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

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

% OF RESPONDENTS WHO CONSIDERED THE FEATURE
VERY IMPORTANT

More than one feature could be selected.

- 91% PERFORMANCE
- 69% EASE OF DEPLOYMENT
- 51% REAL-TIME STREAMING
- 82% ADVANCED ANALYTICS
- 76% EASE OF PROGRAMMING

Source: Databricks, Apache Spark Survey 2016, Report
helps companies speed up their applications by providing ready-to-use accelerators-as-a-service in the cloud

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

www.inaccel.com™, 2018
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

MLlib (machine learning)

Apache Spark

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

www.inaccel.com™, 2018
Accelerators for Spark ML in Amazon AWS in 3 steps

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

2. Import InAccel framework

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

www.inaccel.com™, 2018
Cloud Marketplace: available now

InAccel Products

AWS Marketplace

Amazon EC2 FPGA Deployment via Marketplace

Customers

Scalable to worldwide market

First to provide accelerators for Spark

www.inaccel.com™, 2018
IP cores available in Amazon AWS

Logistic Regression

K-mean 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

www.inaccel.com™, 2018
Communication with Host in Amazon AWS f1.x2 and f1.x16

Accelerators for logistic regression/kmeans

www.inaccel.com ™ , 2018
Performance evaluation

**Execution time for Logistic Regression (seconds) (MNIST 24GB, 500 iterations)**

- r5.x12 (48 cores)
- lnAccel (f1.x2)

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

- r5.x12 (48 cores)
- lnAccel (f1.x2)
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

www.inaccel.com™, 2018
Speedup comparison

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

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

ML Accel

10.2x

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

www.inaccel.com™, 2018
Speed up

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

12.14

f1.x16large (SW + 8 InAccel cores)

f1.x16large (SW)

64 cores + 8 FPGAs with InAccel

64 cores

MLAccel

MLAccel

MLAccel

MLAccel

MLAccel

MLAccel

MLAccel

Speedup of f1.x16 with 8 InAccel FPGA kernels

1.00

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

www.inaccel.com™, 2018
Speedup comparison

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

Cluster of 4 r4 (SW)
- Cluster of 4 r4 (SW + InAccel)
  - r4 (32 cores each – 128 cores total)
  - f1.x2large

Cluster of 4 f1 (SW + InAccel)
- Cluster of 4 f1 (SW + InAccel)

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

- cluster of 4 r4: 1.00
- cluster of 4 f1.x2: 3.18

www.inaccel.com TM, 2018
Platforms

Scalable Resource Manager

Inacel

3x-10x Speedup

2x Lower Cost

Zero-code changes

On-premise

Cloud

Platforms

Scalable
Resource Manager

Inacel

3x-10x Speedup

2x Lower Cost

Zero-code changes

On-premise

Cloud
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`.

Request Calls

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

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Used to store elements of type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharedByteMatrix</td>
<td>byte</td>
</tr>
<tr>
<td>SharedDoubleMatrix</td>
<td>double</td>
</tr>
<tr>
<td>SharedFloatMatrix</td>
<td>float</td>
</tr>
<tr>
<td>SharedIntMatrix</td>
<td>int</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Request</th>
<th>Used to accelerate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradients32</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Centroids32</td>
<td>KMeans</td>
</tr>
<tr>
<td>Black-Scholes</td>
<td>Black-Scholes</td>
</tr>
</tbody>
</table>
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

- 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

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

Single node ML suite
Distributed node ML suite

www.inaccel.com ™, 2018
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.
Adaptable.
Intelligent.
inaccel

XDF XILINX DEVELOPER FORUM

XILINX