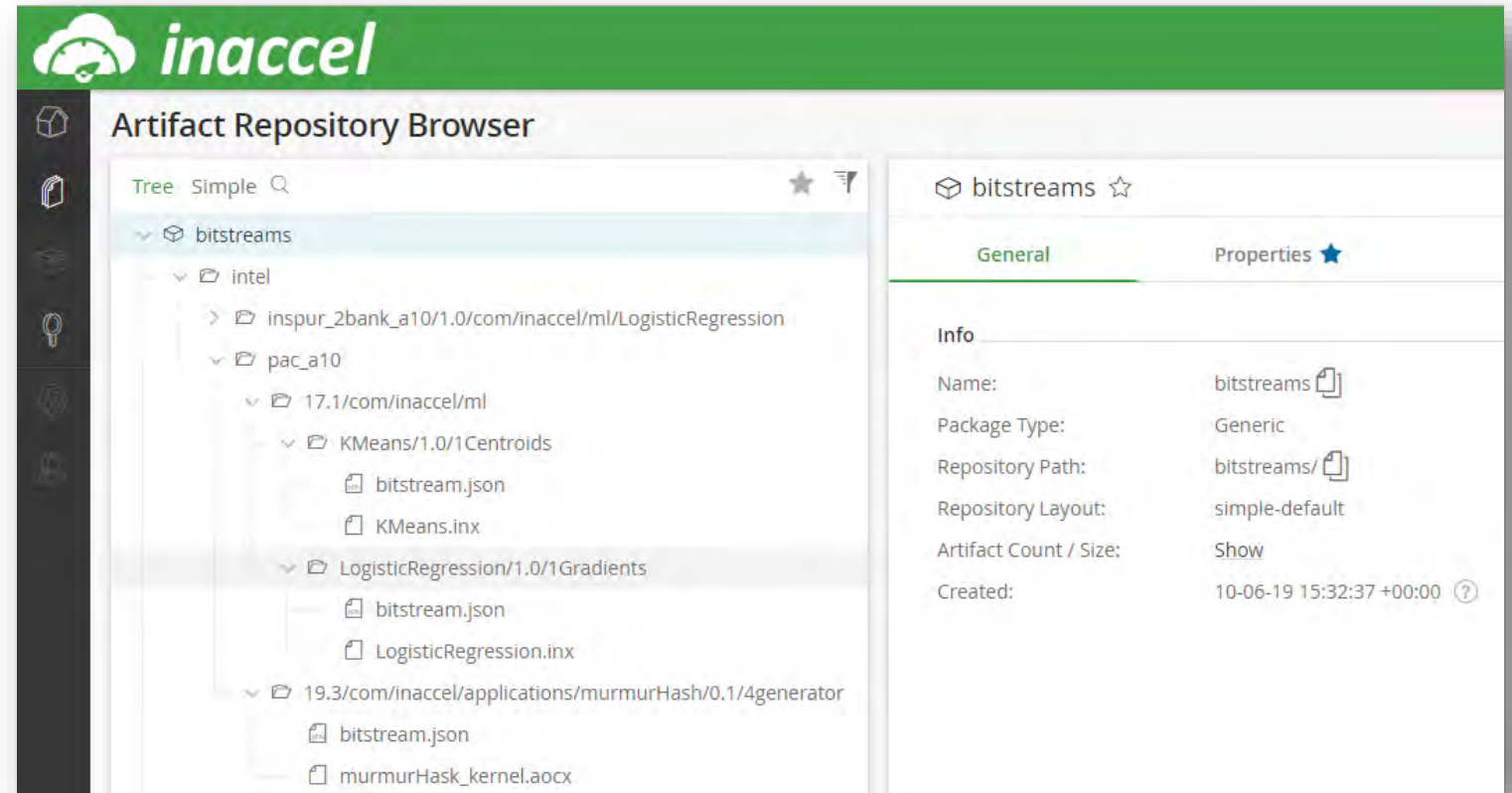
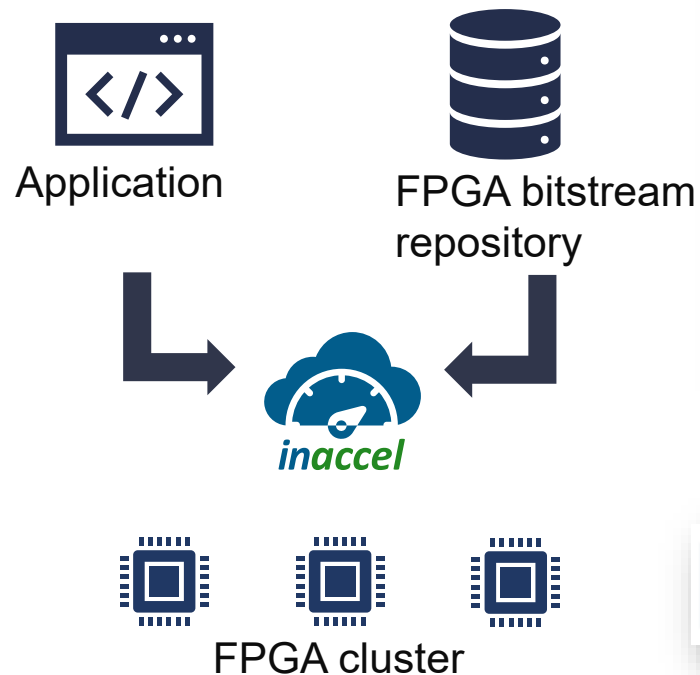
PaaS and SaaS for FPGA clusters



Bitstream repository



- > FPGA Resource Manager is integrated with Jfrog bitstream repository that is used to store FPGA bitstreams

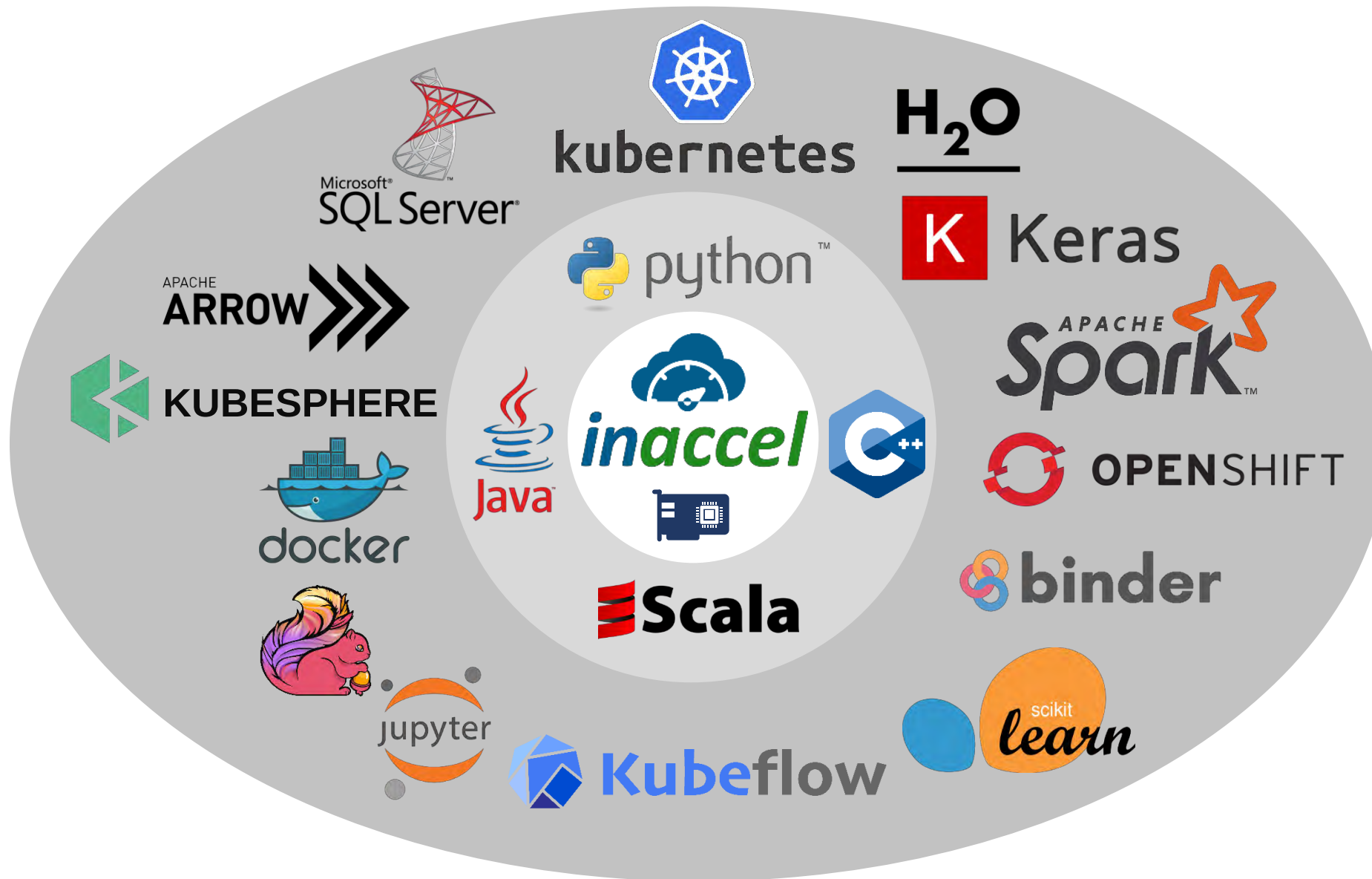


<https://store.inaccel.com>

```
inaccel bitstream install [command options]
```

Single command

Instant Integration with any framework



InAccel portal for Accelerated Machine learning

```
ImageNet Classification with ResNet50

This notebook shows how to classify images using InAccel's Keras-like framework. With this modified Keras APL we can take advantage of FPGA acceleration, while developing a Machine Learning Workflow the same way as before.

ImageNet is a very large collection of human annotated photographs designed by academics for developing computer vision algorithms. The goal of developing the dataset was to provide a resource to promote the research and development of improved methods for computer vision.

Firstly, we import the necessary libraries and load the pretrained ResNet50 model.

[1]: import numpy as np
import time

from inaccel.keras.applications.resnet50 import decode_predictions, ResNet50
from inaccel.keras.preprocessing.image import ImageDataGenerator, load_img

Last executed at 2020-06-23 11:06:12 in 11s

[2]: model = ResNet50(weights='imagenet')

Last executed at 2020-06-23 11:06:12 in 20s

Accelerated Inference

Then, we load thousands of images specifying the number of batches for every process.

[3]: data = ImageDataGenerator(dtype='float32')
images = data.flow_from_directory('imagenet/', target_size=(224, 224), class_mode=None, batch_size=64)

Last executed at 2020-06-23 11:06:18 in 1.26s
Found 22615 images belonging to 498 classes.

Now, it's time to feed the model with the images and predict their class.

We also measure the performance as the number of images processed Per Second.

[4]: begin = time.monotonic()
preds = model.predict(images, workers=10)
end = time.monotonic()

print('Duration for ', len(preds), ' images: %.1f sec' % (end - begin))
print('FPS: %.2f' % (len(preds) / (end - begin)))

Last executed at 2020-06-23 11:06:18 in 11.25s
Duration for 22615 images: 12.787 sec
FPS: 1848.498
```

- **Familiar tools**



- **Faster results**

- 10x faster
- Zero code changes

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

Instant acceleration



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

File Edit View Run Kernel Tabs Settings Help | InAccel Cloud light

/ shared / ml /

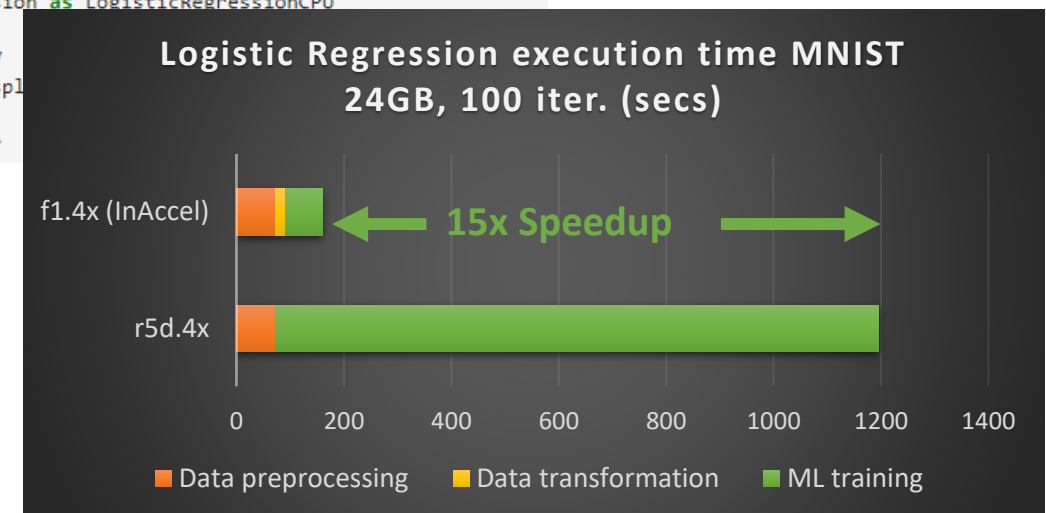
Name	Last Modified
data	12 days ago
img	19 days ago
LogisticRegression.ipynb	a month ago
MLlibPipelines.ipynb	12 days ago
NaiveBayes.ipynb	a month ago
XGBoost.ipynb	19 days ago

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



User-friendly platform for Genomics



<http://genomics.inaccel.com>

The screenshot shows the inaccelstudio web interface. At the top, there's a header with the inaccelstudio logo and a user profile icon. Below the header, there's a text box explaining the Smith-Waterman algorithm. Then, there's a section for uploading files, with buttons for 'Upload (0)' and 'Run'. Below this, there are dropdown menus for 'Target' and 'Query', both set to 'shared/KF793826.1.fasta'. The target and query names are both 'Bottlenose dolphin coronavirus HKU22 isolate CF090331, complete genome'. There are sliders for 'Gap penalty' with 'Opening' set to 5 and 'Extension' set to 2. Below the sliders is a checkbox for 'Enable InAccel'. At the bottom, there's a table with columns for 'Target Name', 'Start', 'End', 'Query Name', 'Start', 'End', 'Score', 'Score²', 'Target End²'. The table is currently empty, and a red rectangle highlights the empty area. Below the table, there's a section titled 'Algorithm' with a description of the Smith-Waterman algorithm and a list of steps.

The Smith-Waterman algorithm performs *local sequence alignment*; that is, for determining similar regions between two strings (**Target, Query**) of *nucleic acid* sequences or *protein* sequences. Instead of looking at the entire sequence, the Smith-Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information (including [Nucleotide](#), [Protein](#) Databases).

Data are available in FASTA file format (e.g. [Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome](#)).

Upload (0) Run

Target: shared/KF793826.1.fasta Bottlenose dolphin coronavirus HKU22 isolate CF090331, complete genome

Query: shared/KF793826.1.fasta Bottlenose dolphin coronavirus HKU22 isolate CF090331, complete genome

Gap penalty:

Opening 5 Extension 2

Enable InAccel

Target Name	Start	End	Query Name	Start	End	Score	Score²	Target End²
-------------	-------	-----	------------	-------	-----	-------	--------	-------------

Algorithm

Let $A = a_1a_2 \dots a_n$ and $B = b_1b_2 \dots b_m$ be the sequences to be aligned, where n and m are the lengths of A and B respectively.

- Determine the substitution matrix and the gap penalty scheme.

Seamless integration with any framework

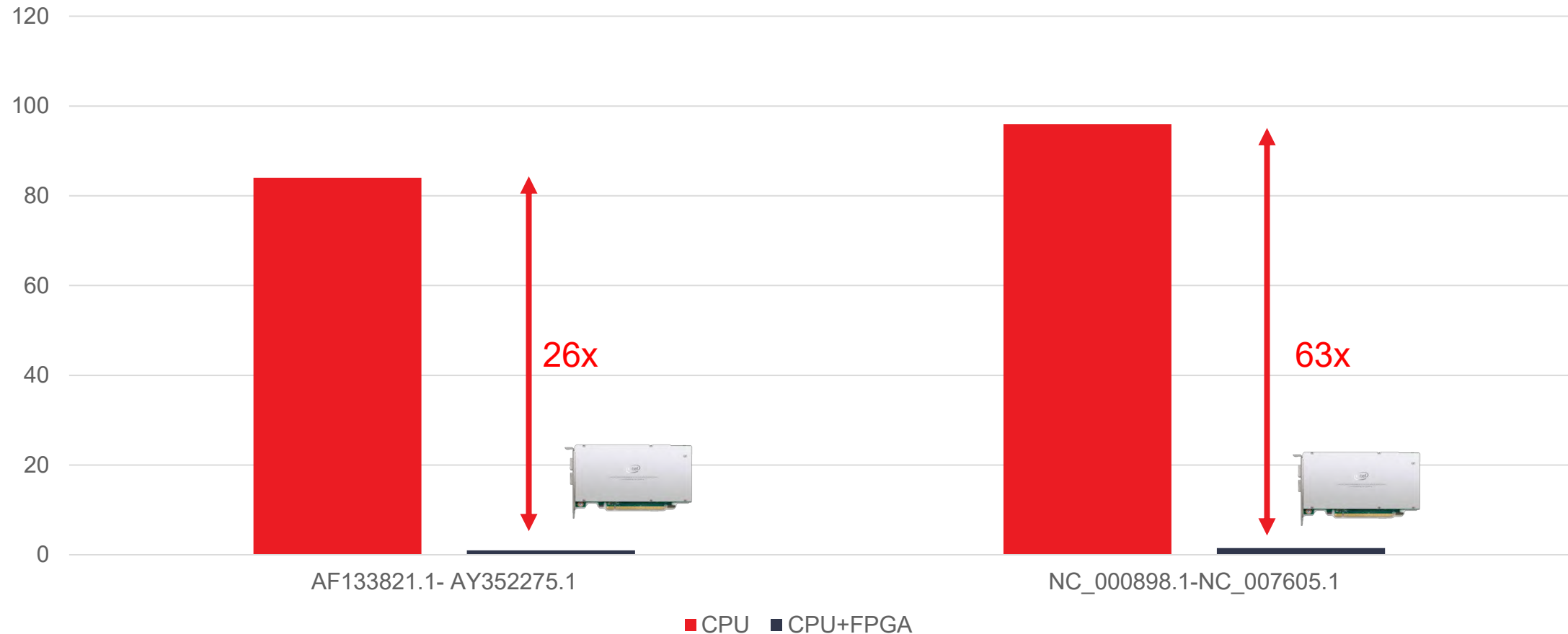


Python/terminal based execution

A screenshot of a Jupyter Notebook interface. The top bar shows "Welcome to InAccele Cloud.ip" and "Smith-Waterman.ipynb". The notebook is in "Python 3" mode. The first cell contains the text "Finally, we check that the results are the same." followed by a code cell with `np.array_equal(c, a + b)`. Below this is a section titled "Genomics" with a description of the Smith-Waterman algorithm and a short example sequence: `AAANTACCTTNTTAGAAGCTGCCCTTTGCTGT`. The following text explains that the `ksw` app can be run with the `-k` flag. The final two cells show terminal output for the `ksw` command: `!ksw -k klib/AF133821.1.fasta klib/NC_000898.1.fasta` and `!ksw klib/AF133821.1.fasta klib/NC_000898.1.fasta`.

Seamless Speedup

Execution time of SW sequence alignment



Reference CPU server: Dell PowerEdge T640 server, Intel® Xeon® Silver processor 4208 2.1 GHz, 8C/16T, 9.6GT/s, 11M cache, RAM: 64 GB DDR, 480 GB SSD SATA
FPGA server: Dell PowerEdge T640 server, Intel Xeon Silver processor 4208 2.1 GHz, 8C/16T, 9.6GT/s, 11M cache, RAM: 64 GB DDR, 480 GB SSD SATA with 1x PAC Intel® Arria® 10 FPGA

Applications



Machine Learning



Deep Learning



Genomics



Quantitative Finance



Video processing

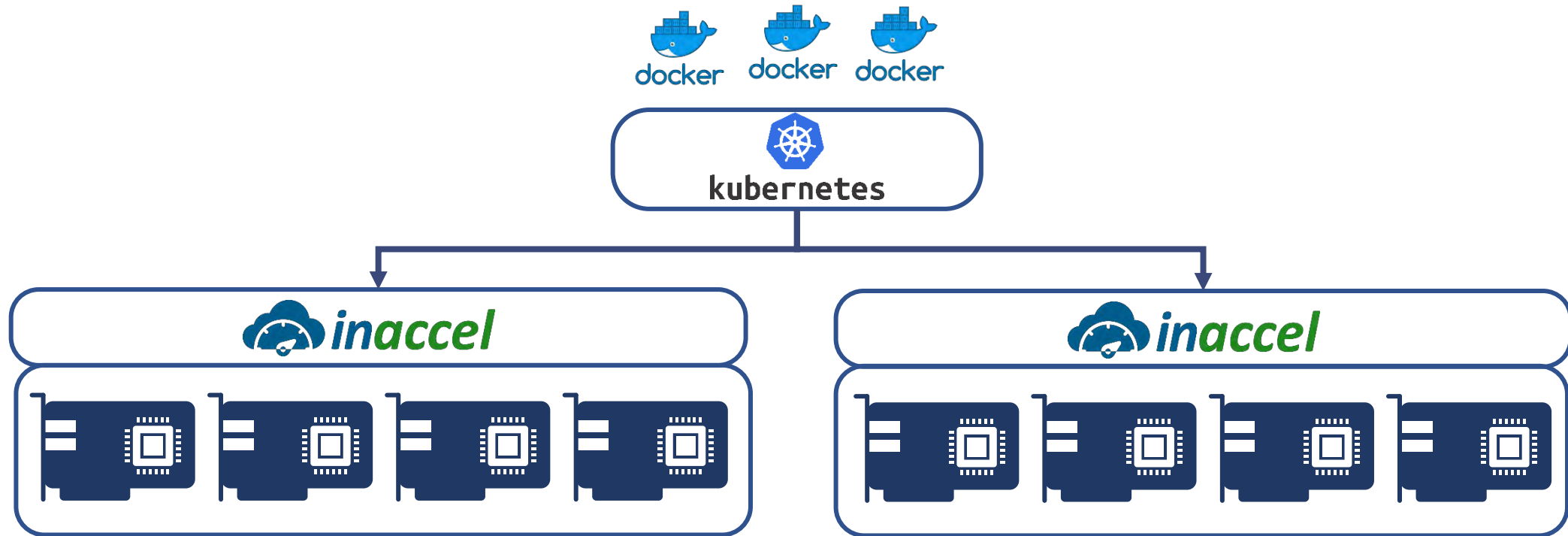


Image processing

InAccel Coral manager - Kubernetes



- > Integrated solution that allows
 - >> Scale Up (1, 2, or 8 FPGAs per server)
 - >> Scale Out to multiple servers



Graphical monitoring tool



Use cases



Machine Learning

INACCEL'S ACCELERATED ML SUITE BOOSTS SPARK ML PERFORMANCE BY AS MUCH AS 7X ON FPGA-BASED ALIBABA CLOUD F1 INSTANCES

Written by Steven Leibson | August 2, 2019



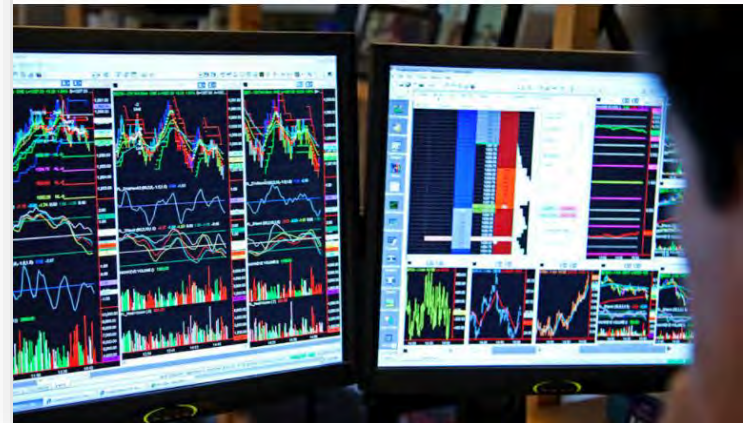
<https://blogs.intel.com/psg/inaccels-accelerated-ml-suite-boosts-spark-ml-performance-by-as-much-as-7x-on-fpga-based-alibaba-cloud-f1-instances/>

Quantitative Finance



FLUMAION ACCELERATES QUANTITATIVE FINANCIAL CALCULATIONS FOR HIGH-FREQUENCY TRADING WITH INTEL® PAC WITH INTEL® ARRIA® 10 GX FPGA AND INACCEL ORCHESTRATOR

Written by Steven Leibson | June 11, 2020



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



Genomics



Solution Brief

FPGA
Genomic Analytics

Acceleration of Sequence Alignment and Variant Calling for Genomic Analytics Using Intel® FPGAs

Introduction

Genomic Analytics aligns a selected genome to a reference genome to detect genetic variants in that selected genome as compared to the reference genome. This technique is fundamental to the diagnosis and cure of rare, inherited diseases as well as for medical breakthroughs and personalized care. The medical industry is progressing towards personalized medicine, which will require the storage of many, many human genomes. There are 3 billion nucleotide base pairs in a human genome, which translates into immense data volumes needed to store everyone's genome.

As the Coronavirus has raced around the world, thousands of genome sequences of the virus have been shared on GISAID¹, an online global platform for genomic data. One Coronavirus genome sequence contains 26K to 32K bases in the RNA strand located inside the coronavirus. These shared sequence variants offer clues about how the virus, named SARS-CoV-2, is spreading and evolving. But because these shared sequences represent a tiny fraction of cases and show few tell-tale differences, they are easy to overinterpret.²

Virologist Eva Broberg of the Centre for Disease Prevention and Control³ states that "there are more plausible scenarios for how the disease reached northern Italy than an undetected spread from Bavaria."⁴ This statement underscores the importance of fast sequence alignment of the Coronavirus mutations.

"The very first SARS-CoV-2 sequence, in early January, answered the most basic question about the disease: What pathogen is causing it? The genomes that followed were almost identical, suggesting the virus, which originated in an animal, had crossed into the human population just once. If it had jumped the species barrier multiple times, the first human cases would show more variety. Some diversity is now emerging. Over the length of its 30,000-base-pair genome, SARS-CoV-2 accumulates an average of about one to two mutations per month. Using these little changes, researchers draw up phylogenetic trees, much like family trees, make connections between cases, and gauge whether there might be undetected spread of the virus."⁵ Fast analysis and tracking of the mutations is critical for better protection of the population as the virus evolves and migrates between the countries and geographies.

Scientists will be scouring the genomic diversity of these viral genome sequences for signs that the virus is getting more dangerous. Caution is warranted. An analysis of 103 genomes published by Lu Jian of Peking University and colleagues on 3 March 2020 in the National Science Review argued that they fell into one of two distinct types, named S and L, and are distinguished by two mutations. Because 70% of sequenced SARS-CoV-2 genomes belong to a newer type L genome, the authors concluded that the type L genome has evolved to become more aggressive and spreads faster.⁶

Genomic scientists and researchers within different groups around the world have been trying to uncover genetic determinants of susceptibility, severity, and outcomes of COVID-19 from the genomes of COVID-19 patients. COVID-19 is the pandemic disease caused by the SARS-CoV-2 coronavirus.

Using genomic analytics on the sequences of novel viruses to track mutations is a computationally intensive algorithm and requires powerful processing platforms for the efficient processing of huge amounts of data. Fast genome sequencing

Authors

Natalia Poliakova
Technical Solution Sales Specialist
Intel Corporation

Ioannis Stamelos
Chief Solution Architect
inAccel

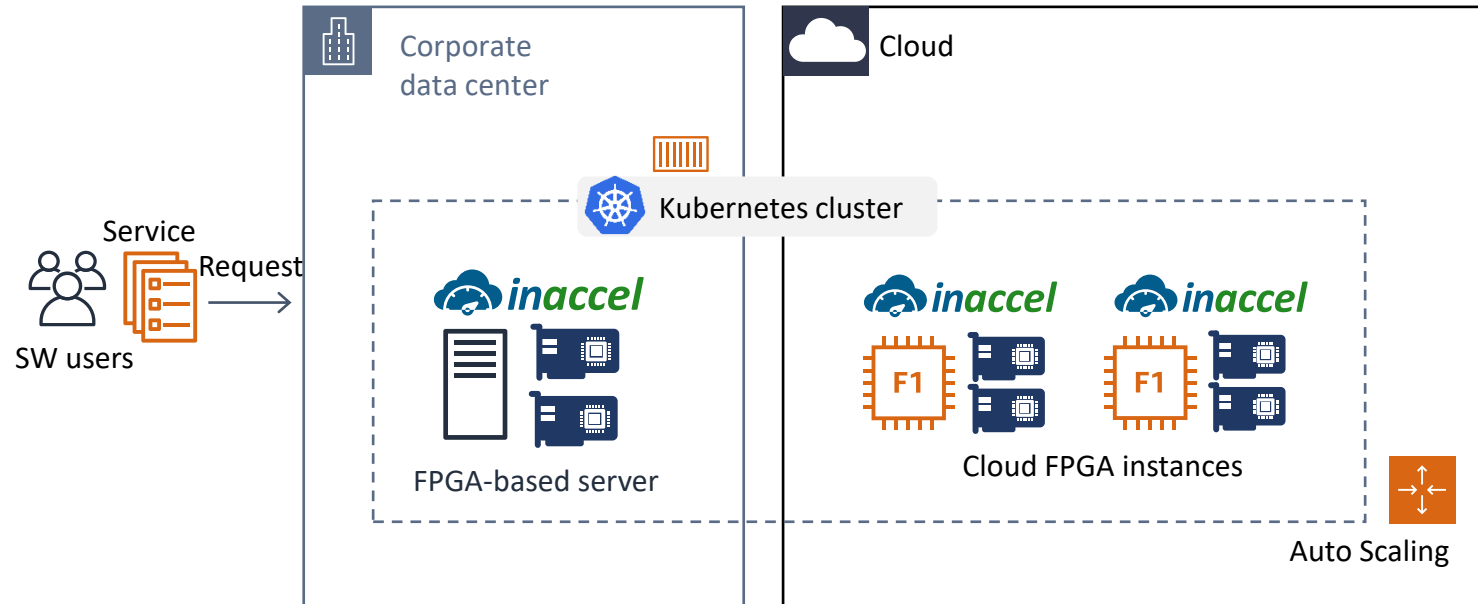
Elias Koromilas
Chief Technology Officer
inAccel

Aspasia Stavrianou
Senior FPGA Engineer
inAccel

Chris Kachris
CEO
inAccel

Calvin Hung
CEO
WASAI Technology

Auto-scaling on cloud and on-prem

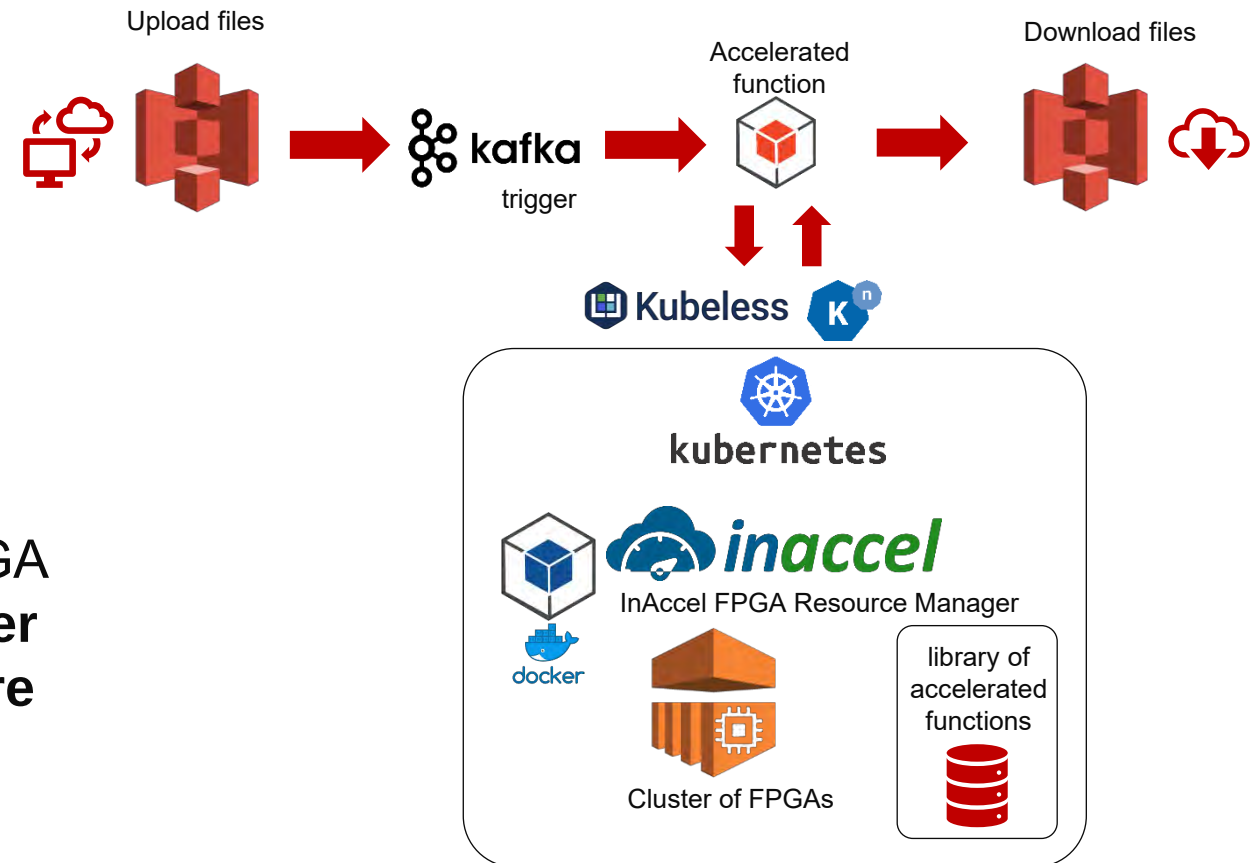


InAccel allows to auto-scale your cluster based on the workload requirement

Serverless deployment



- > Integrated framework for serverless deployment
- > Compatible with Kubernetes
- > Compatible with Kubeless, Knative
- > Users only have to **upload the images** on the storage bucket and then InAccel's FPGA Manager **automatically deploy the cluster of FPGAs**, process the data and then **store back the results** on the storage bucket.
- > Users do not have to know anything about the FPGA execution.

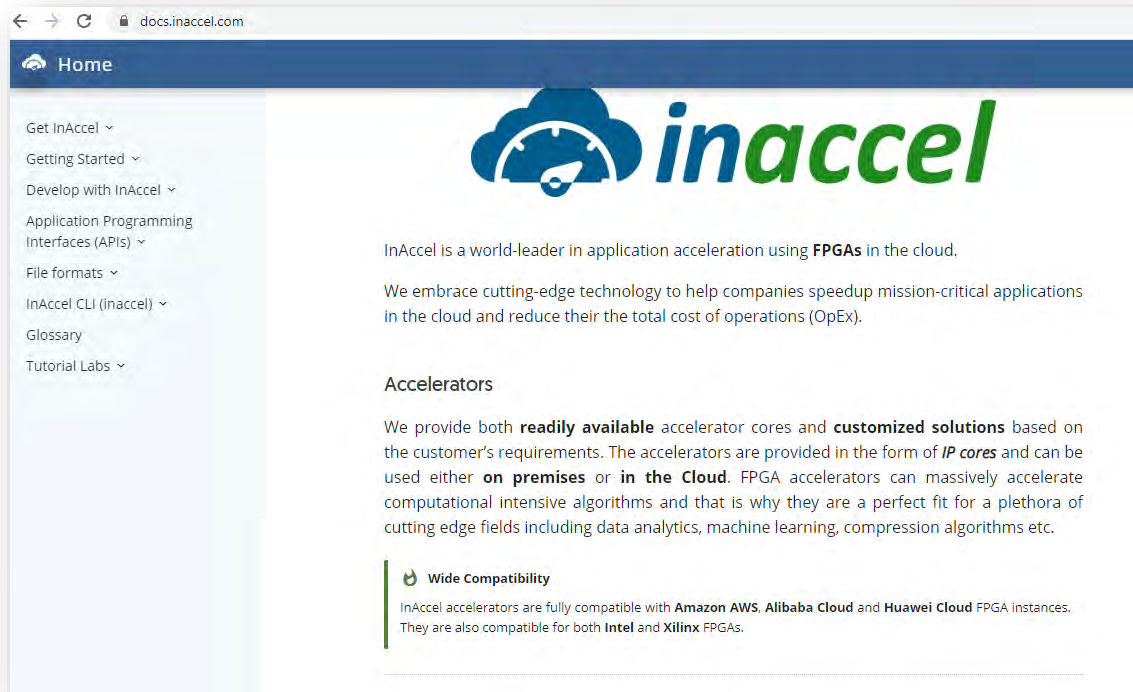


<https://medium.com/@inaccel/fpgas-goes-serverless-on-kubernetes-55c1d39c5e30>

Test it on your prem or on your browser

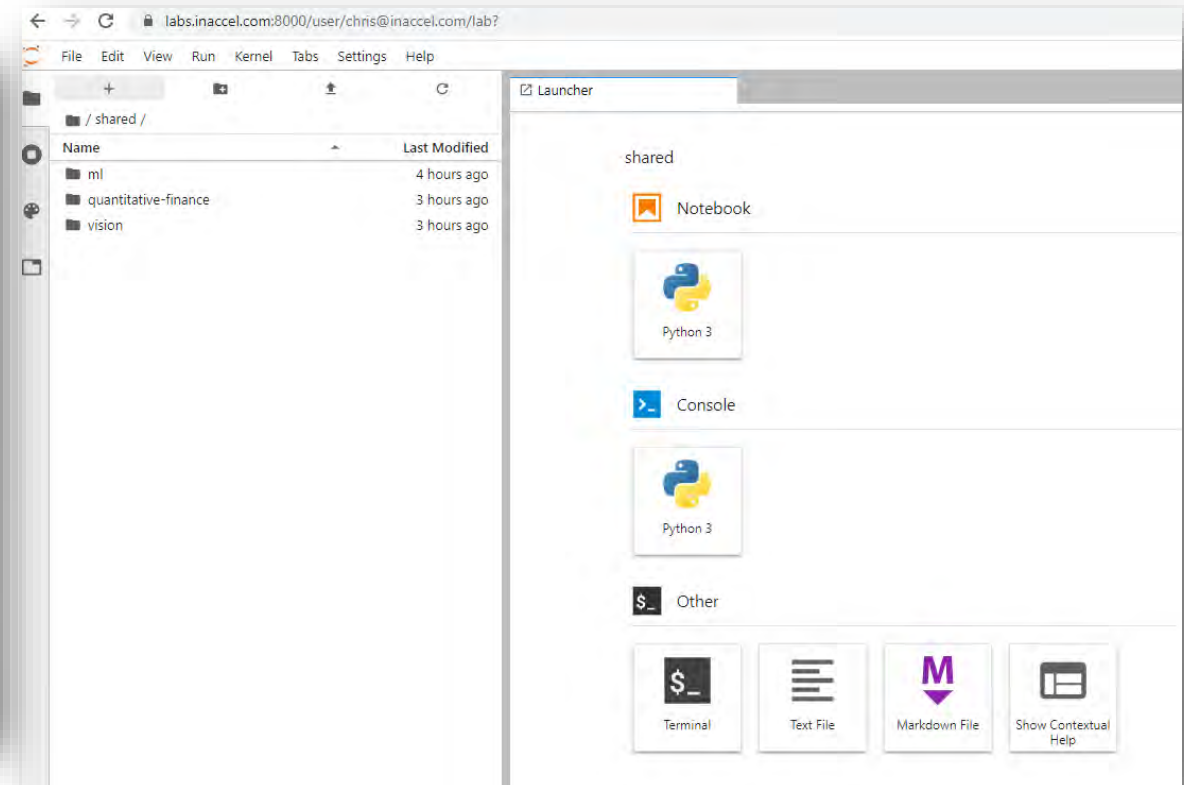


On-prem



<https://docs.inaccel.com/>

Online - Browser



<https://labs.inaccel.com:8000/>

InAccel, Inc. Corporate overview

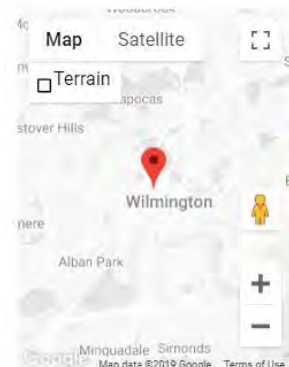


> Founded in January 2018

> Membership:



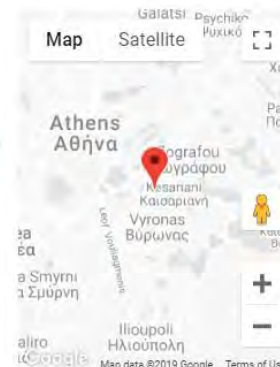
Microsoft
for Startups



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



MORPHEMIC is newly funded EU Horizon 2020 project.
Start date: 01/01/2020, end date: 31.12.2022. Grant
agreement ID: 87164312



Application Acceleration, seamlessly

www.inaccel.com

info@inaccel.com

USA:

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

Europe (Design Center):

Formionos 47
Kesariani 116 33
Athens, Greece