Apache Spark MLLib Accelerator
Real-Time Big Data Analytics

XELERA

- **50x speed-up vs. CPU**
- **Fully integrated into Apache Spark framework**
- **Plug-and-play acceleration**

INTRODUCTION

The interest in developing real-time applications and advanced analytics is on the rise. True real-time processing of large data volumes, for example in real-time decision processes and live predictive analyses, is becoming an increasingly important feature of modern business applications. At the same time, more sophisticated data analytics algorithms are integrated in business applications so as to extract more accurate information from larger amounts of data.

Apache Spark is a popular Big Data framework used by a large number of companies across various industries. Users deploy Spark because it provides faster data processing than other Big Data frameworks and can easily scale to very large data sets. Secondly, Spark's machine learning library, MLLib, provides a rich collection of advanced analytics algorithms and tools.

Many components of MLLib are computationally demanding and require enormous processing time for larger data sets, which hampers real-time capabilities. The standard Spark deployment works on servers with multi-core CPUs. Xelera Analytics provides a software plugin to Apache Spark which offloads computationally expensive parts of Spark MLLib to FPGA accelerators, reducing the end-to-end compute time by 50x and enabling real-time capabilities.
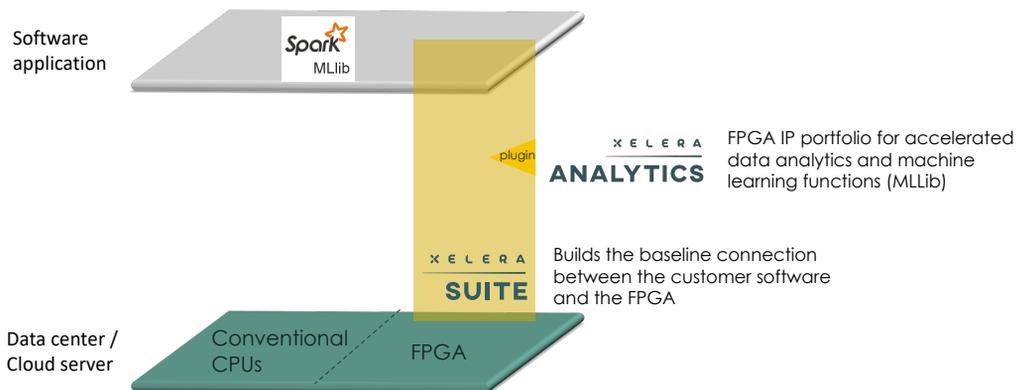
PRODUCT OVERVIEW

The Xelera Analytics software in this solution uses Xilinx Virtex Ultrascale+ FPGAs. The solution can be deployed in the public cloud (AWS, Nimbix, OTC) and on-premise using the Xilinx ® Alveo™ Data Center accelerator card.

SOLUTION OVERVIEW

Spark MLLib Accelerator

Xelera Suite (baseline middleware) and Xelera Analytics (plugin) enable acceleration for Apache Spark MLLib
- Fully integrated into Spark framework
- 50x speed-up vs. CPU
- Plug-and-play acceleration

Software application

Spark MLLib

plugin

XELERA ANALYTICS

FPGA IP portfolio for accelerated data analytics and machine learning functions (MLLib)

XELERA SUITE

Builds the baseline connection between the customer software and the FPGA

Data center / Cloud server

Conventional CPUs

FPGA

Apache Spark MLLib Example: Random Forest Algorithm
- Supervised machine learning
- Train with historical data to make predictions
- Ensemble of randomized decision trees
- Focus on accelerating training process
- Data setup (this benchmark):
  - 9 features
  - Each feature has 1.4M values
  - Ensemble of 10 trees

Implementation & Results
Xilinx Virtex Ultrascale+ FPGA
- Heavily vectorized (32 parallel entries)
- Deeply pipelined

Comparison
- Node by node
- Standard Apache Spark deployment on 24 CPU cores
  (48 threads)

Memory bank0	Memory bank1	Memory bank2	Memory bank3
FPGA RF core pipeline			

Implementation	Execution time per forest	Acceleration factor
FPGA-accelerated Spark MLLib	29.95 s	49.3
Standard Spark MLLib	1477.02 s	

CONCLUSION

With the Xelera Analytics software for Virtex Ultrascale+ FPGAs in the cloud and on the Xilinx ® Alveo™ Data Center accelerator card, customers can use Apache Spark MLLib in the usual way – without any knowledge about hardware accelerators – and benefit from 50x acceleration. This enables real-time business applications with advanced analytics.

TAKE THE NEXT STEP

Apart from Apache Spark, Xelera Analytics is integrated into several other software frameworks. Learn more at Xilinx Alveo Accelerator Card page and www.xelera.io/alveo

≨ XILINX®

✗ E L E R A