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.

- Real-time Streaming: 51%
- Performance: 91%
- Ease of Deployment: 69%
- Advanced Analytics: 82%
- Ease of Programming: 76%

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 speed up their applications by providing ready-to-use accelerators-as-a-service in the cloud
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

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
Cloud Marketplace: available now

Customers

AWS Marketplace

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
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.inaccel.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

Cluster of 4 f1 (SW) vs. Cluster of 4 f1 (SW + InAccel)

- 4x f1.x2large (32 cores)
- 4x f1.x2large (32 cores + InAccel)

Speedup on cluster of f1.x2 using InAccel: 10.2x
Speed up

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

- f1.x16large (SW)
- f1.x16large (SW + 8 InAccel cores)

64 cores

64 cores + 8 FPGAs with InAccel

Speedup of f1.x16 with 8 InAccel FPGA kernels

1.00

12.14

f1.16xlarge (sw)
f1.16xlarge (hw)
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)

3.18

1.00

cluster of 4 r4  cluster of 4 f1.x2
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

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
**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.