FPGA Acceleration of Apache Spark ML on the Cloud, Instantly

Dr. Chris Kachris
CEO, co-founder
www.inaccel.com

…or

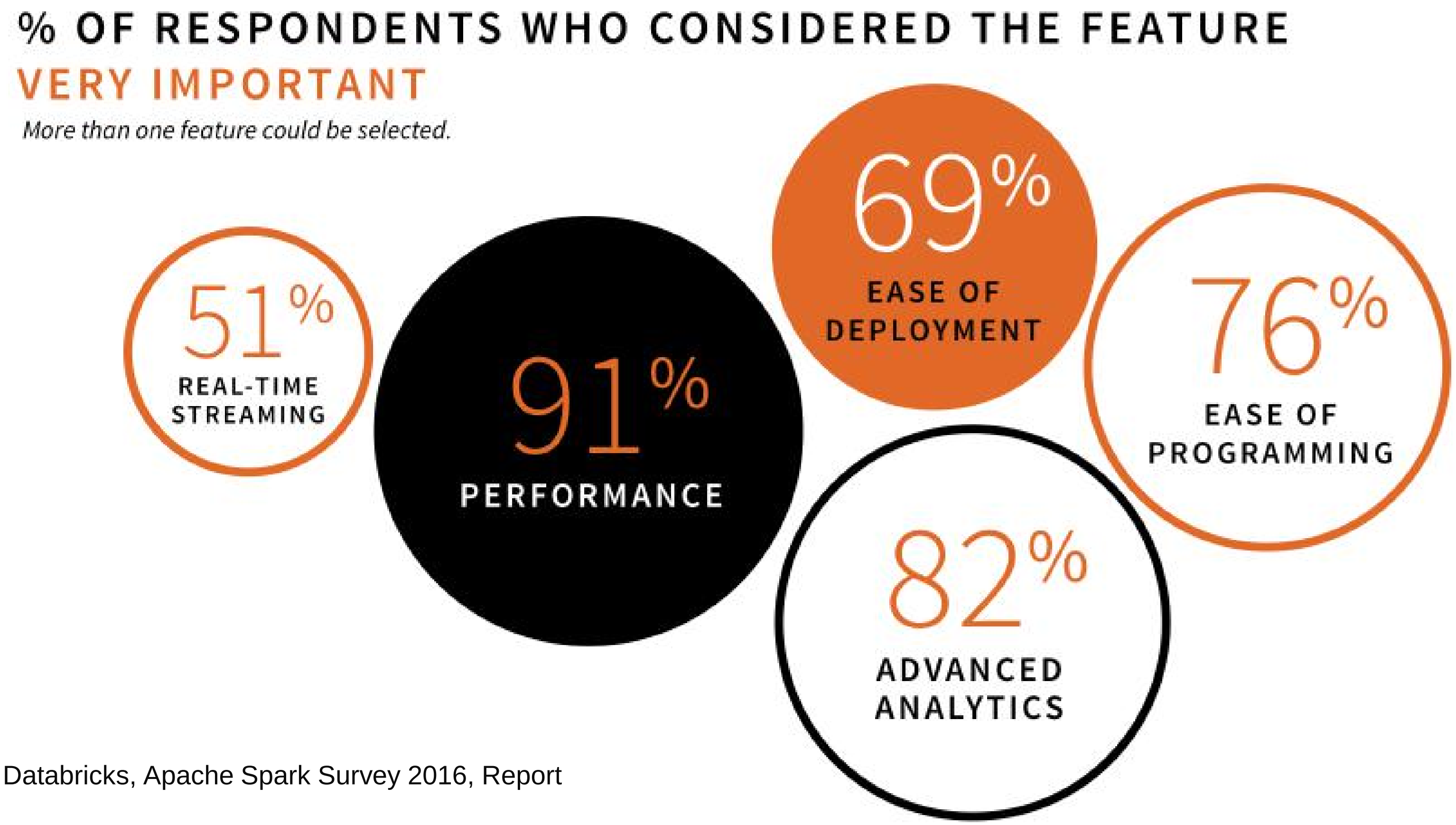
How to speedup your Spark ML applications

with the same cost

with the same code

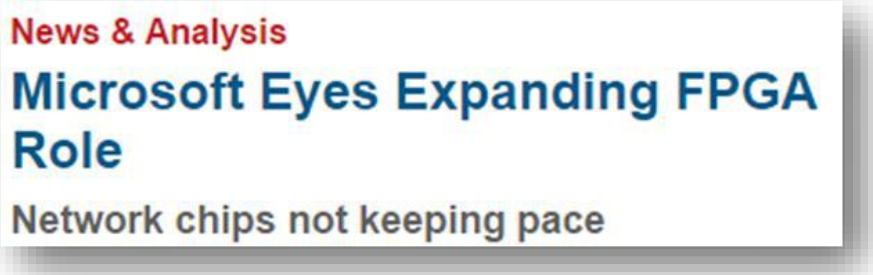
Why acceleration

- 91% of Spark users for Big Data analytics care about Performance



Source: Databricks, Apache Spark Survey 2016, Report

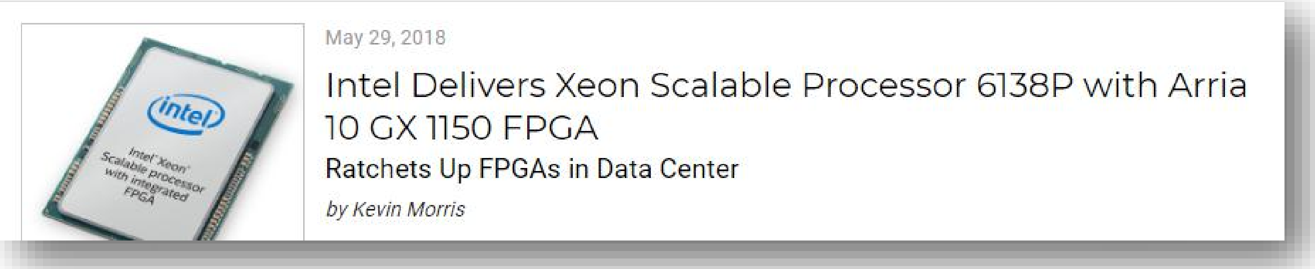
FPGAs in the news



News & Analysis

Microsoft Eyes Expanding FPGA Role

Network chips not keeping pace

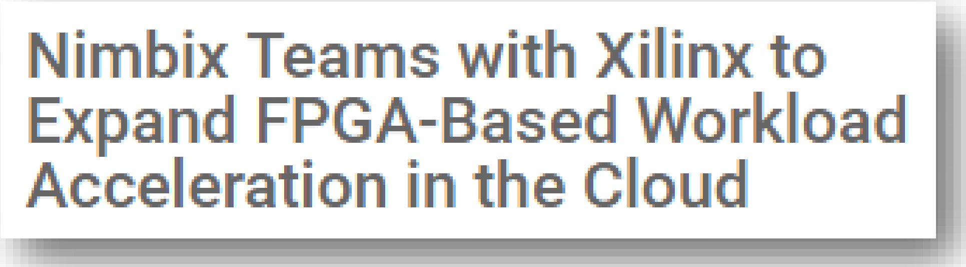


May 29, 2018

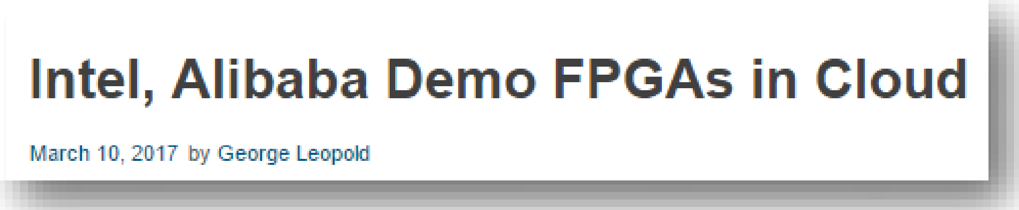
Intel Delivers Xeon Scalable Processor 6138P with Arria 10 GX 1150 FPGA

Ratchets Up FPGAs in Data Center

by Kevin Morris

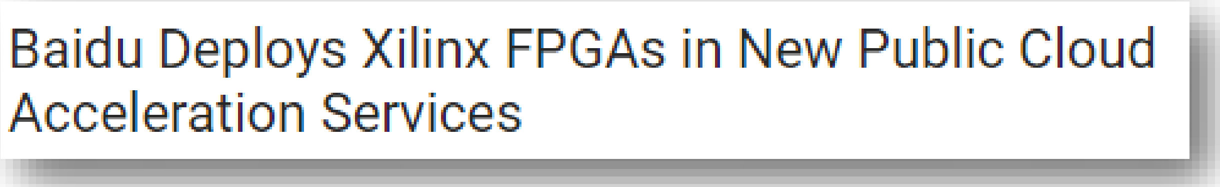


Nimbix Teams with Xilinx to Expand FPGA-Based Workload Acceleration in the Cloud

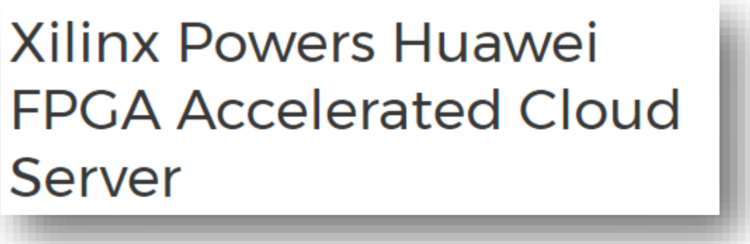


Intel, Alibaba Demo FPGAs in Cloud

March 10, 2017 by George Leopold



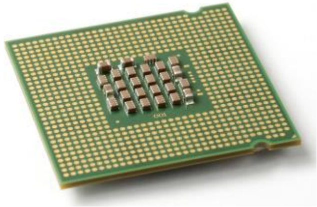
Baidu Deploys Xilinx FPGAs in New Public Cloud Acceleration Services



Xilinx Powers Huawei FPGA Accelerated Cloud Server

Available Platforms

Flexibility



CPUs

+ Flexible & Cheap

- low performance

GPUs

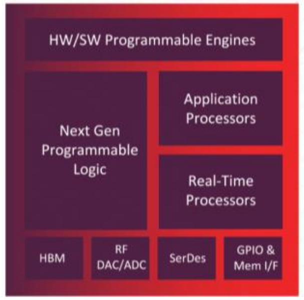
+ Flexible

- Expensive & hard to program

Specialized chips/FPGA

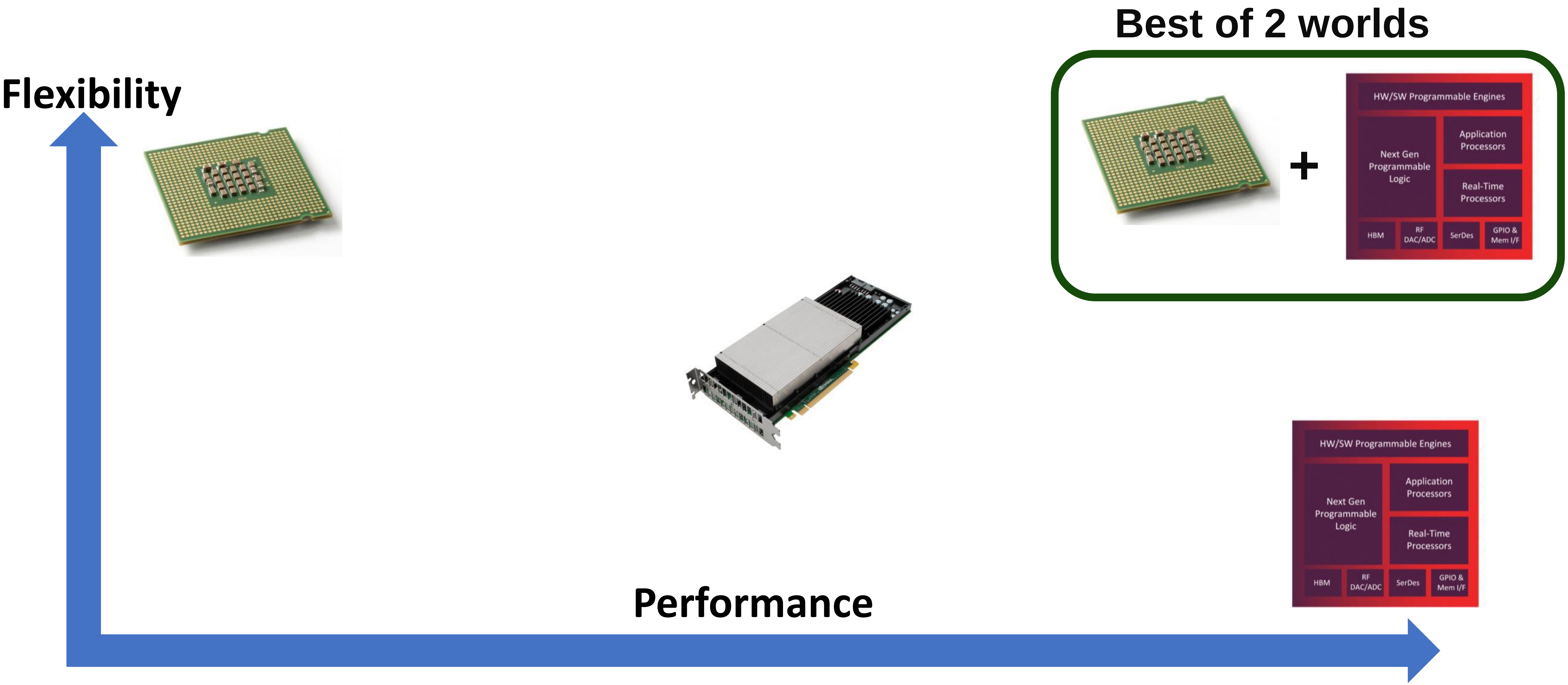
+ High Performance

- low flexibility



Performance

Available Platforms



Best of 2 worlds

Flexibility

+

HW/SW Programmable Engines

Next Gen Programmable Logic

Application Processors

Real-Time Processors

HBM

RF DAC/ADC

SerDes

GPIO & Mem I/F

HW/SW Programmable Engines

Next Gen Programmable Logic

Application Processors

Real-Time Processors

HBM

RF DAC/ADC

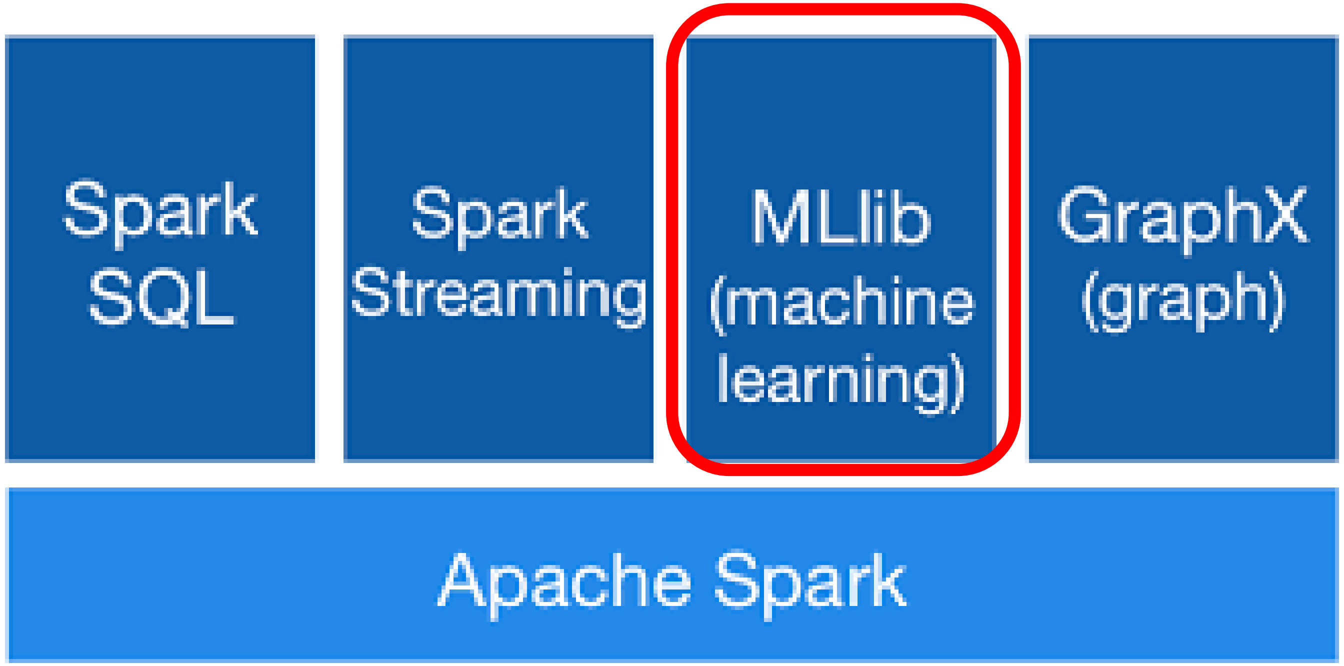
SerDes

GPIO & Mem I/F

Performance

Apache Spark

- Spark is the most widely used framework for Data Analytics
- Develop hardware components as IP cores for widely used applications
  - Spark
    - Logistic regression
    - Recommendation
    - K-means
    - Linear regression
    - PageRank
    - Graph computing



Market size

- The **data center accelerator market** is expected to reach **USD 21.19 billion by 2023** from USD 2.84 billion by 2018, at a CAGR of **49.47%** from 2018 to 2023.

- The market for FPGA is expected to grow at **the highest CAGR during the forecast period** owing to the increasing adoption of FPGAs for the acceleration of enterprise workloads.

[Source: Data Center Accelerator Market by Processor Type (CPU, GPU, FPGA, ASIC)- Global Forecast to 2023, Research and Markets] inaccel

helps companies speedup their applications

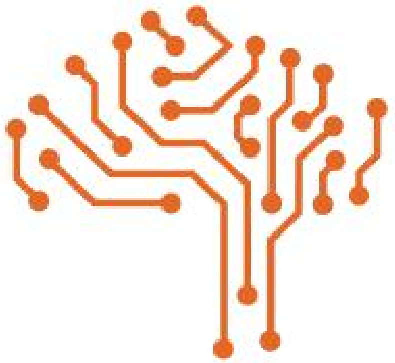
by providing ready-to-use accelerators-as-a-service in the cloud

- 3x-10x Speedup
- 2x Lower Cost
- Zero code changes

Acceleration for machine learning

inaccel offers **Accelerators-as-a-Service** for Apache Spark in the cloud (e.g. Amazon AWS f1) using FPGAs

APACHE Spark™



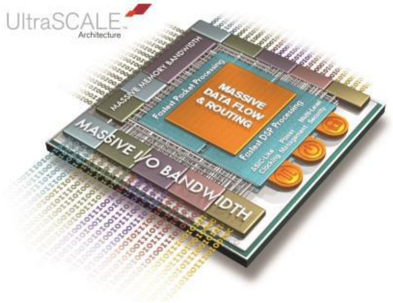
ADVANCED ANALYTICS USERS (MLLIB)
IN PRODUCTION

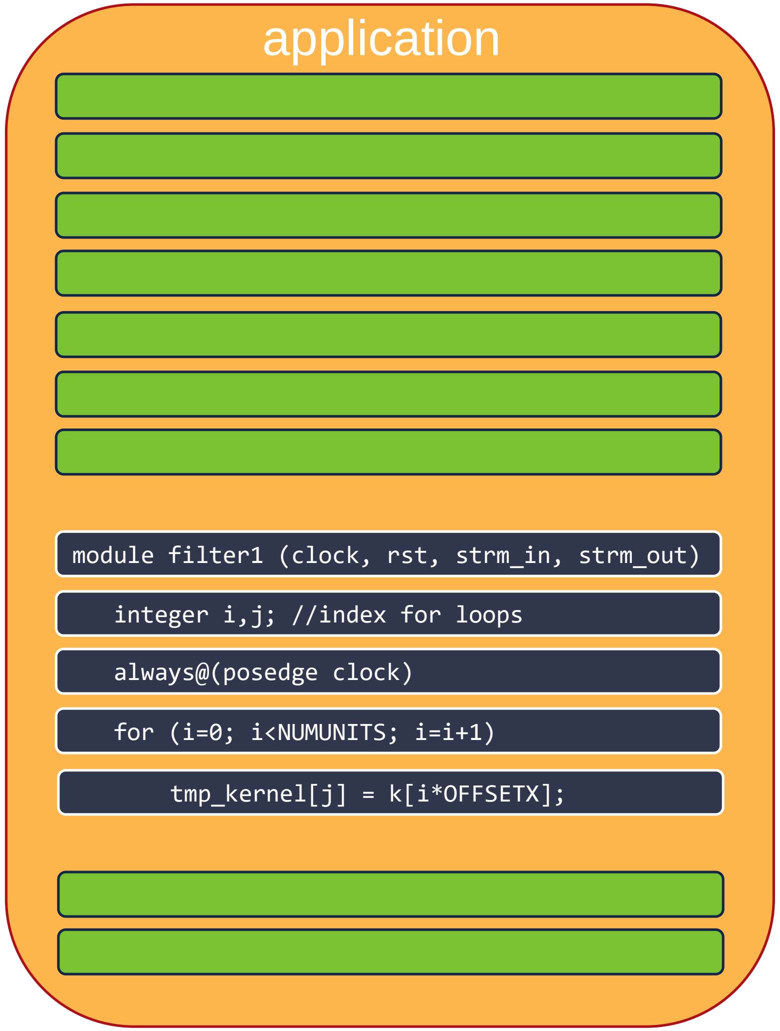
+38%

2015	2016
13% OF RESPONDENTS	18% OF RESPONDENTS

Hardware acceleration

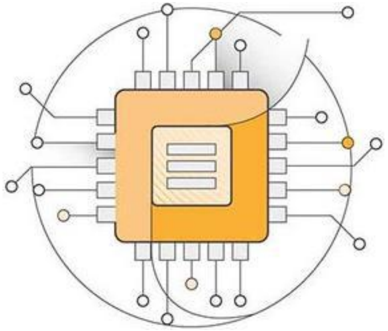
FPGA handles compute-intensive, deeply pipelined, hardware-accelerated operations





```
module filter1 (clock, rst, strm_in, strm_out)
   integer i,j; //index for loops
   always@(posedge clock)
   for (i=0; i<NUMUNITS; i=i+1)
       tmp_kernel[j] = k[i*OFFSETX];
```

200 sec

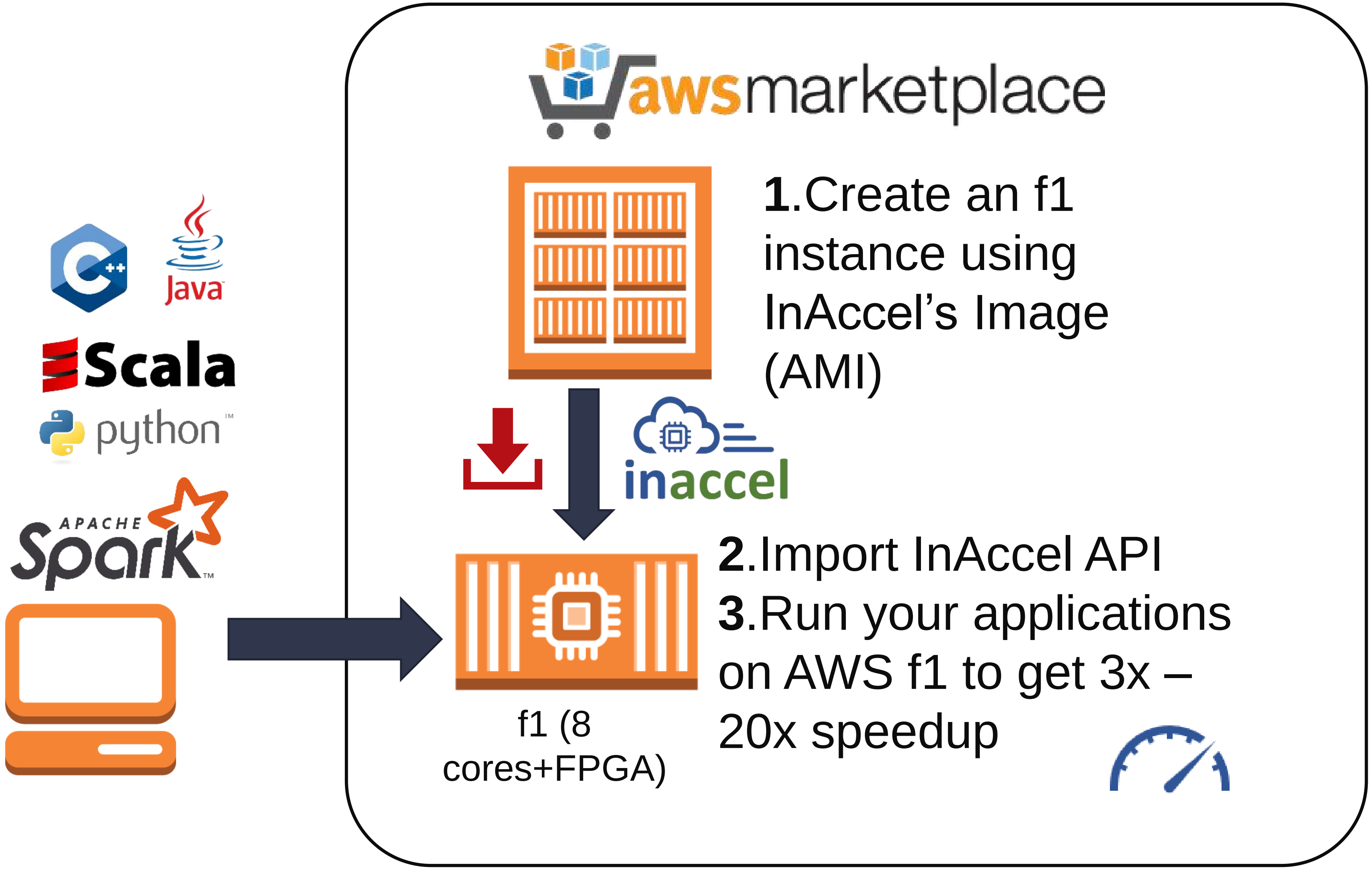


CPU handles the rest

80 sec **InAccel** 800 sec

Source: amazon, Inc.

Accelerators for Spark ML in Amazon AWS in 3 steps



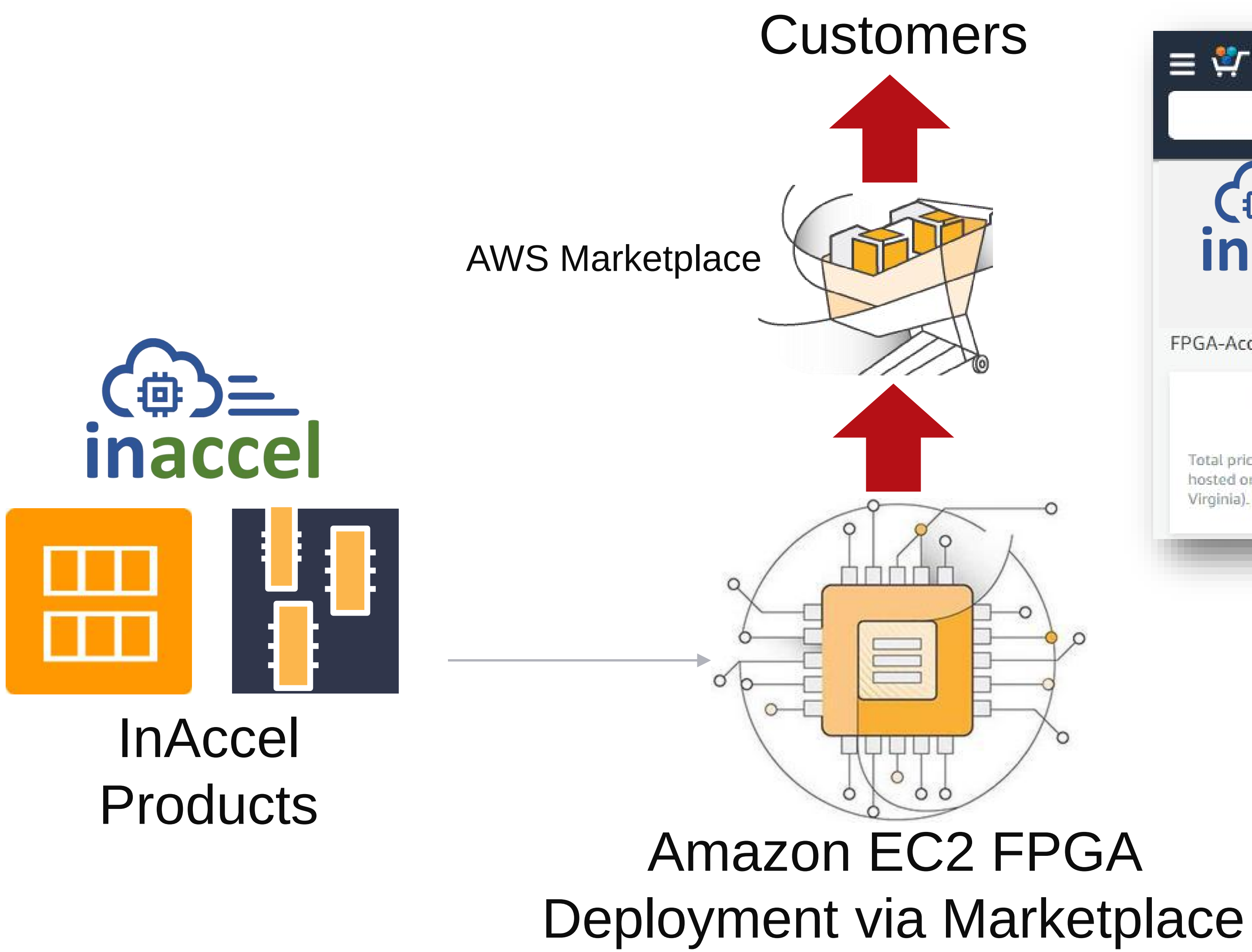
1.Create an f1 instance using InAccel's Image (AMI)

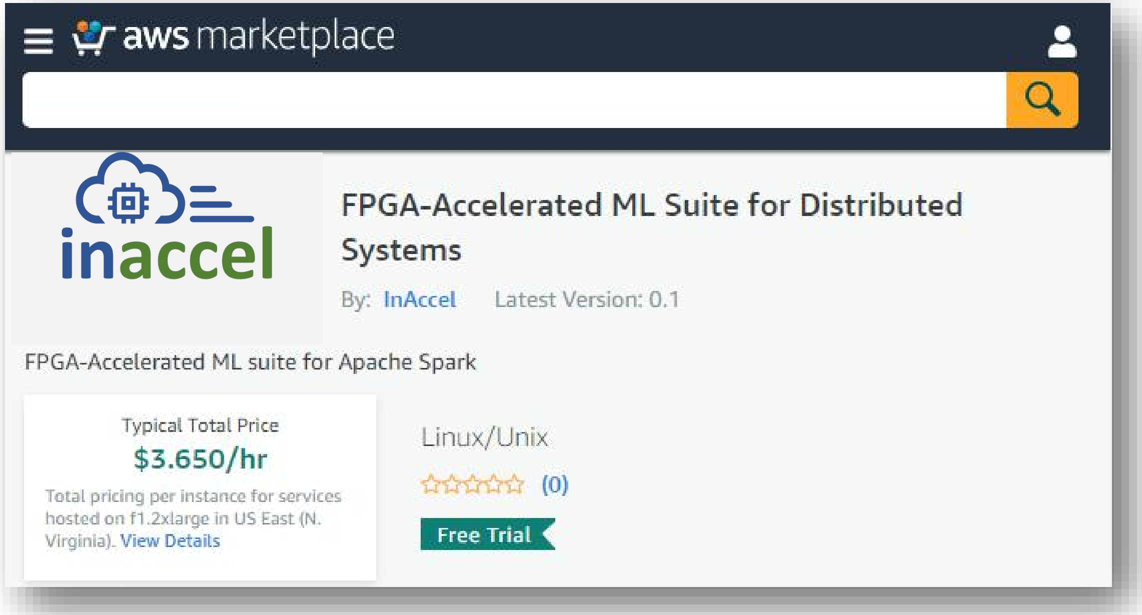
2.Import InAccel API

3.Run your applications on AWS f1 to get 3x – 20x speedup

f1 (8 cores+FPGA)

Cloud Marketplace: available now





FPGA-Accelerated ML Suite for Distributed Systems

By: InAccel Latest Version: 0.1

FPGA-Accelerated ML suite for Apache Spark

Typical Total Price
$3.650/hr
Total pricing per instance for services hosted on f1.2xlarge in US East (N. Virginia). View Details

Linux/Unix

☆☆☆☆☆ (0)

Free Trial

Scalable to worldwide market

First to provide accelerators for Spark

IP cores available in Amazon AWS

Logistic Regression

Gradient Descent IP block for faster training of machine learning algorithms.

K-mean clustering

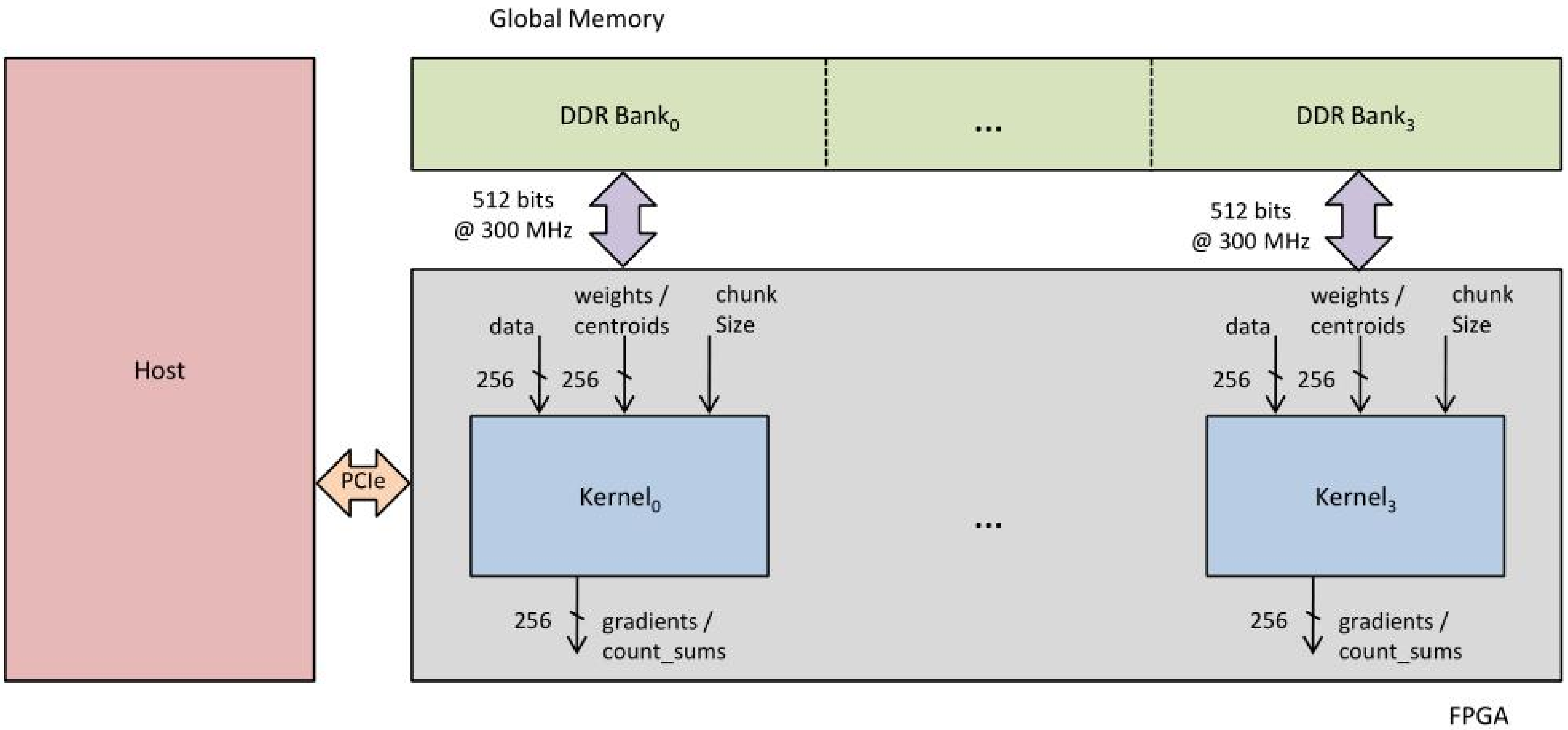
K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.

Recommendation Engines (ALS)

Alternative-Least-Square IP core for the acceleration of recommendation engines based on collaborative filtering.

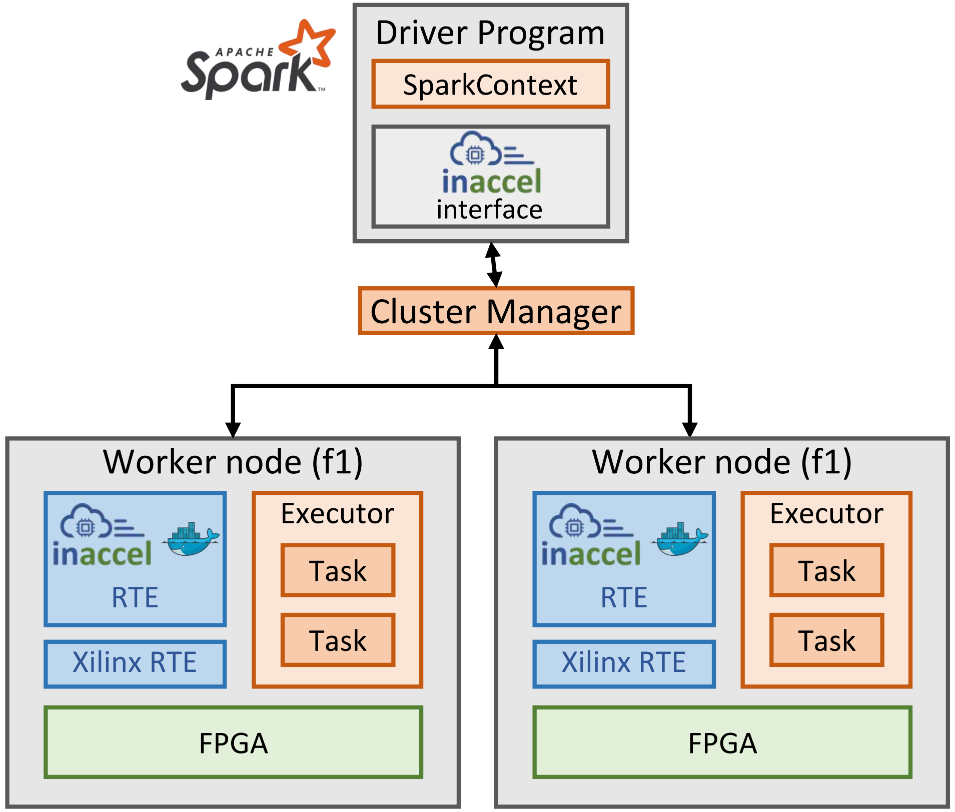
Available in Amazon AWS marketplace for free trial: www.inaccel.com

Communication with Host in Amazon AWS f1.x2 and f1.x16



Accelerators for logistic regression/kmeans

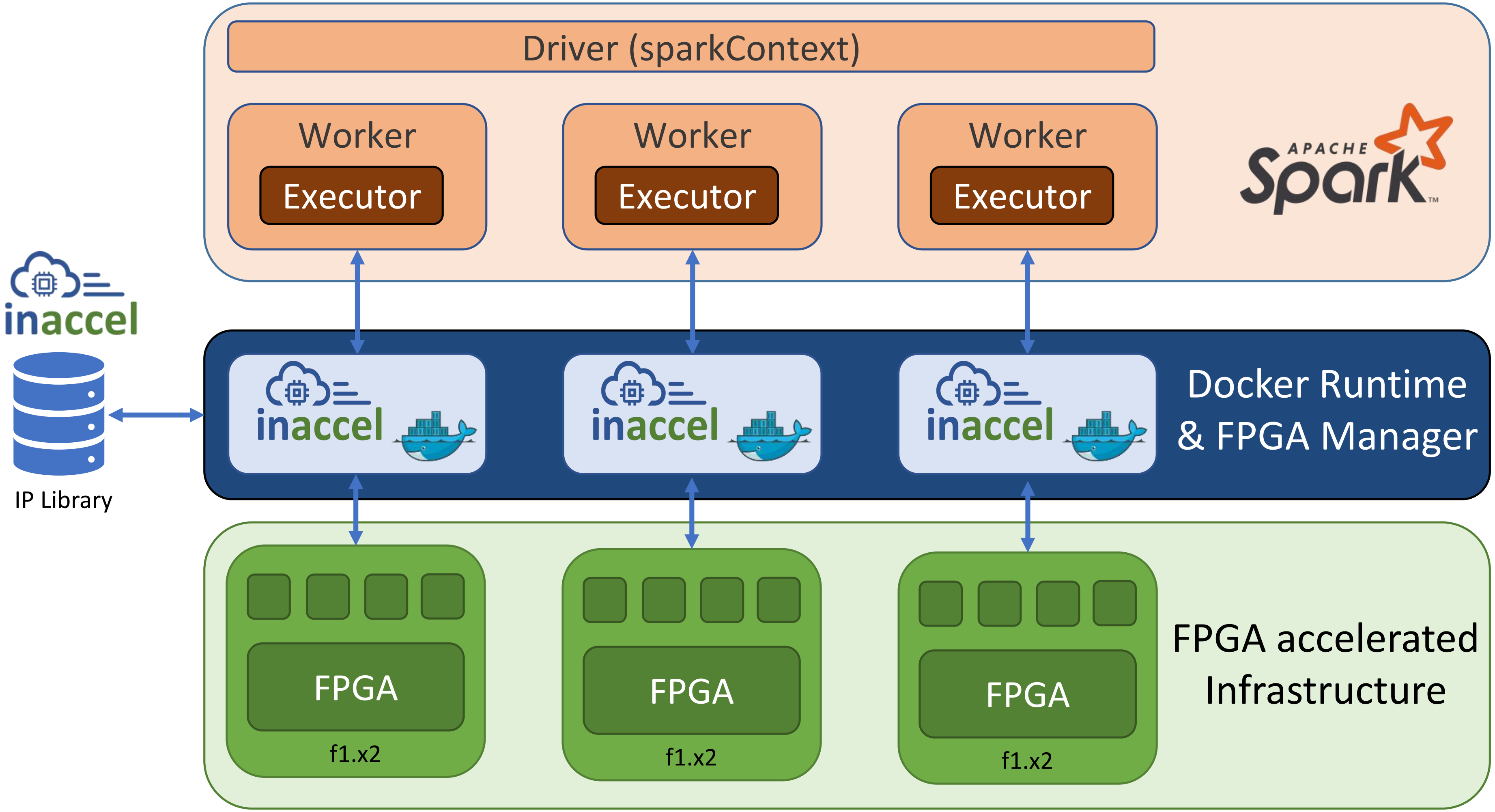
Docker-based implementation for easy integration



- Inaccel’s FPGA manager docker container comprises both an FPGA manager to schedule, orchestrate, and monitor the execution of the accelerated applications but also the required FPGA runtime system.
- The dockerized runtime system detects the FPGA platform (aws F1) and manages the interaction/communication with the FPGA (i.e., loading the accelerator, transferring input data and results), making it transparent to the application.
- Docker swarm, Kubernetes, naïve execution

Cluster mode

- Cluster mode



Demo on Amazon AWS

Intel 36 cores Xeon on Amazon AWS

c4.8xlarge $1.592/hour

8 cores + inaccel

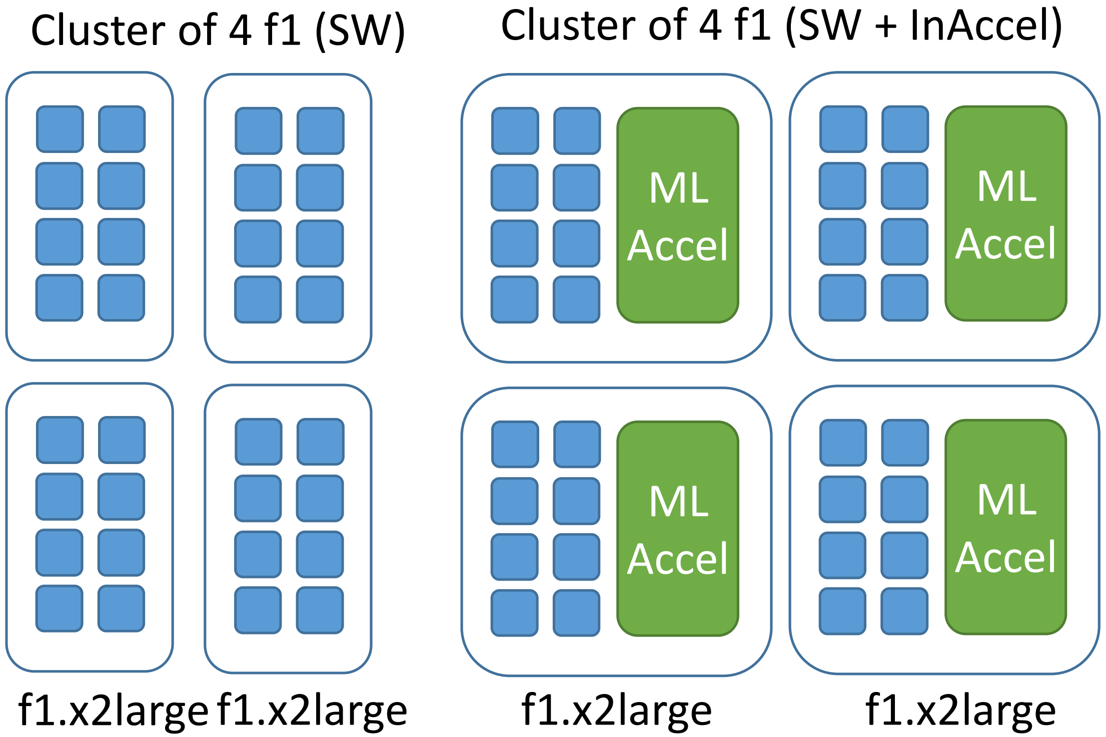
in Amazon AWS FPGA

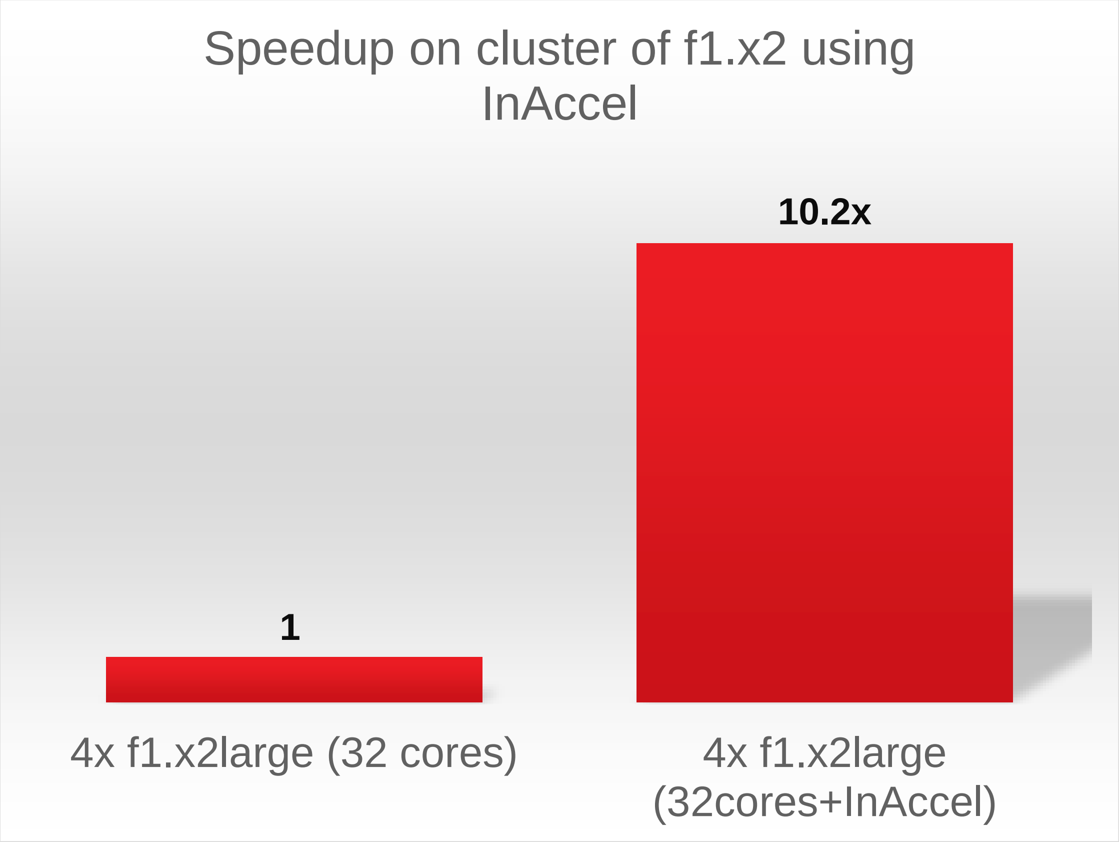
f1.2xlarge $1.65/hour + inaccel

Note: 4x fast forward for both cases

Speedup comparison

> **Up to 10x speedup compared to 32 cores based on f1.x2**

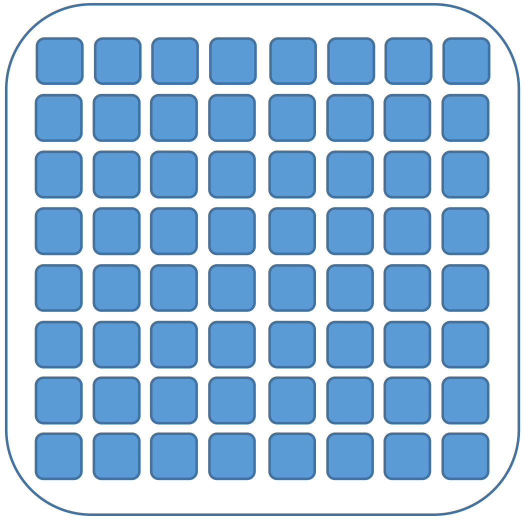




Speed up

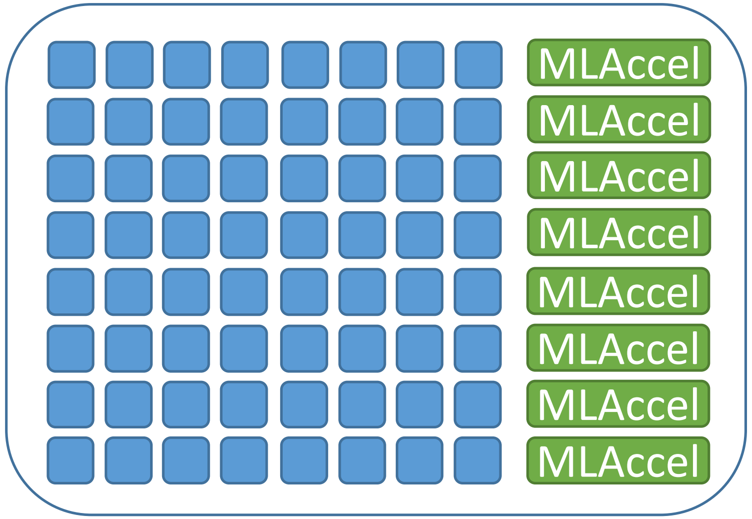
> **Up to 12x speedup compared to 64 cores on f1.x16**

f1.x16large (SW)

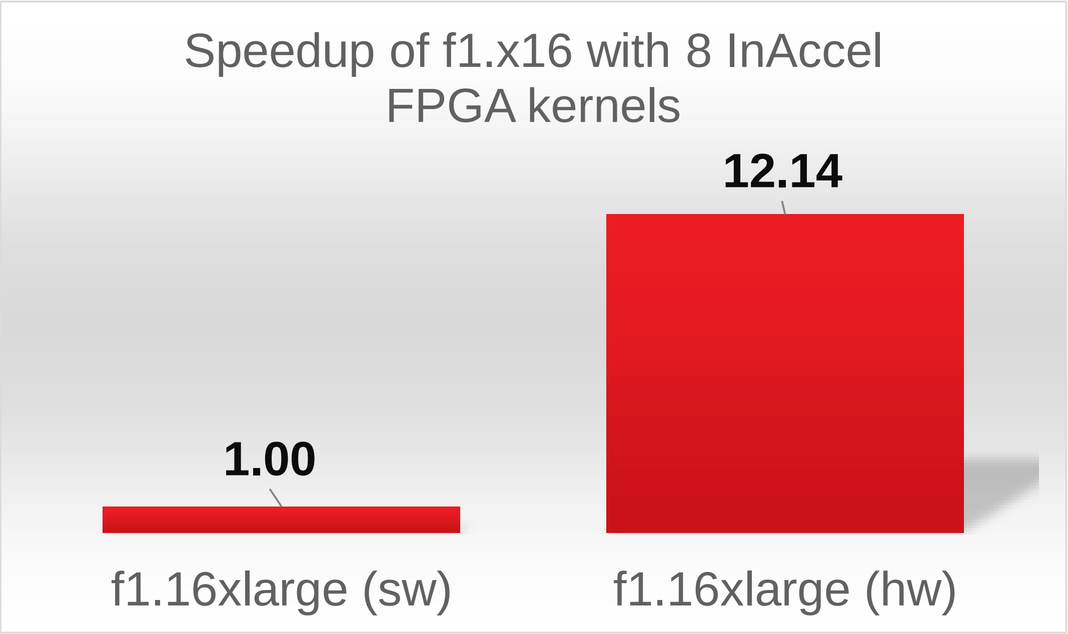


64 cores

f1.x16large (SW + 8 InAccel cores)

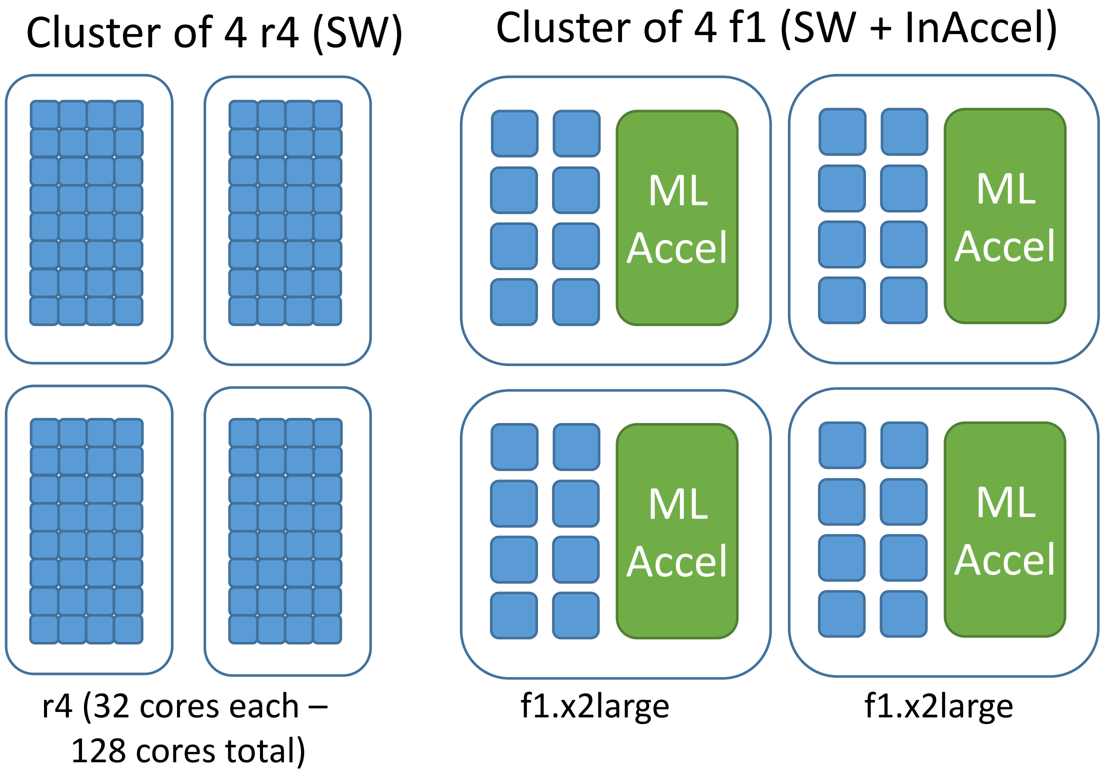


64 cores + 8 FPGAs with InAccel


Speedup of f1.x16 with 8 InAccel FPGA kernels

Speedup comparison

- 3x Speedup compared to r4
- 2x lower OpEx



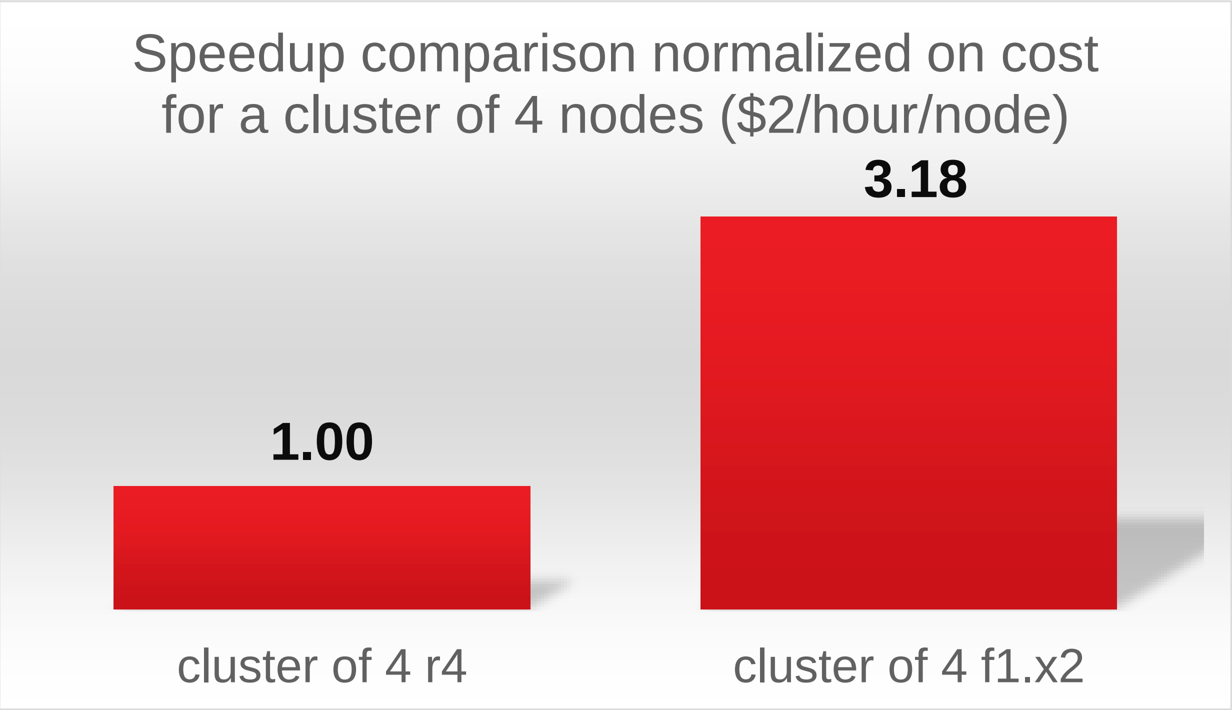
Cluster of 4 r4 (SW)

Cluster of 4 f1 (SW + InAccel)

r4 (32 cores each – 128 cores total)

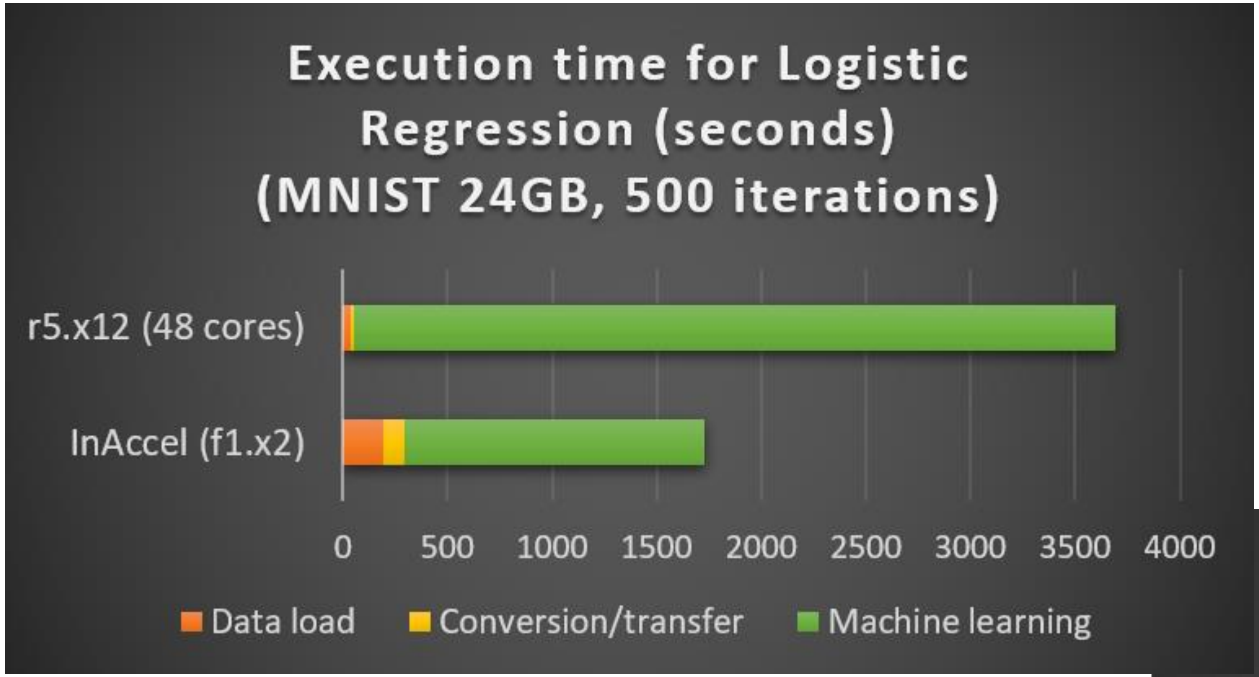
f1.x2large

f1.x2large

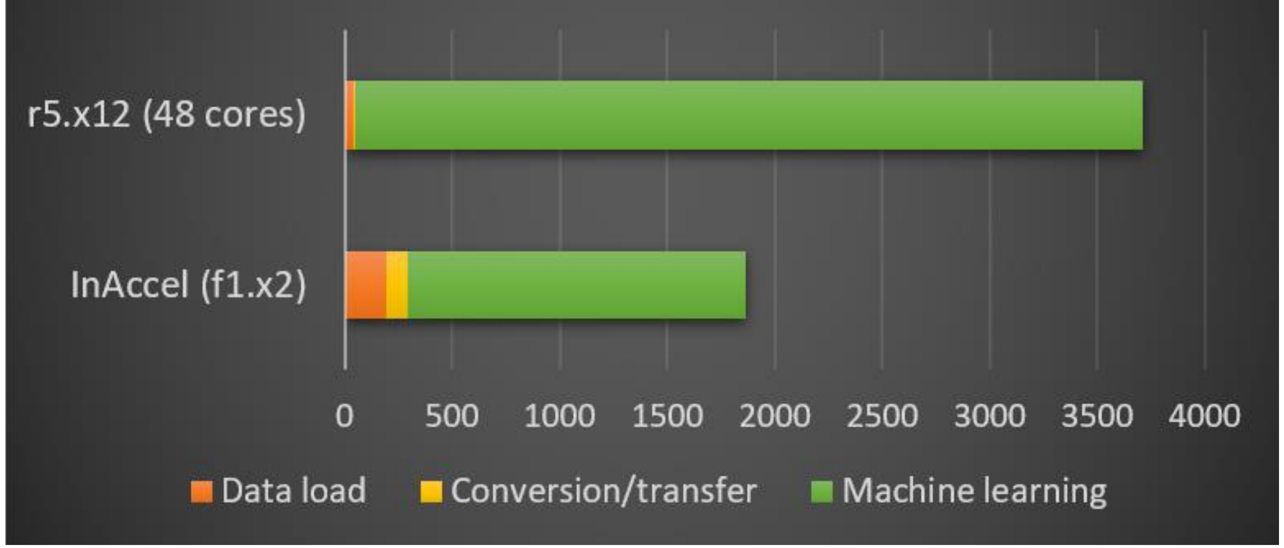


Speedup comparison normalized on cost for a cluster of 4 nodes ($2/hour/node)

Performance evaluation

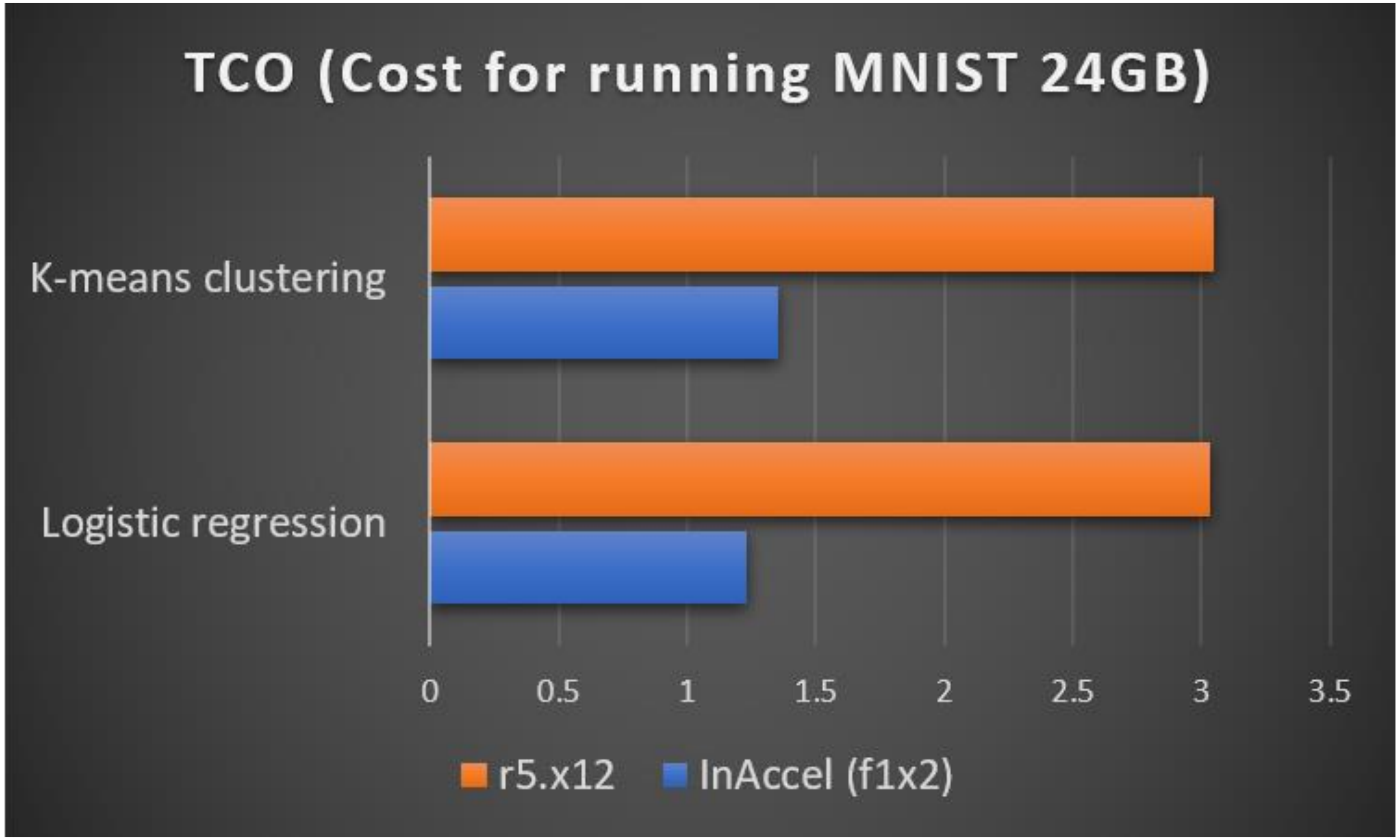


Execution time for K-Means (seconds)
(MNIST 24GB, 500 iterations)



Cost reduction

- Up to 3x lower cost to train your ML model



Try for free on Amazon AWS

Single node version

- Single-node Machine learning accelerators for Amazon f1.x2large instances providing APIs for C/C++, Java, Python and Scala for easy integration

Single node ML suite

Distributed version for Apache Spark

- Machine learning accelerators for Apache Spark providing all the required APIs and libraries for the seamless integration in distributed systems

Distributed node ML suite

InAccel unique Advantages

Compatible with Amazon AWS

All accelerators are compatible with the Amazon AWS F1 instances. AWS compatibility allows easy and fast deployment of the accelerators and seamless integration with your current AWS applications.

Seamless integration with your code

InAccel provides all the required APIs for the seamless integration of the accelerators without any modifications on your original code.

Acceleration of your code

Accelerators from InAccel provide up to 2x-10x speedup compared to contemporary processors in typical servers.

XILINX
ALLIANCE PROGRAM
MEMBER

aws partner network

awsmarketplace

Adaptable.

Intelligent.

inaccel