



FPGA Acceleration of Apache Spark on the Cloud, Instantly



Dr. Chris Kachris
CEO, co-founder
Oct 2 2018



...or

How to **speedup** your Spark ML

applications

with the same cost

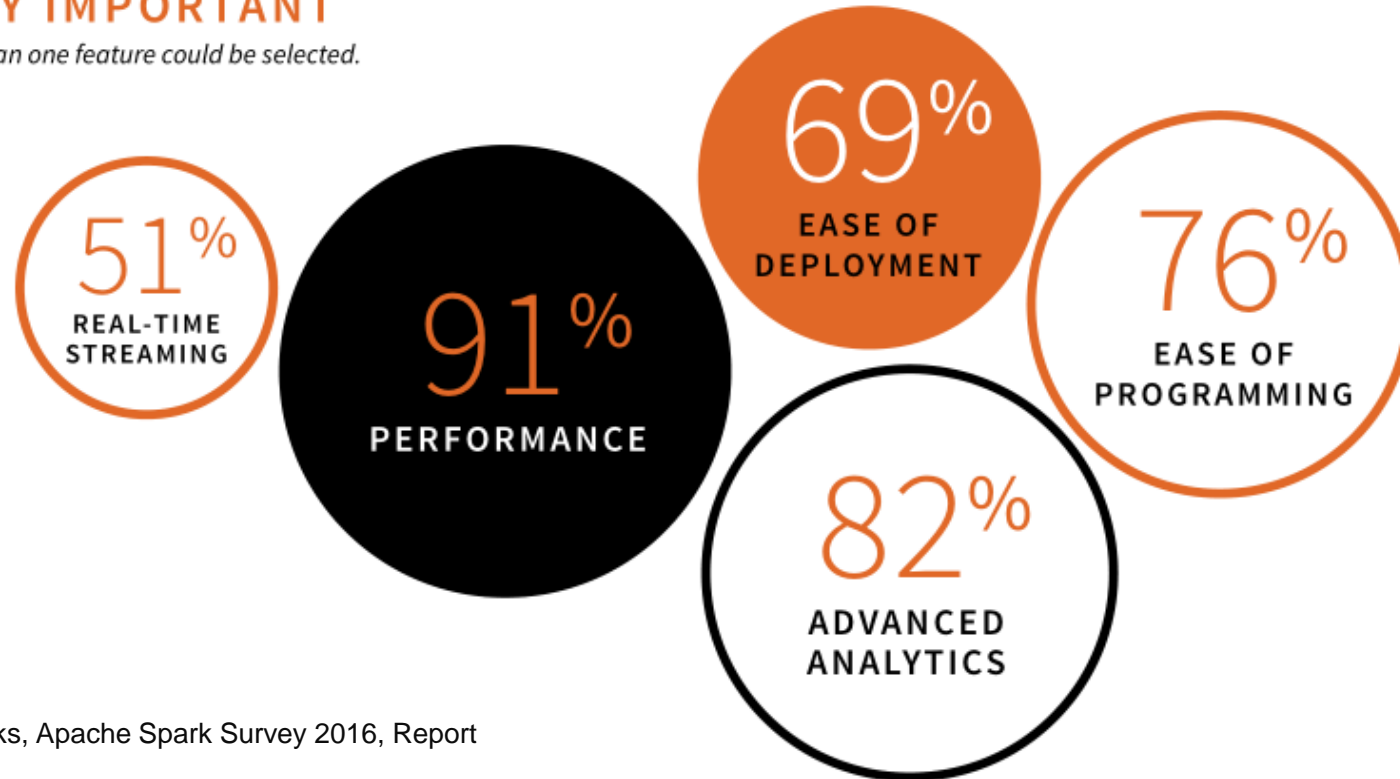
with the same code

Why acceleration

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

% OF RESPONDENTS WHO CONSIDERED THE FEATURE
VERY IMPORTANT

More than one feature could be selected.



Source: Databricks, Apache Spark Survey 2016, Report

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]



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



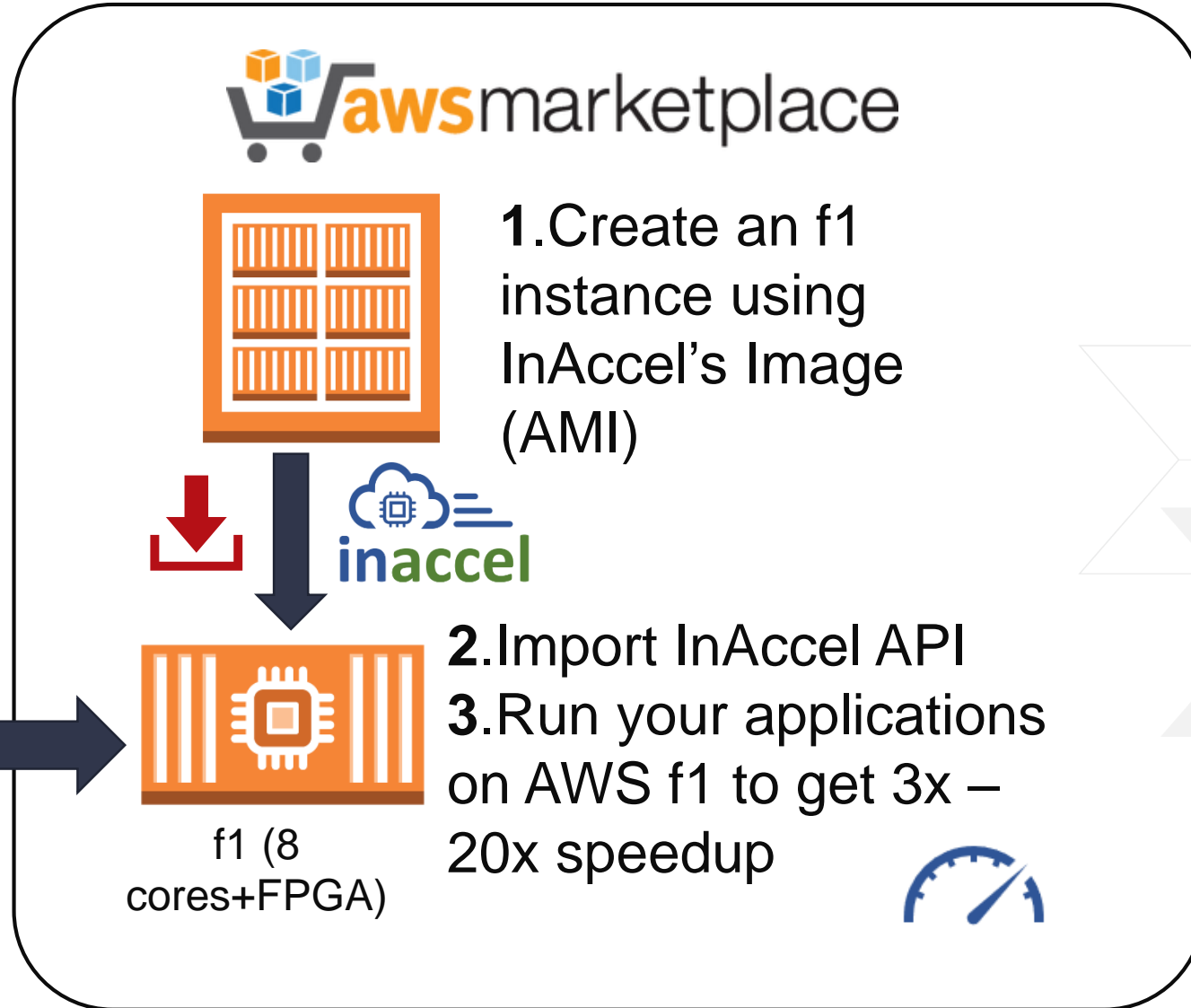
ADVANCED ANALYTICS USERS (MLLIB)
IN PRODUCTION

+ 38%

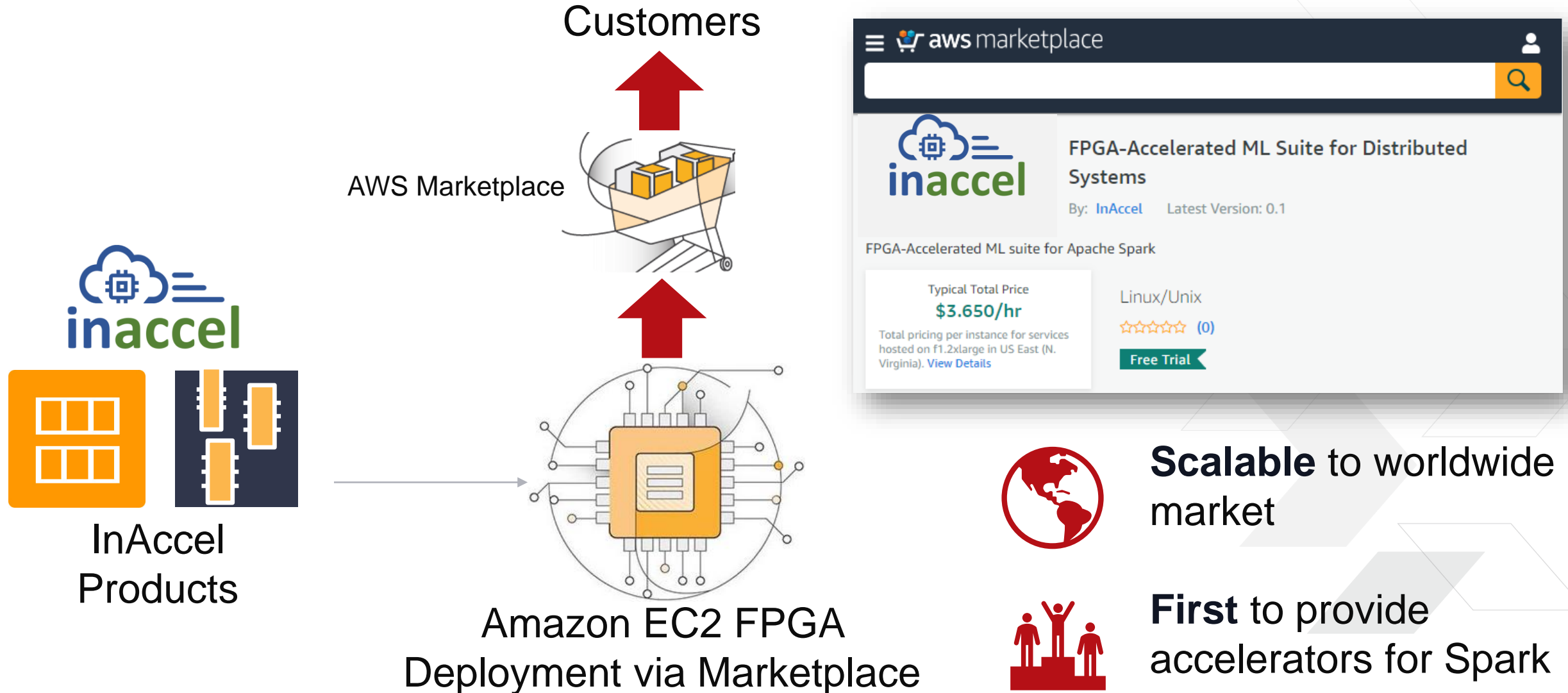
2015
13%
OF RESPONDENTS

2016
18%
OF RESPONDENTS

Accelerators for Spark ML in Amazon AWS in 3 steps

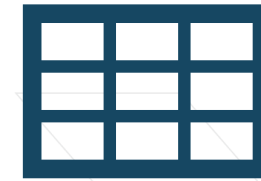


Cloud Marketplace: available now



IP cores available in Amazon AWS

Logistic Regression
K-means clustering
Recommendation Engines (ALS)



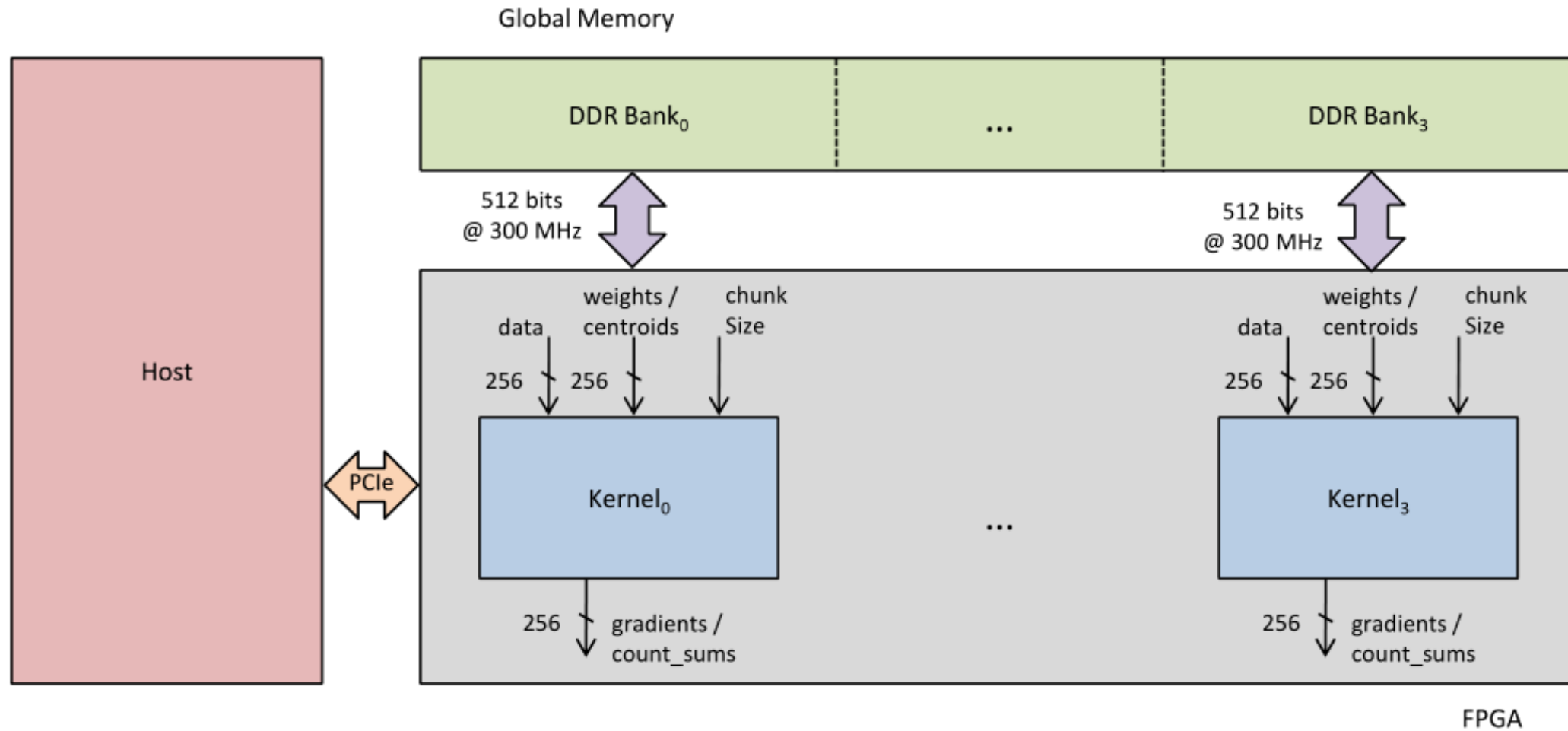
Gradient Descent IP block for faster training of machine learning algorithms.

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

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

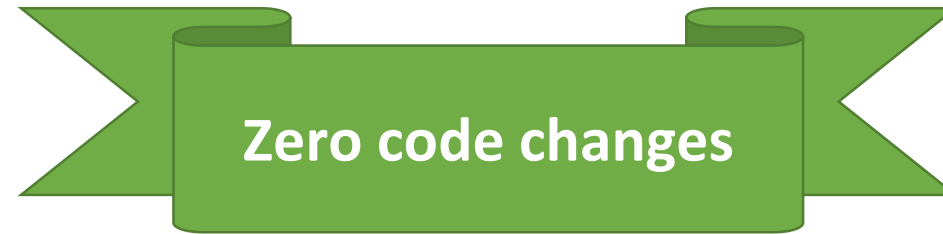
Available in Amazon AWS marketplace for free trial: www.inacel.com

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

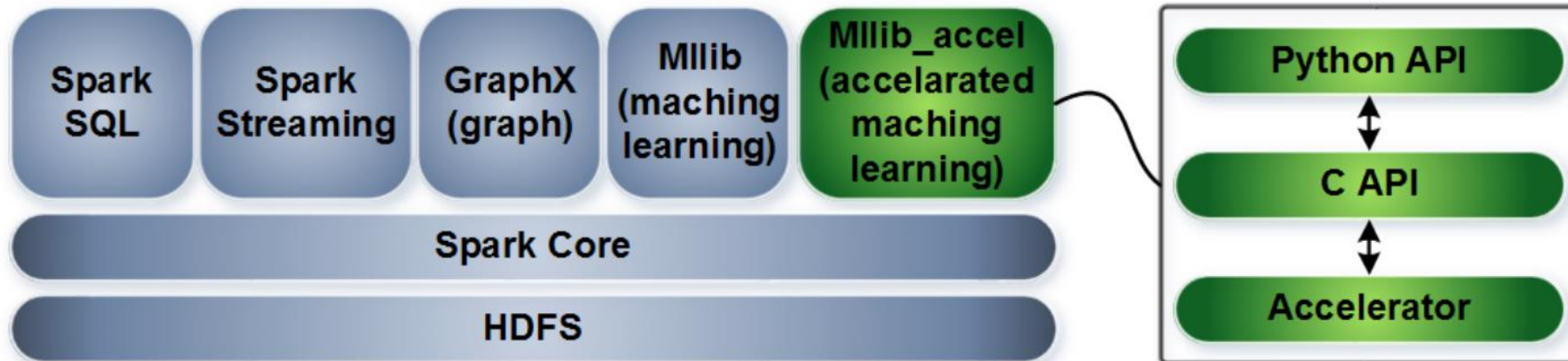


Accelerators for logistic regression/kmeans

Zero code changes




> Only replacement of the library is required



Demo on Amazon AWS



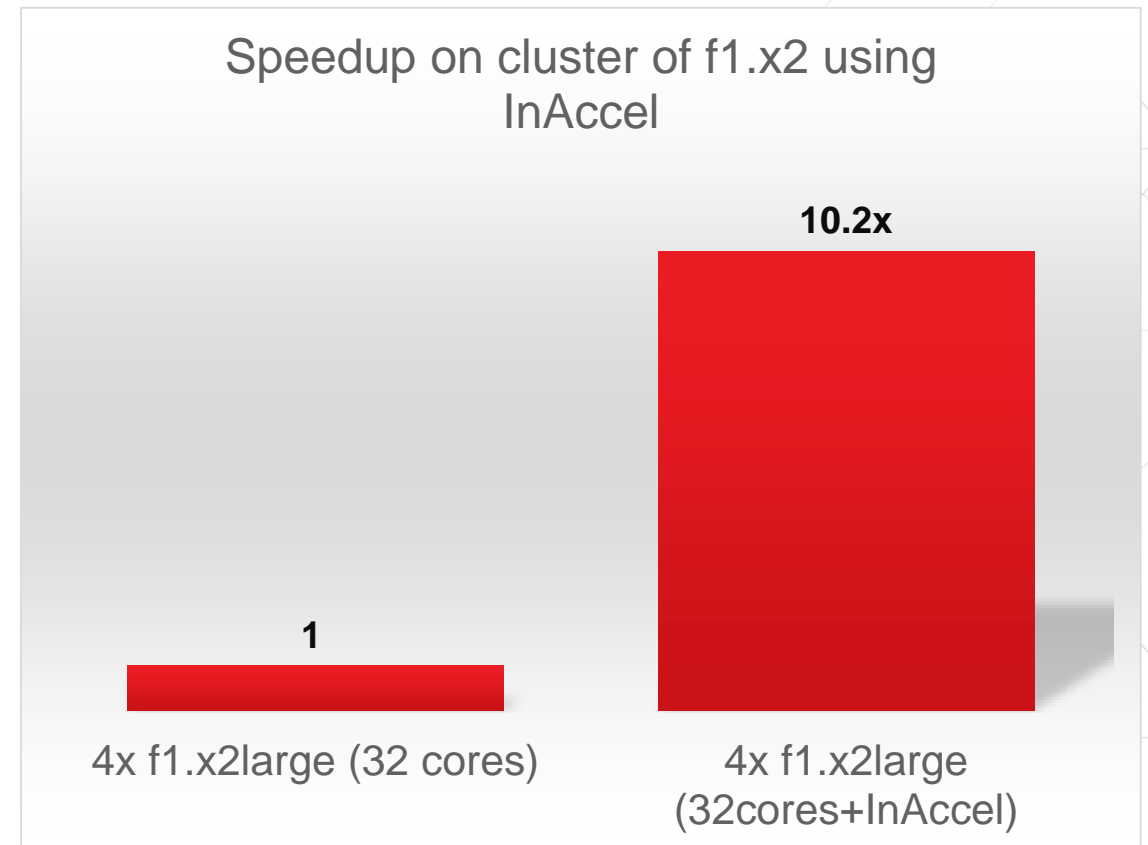
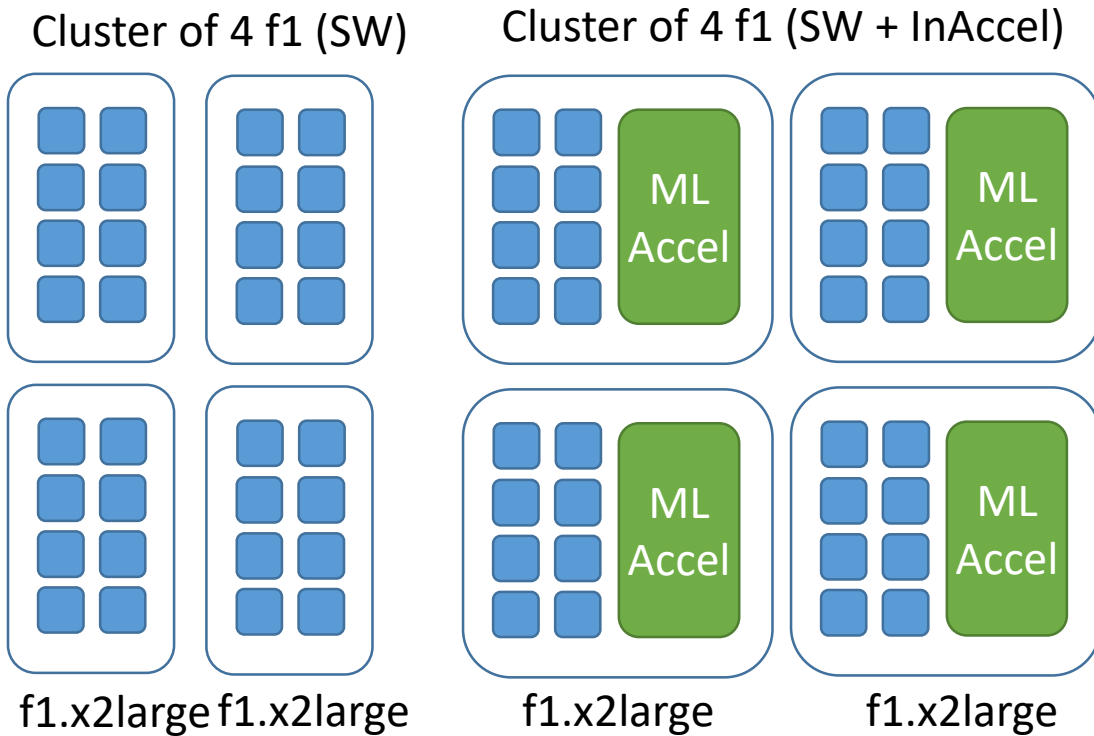
Intel 36 cores Xeon on Amazon AWS
c4.8xlarge \$1.592/hour

8 cores +  inaccel
in Amazon AWS FPGA
f1.2xlarge \$1.65/hour + inaccel

Note: 4x fast forward for both cases

Speedup comparison

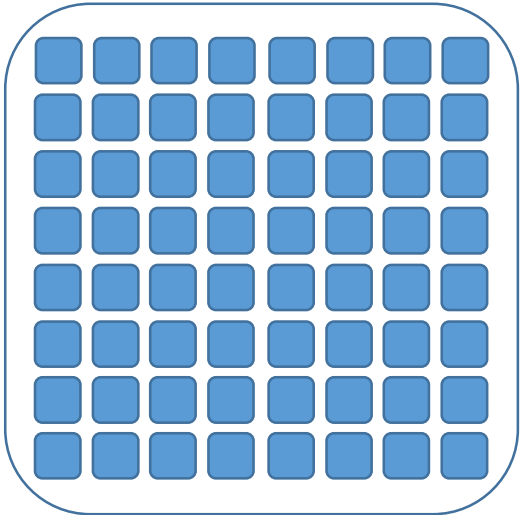
> Up to 10x speedup compared to 32 cores based on f1.x2



Speed up

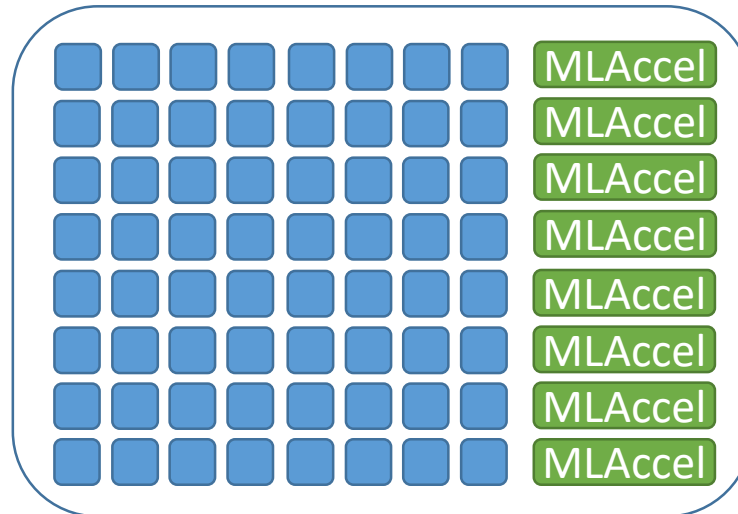
> Up to 12x speedup compared to 64 cores on f1.x16

f1.x16large (SW)



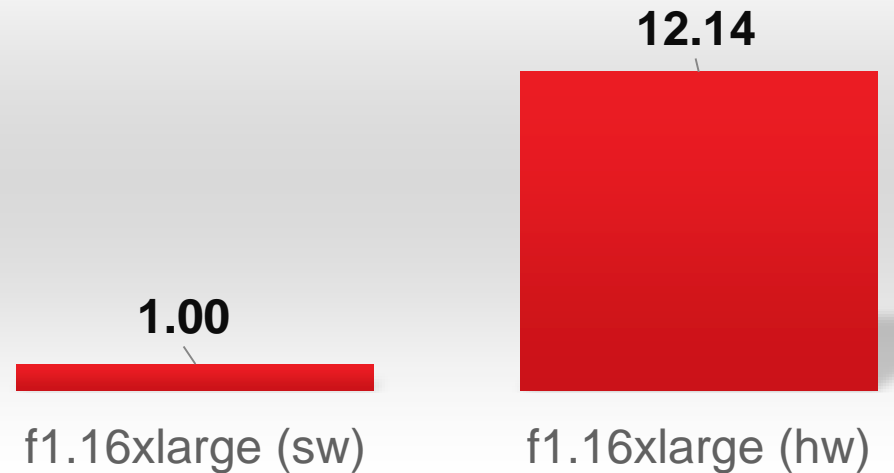
64 cores

f1.x16large (SW + 8 InAccel cores)



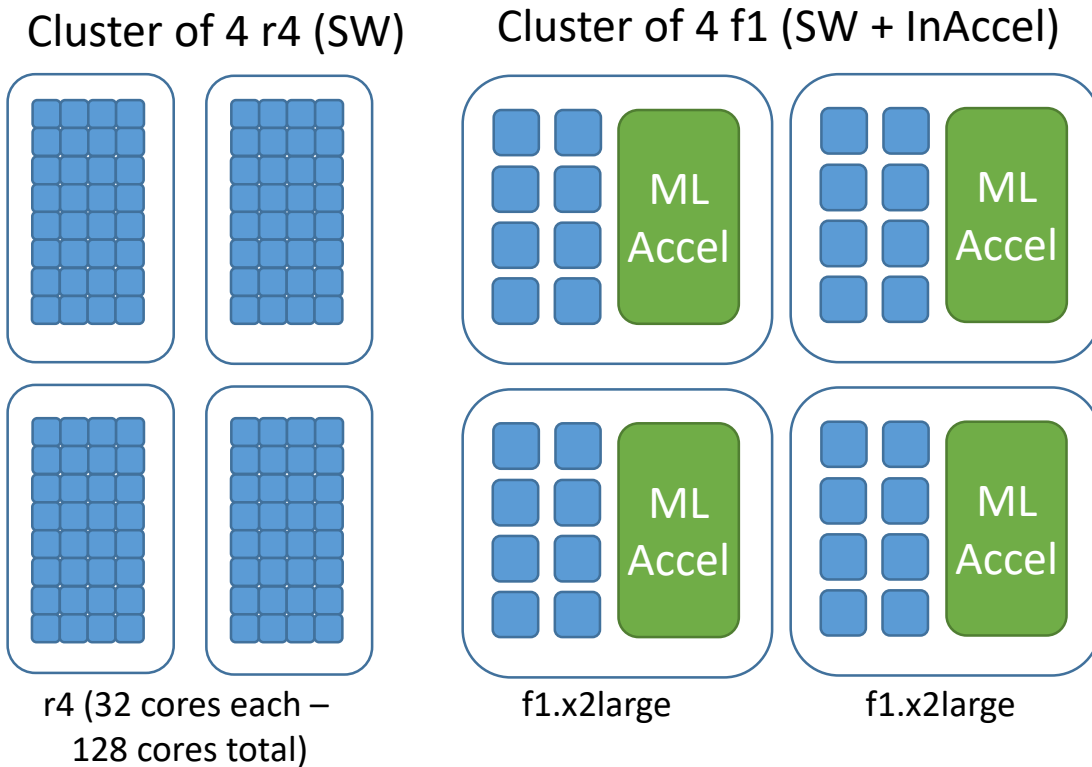
64 cores + 8 FPGAs with InAccel

Speedup of f1.x16 with 8 InAccel FPGA kernels

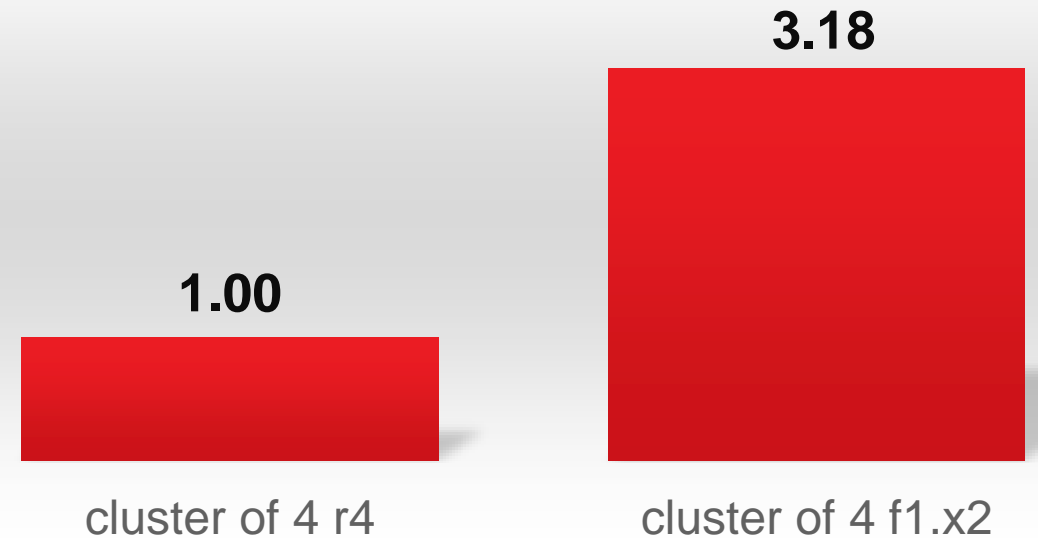


Speedup comparison

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



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



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
DEVELOPER
FORUM

