XDF XILINX DEVELOPER FORUM

Acceleration of Spark ML on the Cloud using container-based FPGAs

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

XILINX

Outline

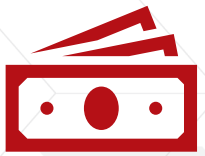
- **A use case on Machine learning acceleration on the Cloud**
  - Data scientists/engineers

- **An FPGA Manager to scale your FPGA design on the cloud**
  - FPGA engineers

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.

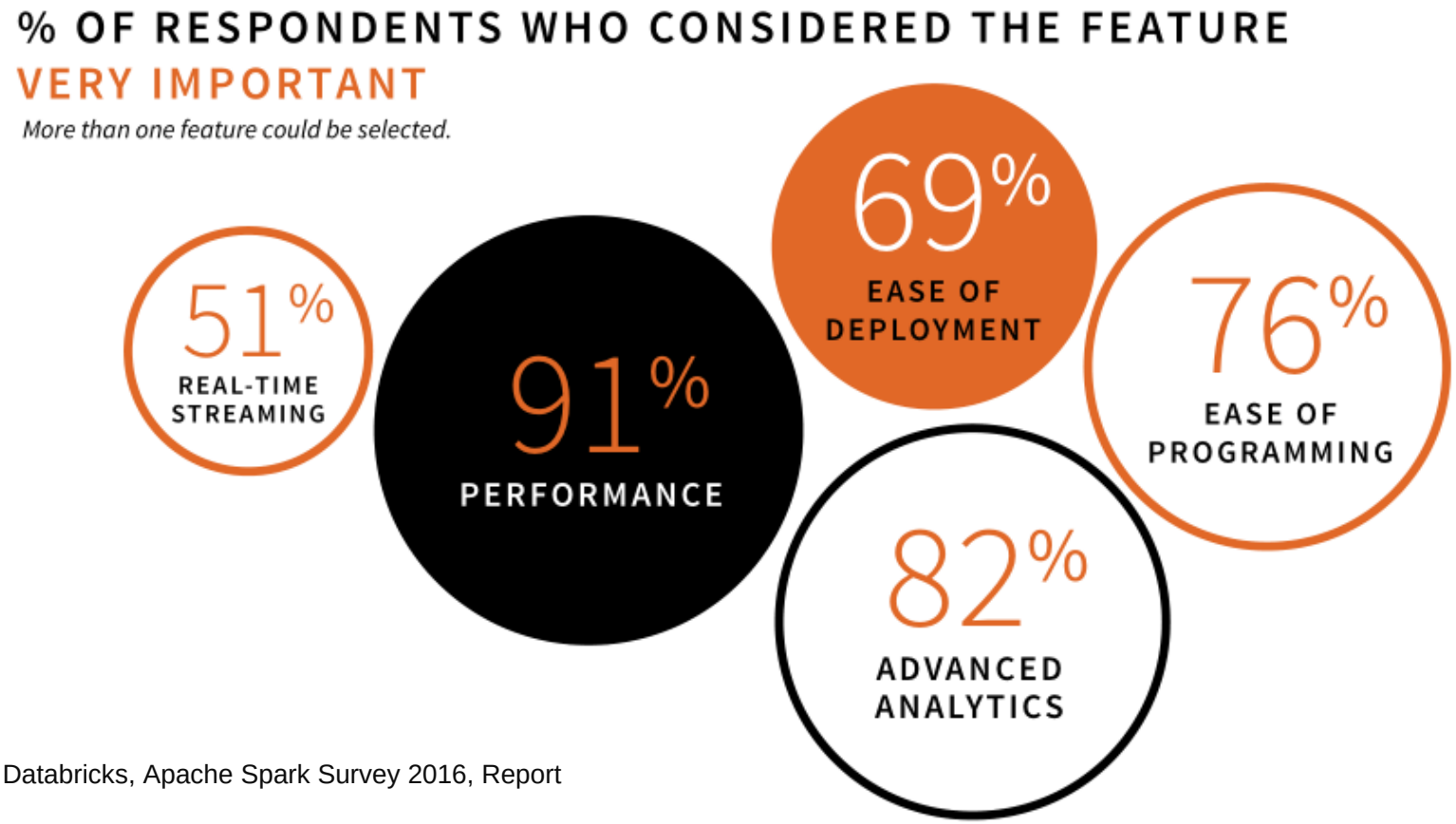
- The market for FPGA is expected to grow at **the highest CAGR during the forecast period** owing to the increasing adoption of FPGAs for the acceleration of enterprise workloads.



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

Why acceleration

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



Source: Databricks, Apache Spark Survey 2016, Report

inaccel

helps companies speedup their applications

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

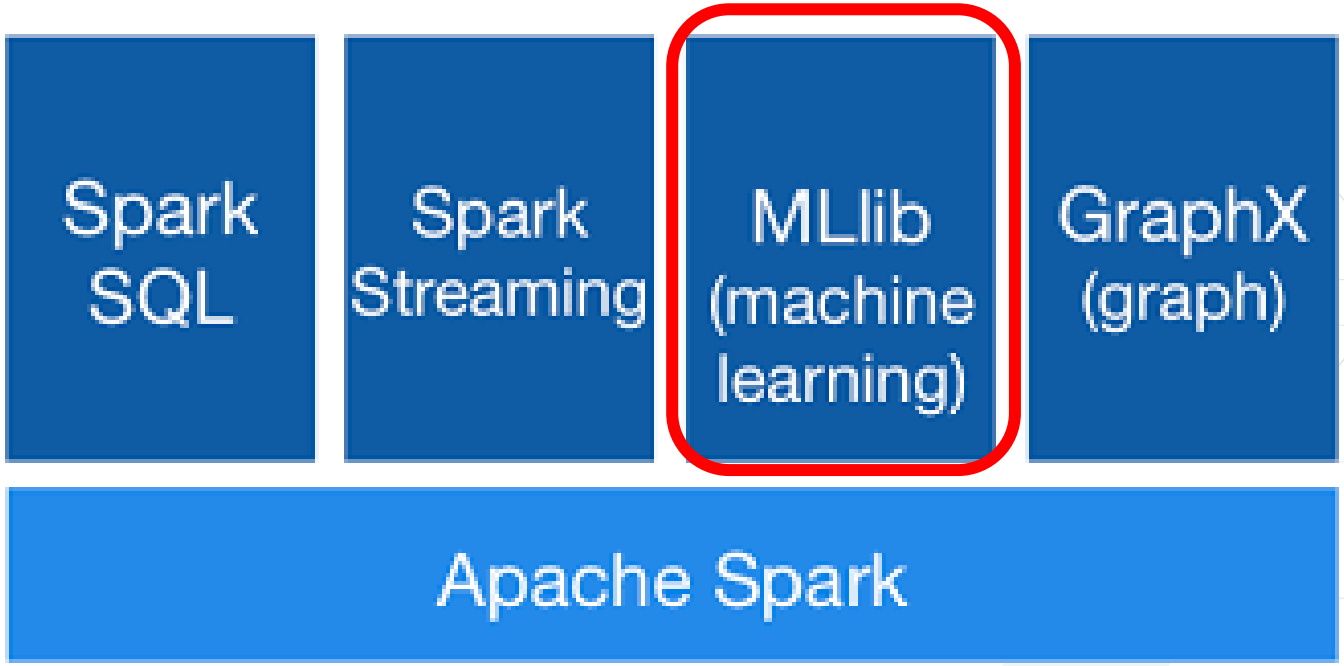
3x-10x Speedup

2x Lower Cost

Zero code changes

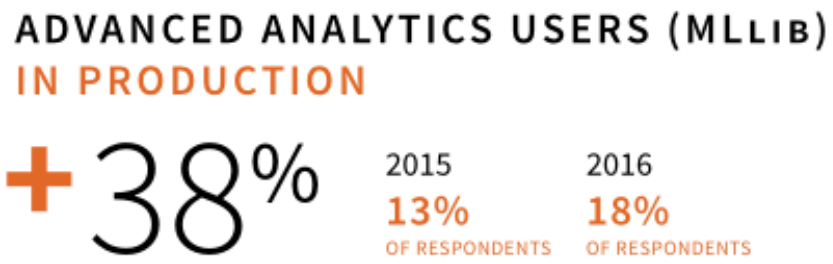
Apache Spark

- **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

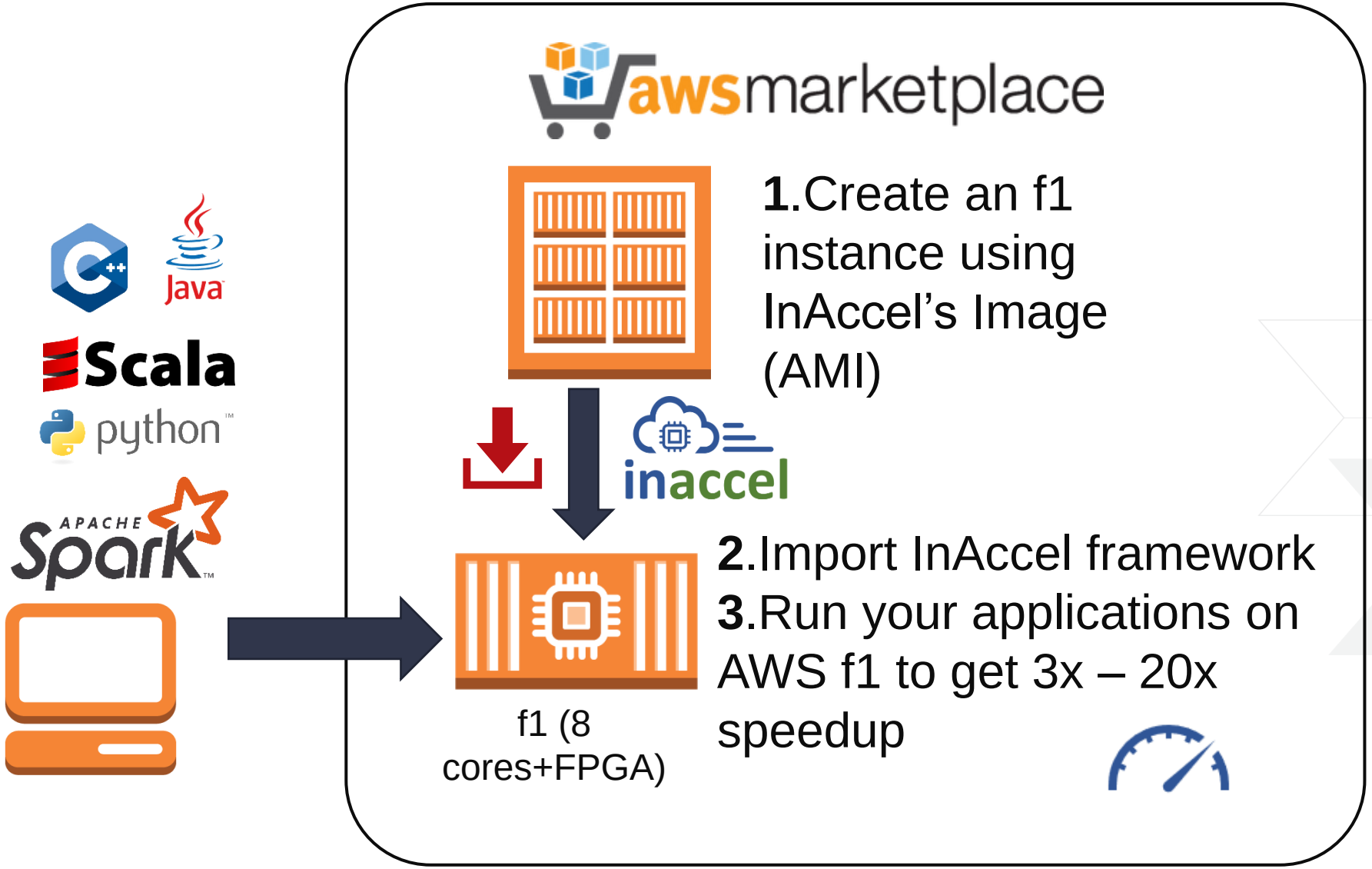


Acceleration for machine learning

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



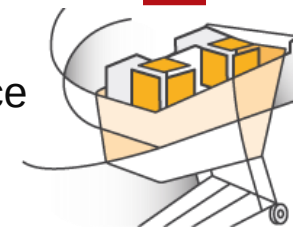
Accelerators for Spark ML in Amazon AWS in 3 steps

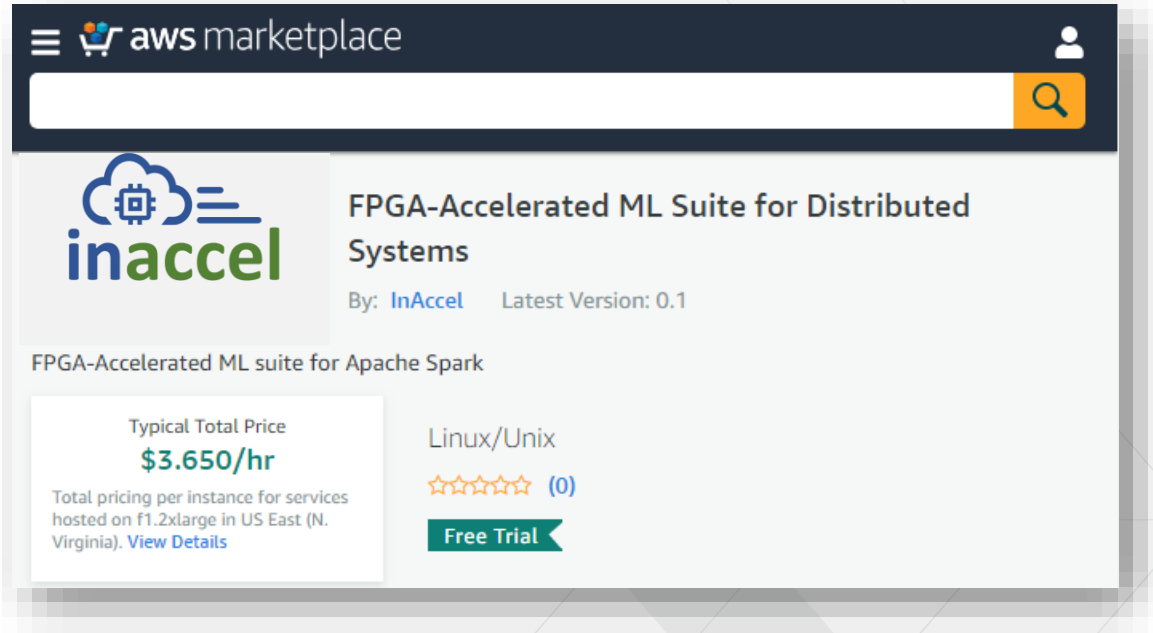


Cloud Marketplace: available now

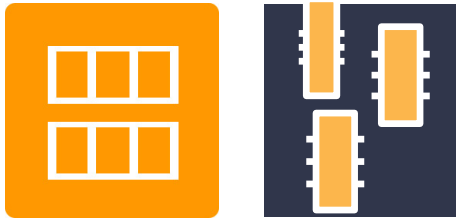
Customers

AWS Marketplace

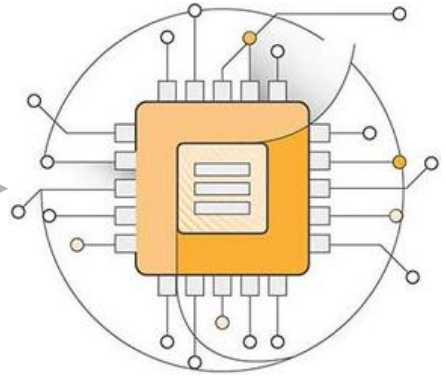




inaccel



InAccel Products



Amazon EC2 FPGA Deployment via Marketplace

Scalable to worldwide market

First to provide accelerators for Spark

IP cores available in Amazon AWS

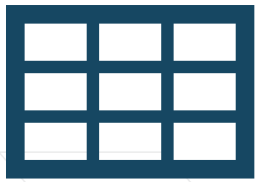
Logistic Regression

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.

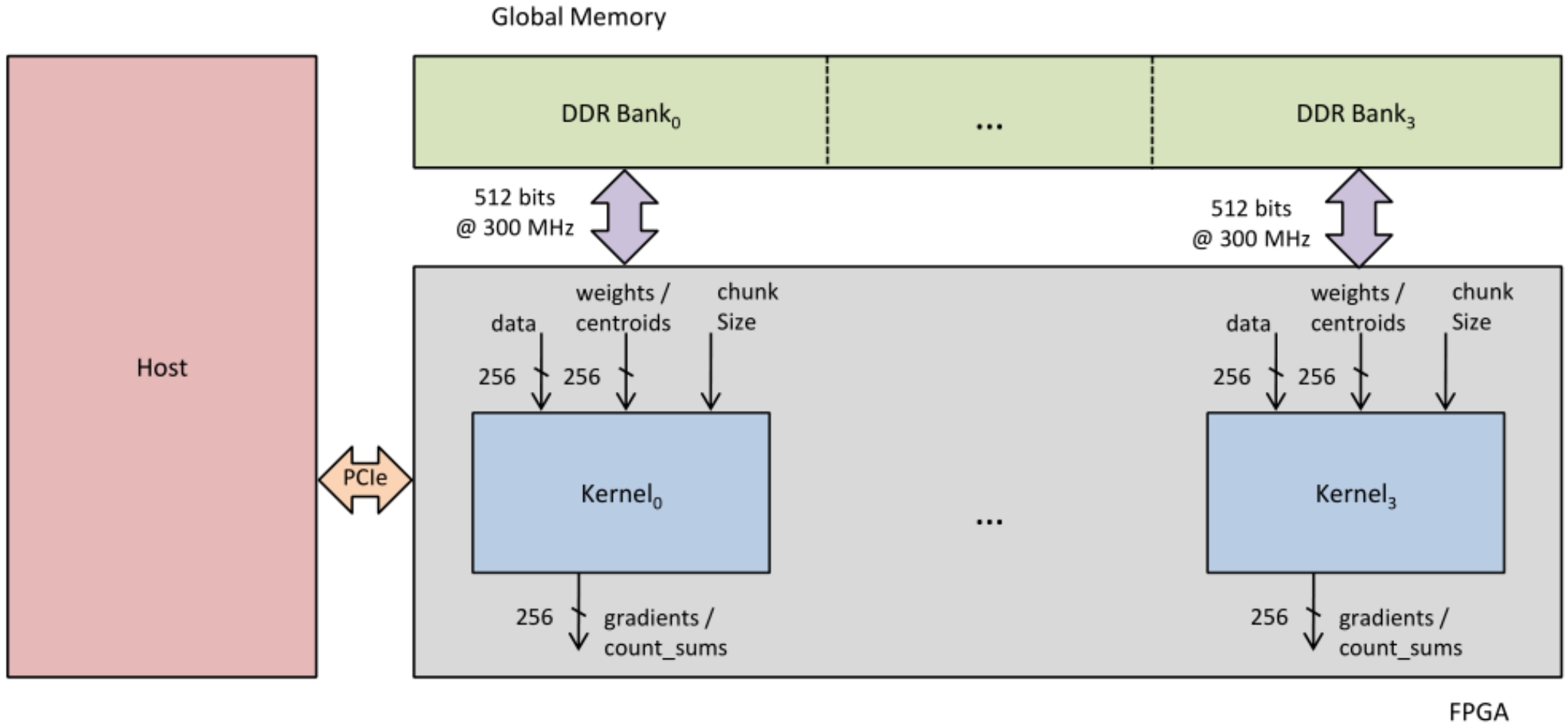
Recommendation Engines (ALS)



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

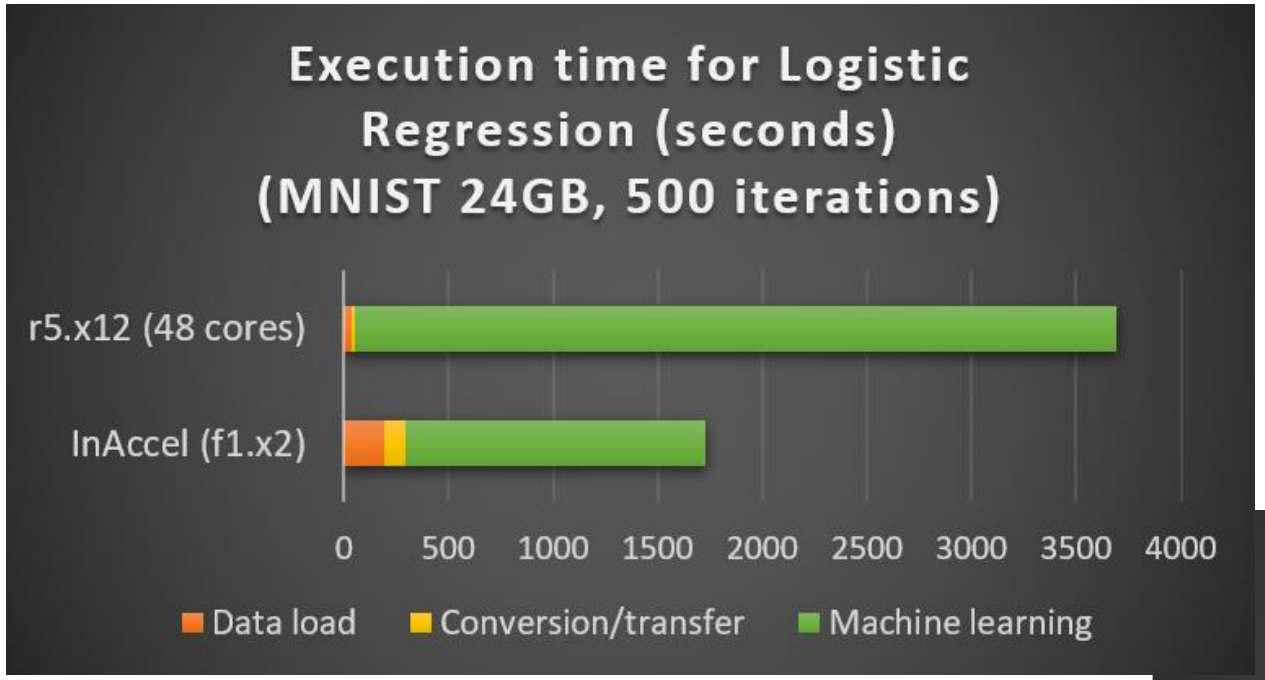
Available in Amazon AWS marketplace for free trial: www.inaccel.com

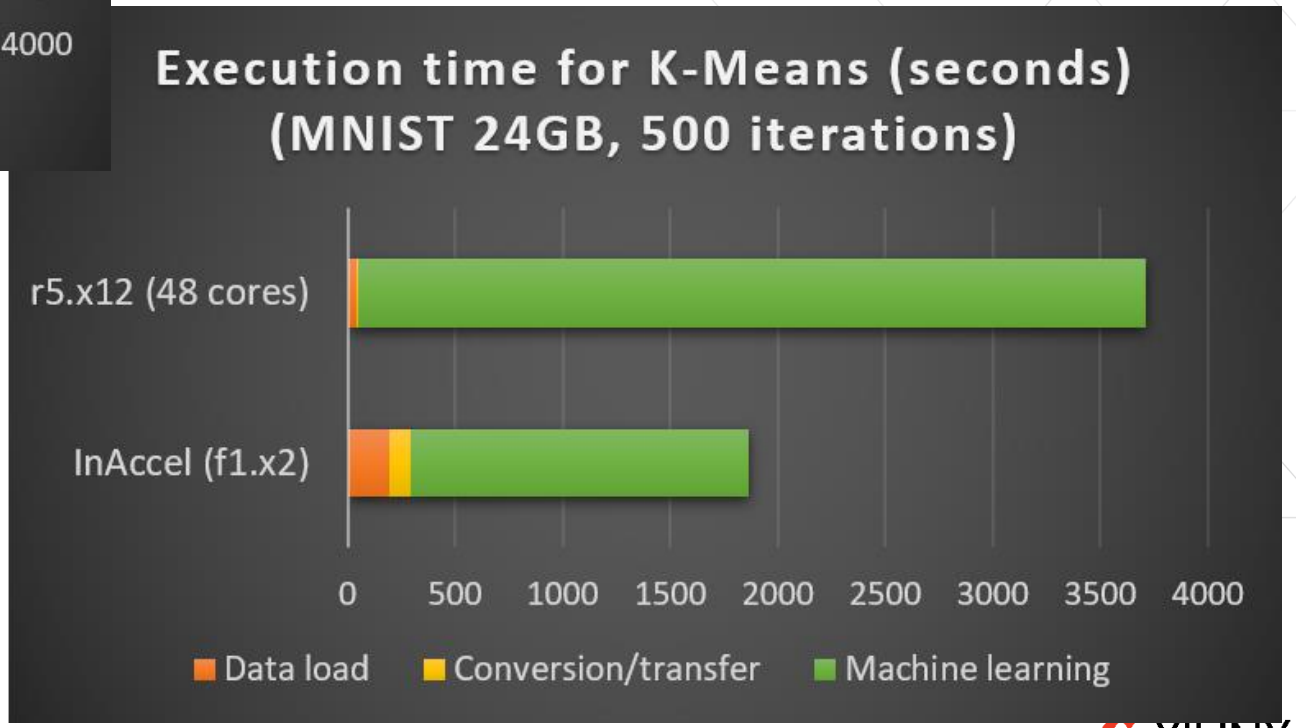
Communication with Host in Amazon AWS f1.x2 and f1.x16



Accelerators for logistic regression/kmeans

Performance evaluation





Demo on Amazon AWS

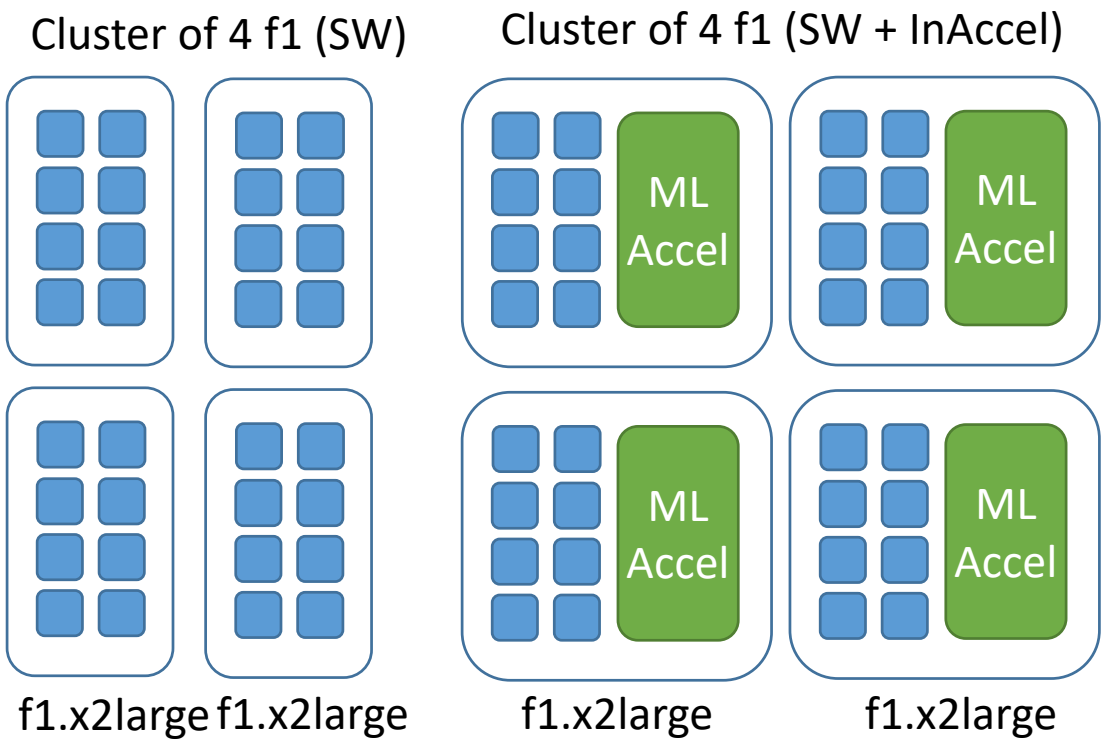
Intel 36 cores Xeon on Amazon AWS
c4.8xlarge $1.592/hour

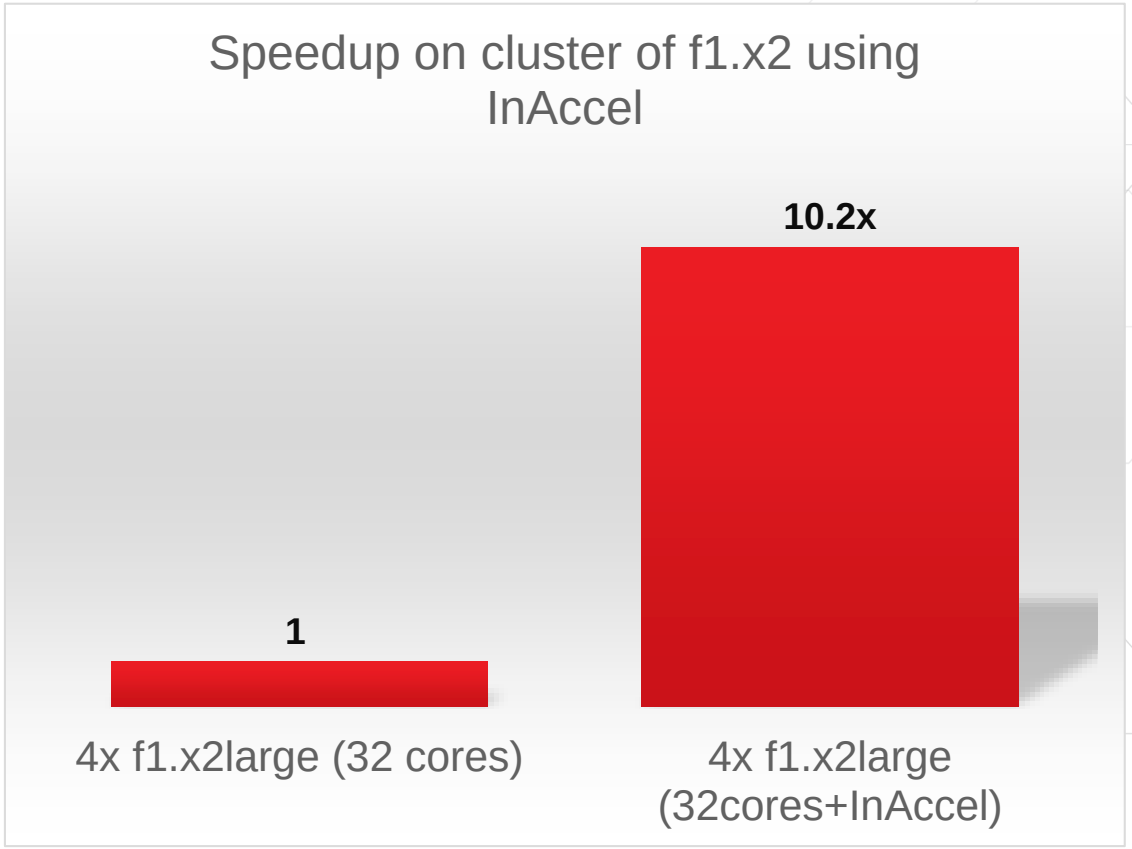
8 cores + inaccel
in Amazon AWS FPGA
f1.2xlarge $1.65/hour + inaccel

Note: 4x fast forward for both cases

Speedup comparison

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

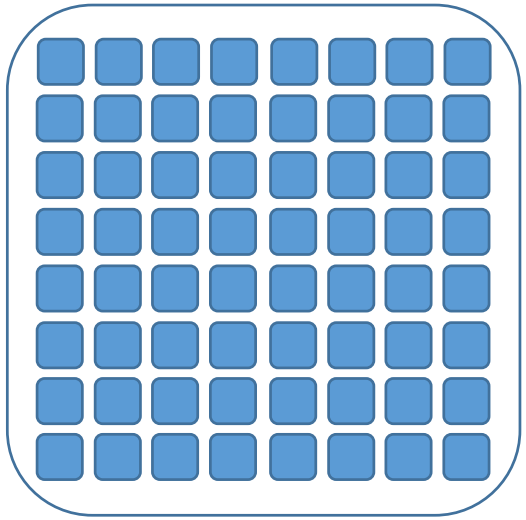




Speed up

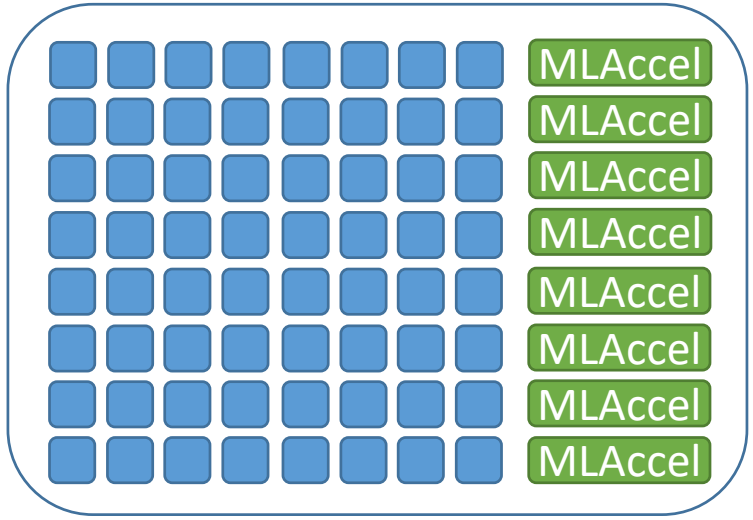
> **Up to 12x speedup compared to 64 cores on f1.x16**

f1.x16large (SW)

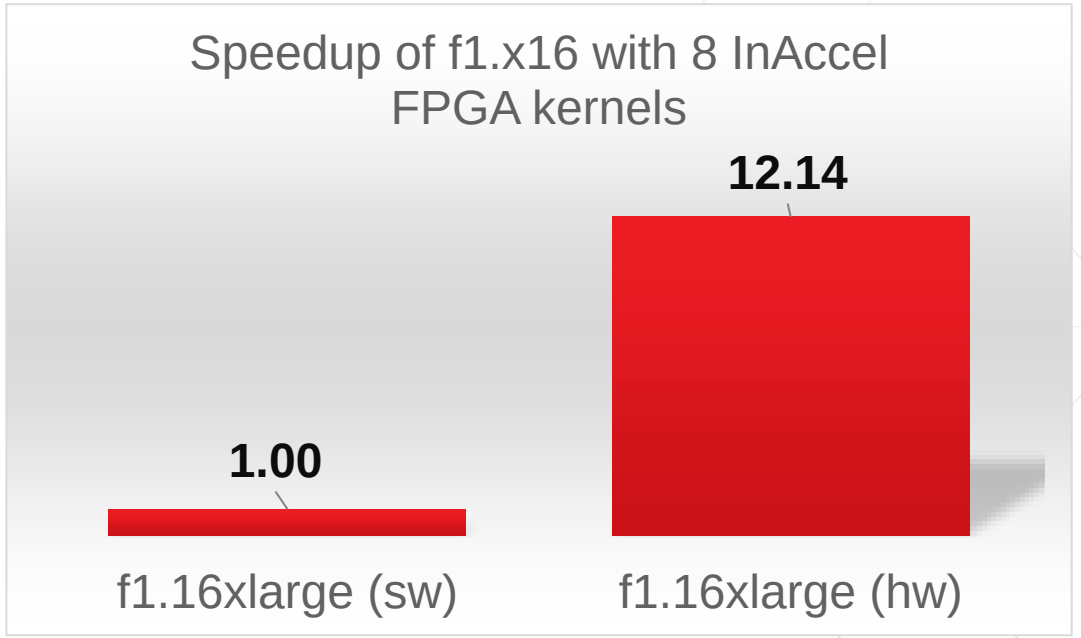


64 cores

f1.x16large (SW + 8 InAccel cores)



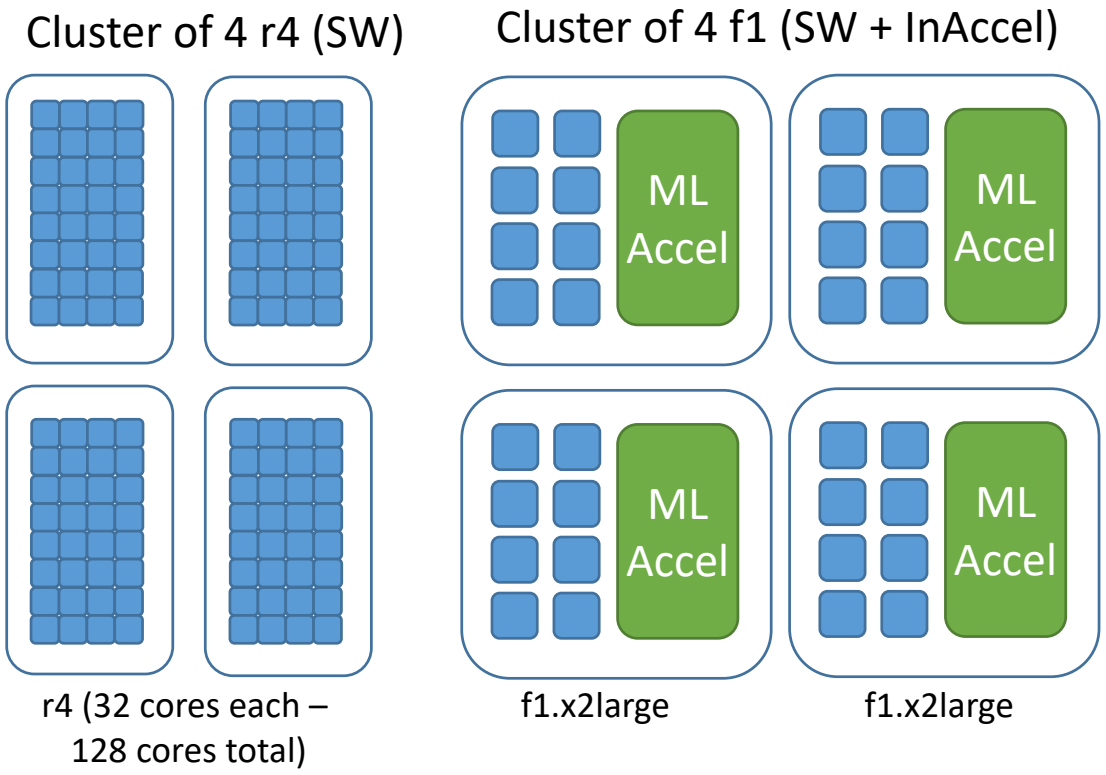
64 cores + 8 FPGAs with InAccel

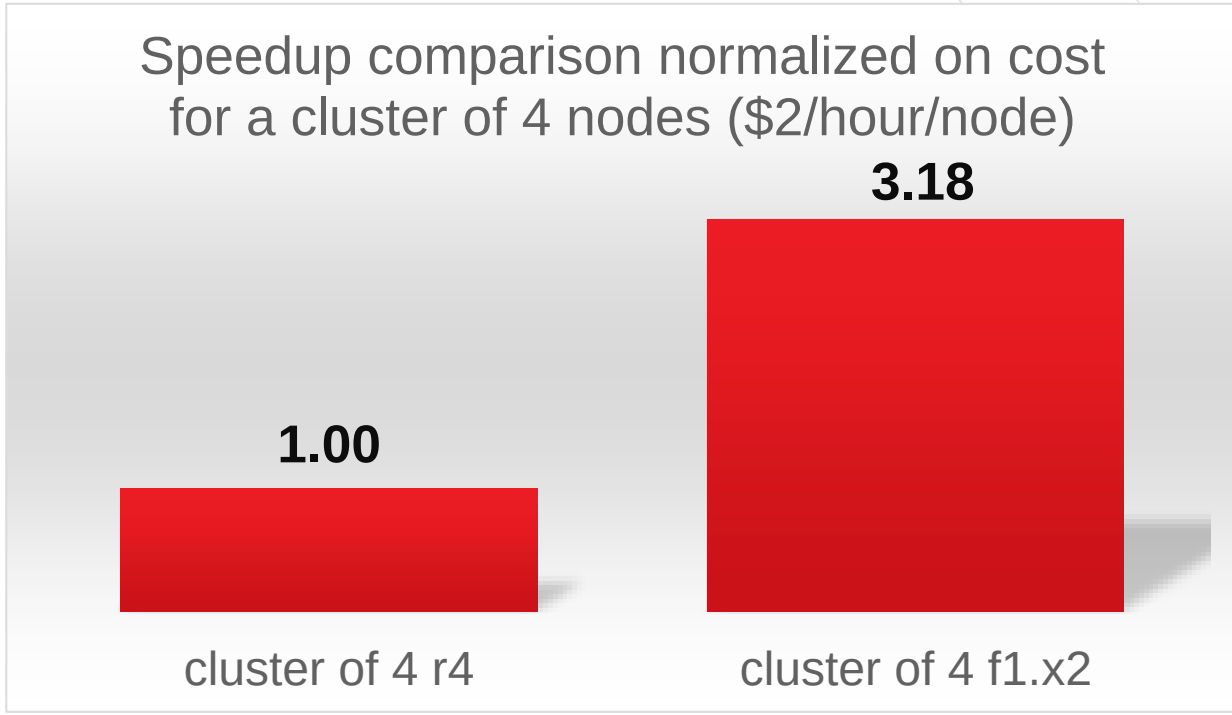


Speedup comparison

> 3x Speedup compared to r4

> 2x lower OpEx





Platforms

inaccel

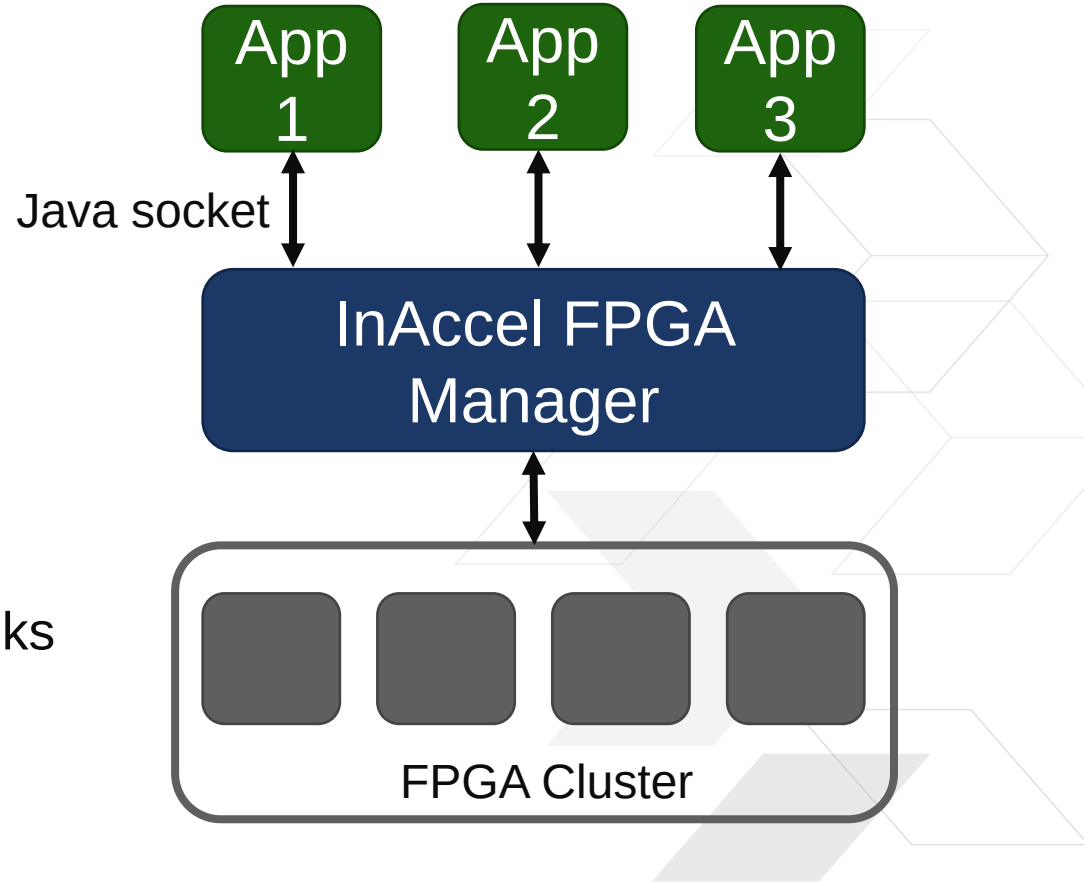
3x-10x Speedup

2x Lower Cost

Zero-code changes

InAccel's Coral FPGA Manager

High-level abstraction layer to utilize and manage an FPGA cluster



- **Resource Management**
  - Automatic configuration and management of the FPGA bitstreams and memory
- **Scheduling**
  - Automatic serialization and scheduling of the tasks send to the FPGA cluster
  - Scale to f1.x2, f1.x4, f1.x16 automatic
- **"Virtualization"**
  - Automatic serialization from multiple applications

FPGA Manager API

Memory Calls

> To make things easier we have incorporated a new **SharedMatrix** class that is basically backed up by a **Java ByteBuffer**.

Subclass	Used to store elements of type
SharedByteMatrix	byte
SharedDoubleMatrix	double
SharedFloatMatrix	float
SharedIntMatrx	int

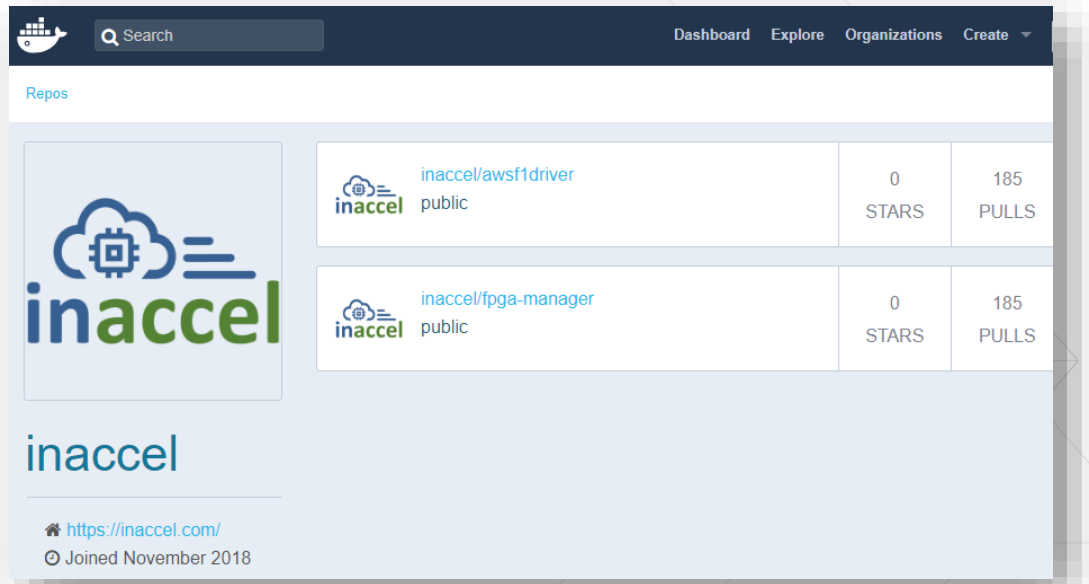
Request Calls

> Request calls are responsible for sending new tasks to the FPGA manager. All the requests are static methods of **InAccel class**.

Request	Used to accelerate:
Gradients32	Logistic Regression
Centroids32	KMeans
Black-Scholes	Black-Scholes

FPGA Manager deployment

- Easy deployment through dockers
- https://hub.docker.com/u/inaccel/
- Price for 3rd parties: $0.5/hour/node
- Free evaluation / limited features



inaccel

FPGA Manager

+

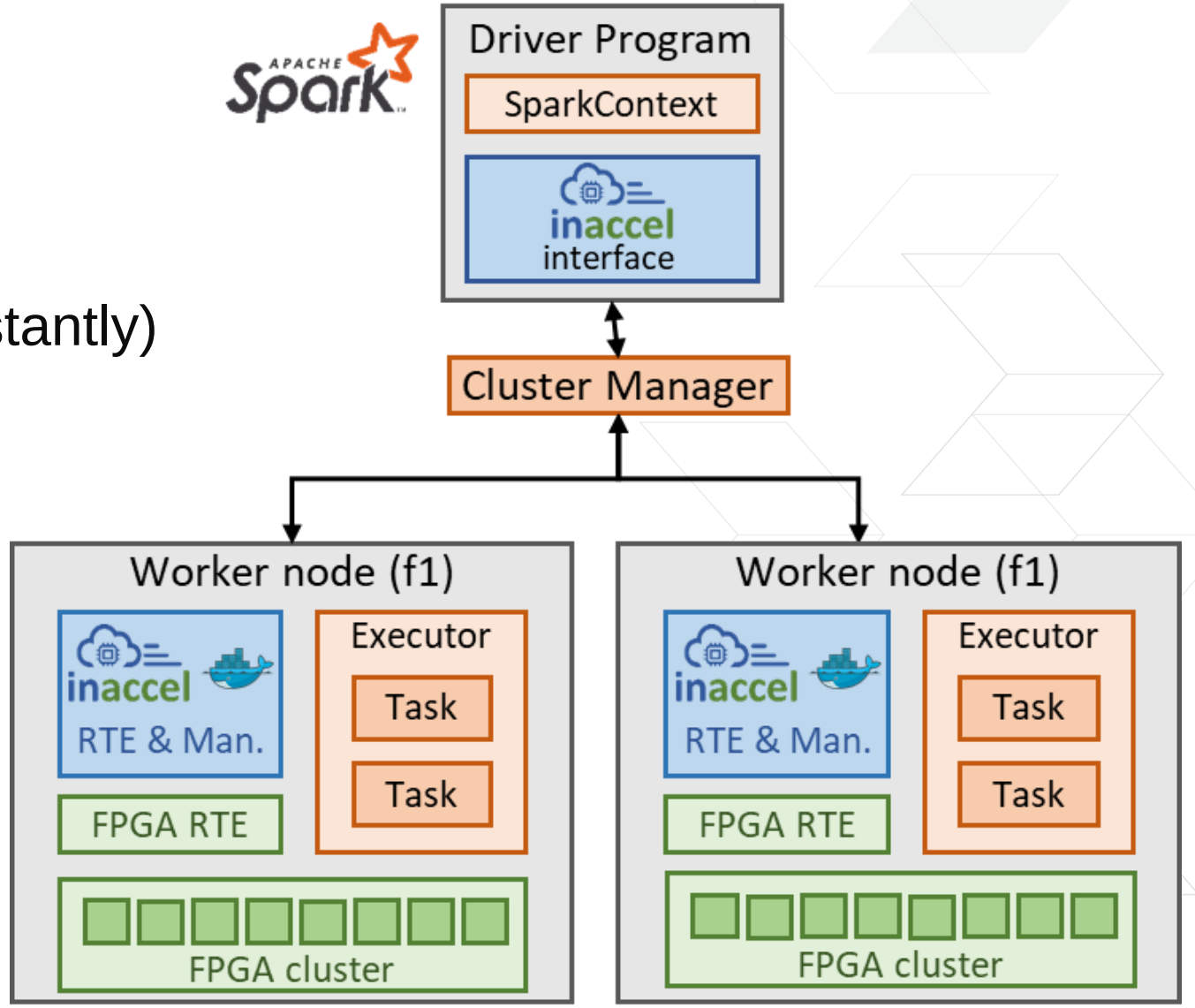
docker

=

- Easy deployment
- Easy scalability
- Easy integration

InAccel's Run-time Engine

- **Runtime engine that allows**
  - Scale Up (1, 2, or 8 FPGAs instantly)
  - Scale Out (using Spark API)
  - Seamless integration
  - Docker-based deployment



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.