Solution Brief

Accelerating Quantitative Finance with FPGA-Based Acceleration Cards

Authors:

Ioannis Stamelos
Elias Koromilas
Chris Kachris
InAccel

Fred Tsang
Raj Subramani
Flumaion

Graham McKenzie
Natalia Poliakova
Intel

If you are responsible for building, testing, maintaining, or deploying trading/risk models utilizing financial algorithms:

- **As a business strategist or trader**: You will better understand how to apply the latest technologies for financial back-testing to successfully generate more profitable and/or less risky, robust trading models and increase competitive edge.

- **As a technology decision-maker**: You will learn how to incorporate a cost-effective risk analytics framework into your technology stack for back-testing financial models that simultaneously breaks through these historical bottlenecks or constraints:
  - Costs
  - I/O bandwidth
  - Complex, time-consuming computation
  - Time to market

Speedup your quantitative finance applications with the power of FPGAs using InAccel's orchestrator for seamless integration

Executive Summary

Quantitative finance is the use of mathematical models and extremely large datasets to analyze financial markets and securities. Common examples include the pricing of derivative securities such as options, and risk management, especially as it relates to portfolio management applications.

Quantitative finance are computationally intensive applications and the total execution time and the OpEx is mission-critical for many companies.

Hardware accelerators, like FPGAs, can offer significant performance but have historically lacked the required framework for easy integration with high-level frameworks, and easy deployment.

In this solution brief we describe how Flumaion utilized the power of the Intel® Programmable Acceleration Card (Intel® PAC) for faster execution of quantitative finance applications and the convenience of InAccel's orchestrator that allows easy integration and deployment of FPGA-based accelerators.

We show how InAccel's orchestrator allows **easy deployment, scaling, resource management, and task scheduling** for FPGAs making it easier than ever, the deployment and the utilization of FPGA for financial applications.

Industry Challenge in Quantitative Financial applications

Traditional financial workloads have huge data sets that need to be processed and pored through. An example workload is back-testing, which is a simulated trade on reliable historical data. It is a computationally intensive task because of the sheer volume of data.

For these high-frequency trading and risk management applications, the software demands are time and resource intensive. As a result, the hardware must keep up with the computing demands and constantly changing industry parameters. Hardware accelerators like FPGAs offer significant acceleration for these tasks resulting in faster execution time.

The Solution using FPGA-based Accelerators

Intel® FPGA-based acceleration platforms, also referred to as Intel PAC, provide FPGA solutions in PCI Express (PCIe) form factor that are verified and qualified by major server OEMs ready for integration into compute grids. High-level design languages such as OpenCL™, C++ and Data Parallel C++ (part of the Intel® OneAPI initiative) simplify the transition from software programming to FPGA programming, thus enabling acceleration with low latency of key workloads within the industry.

FPGA Acceleration Libraries (FAL) bring the programming experience another step closer to software development by providing highly optimized, common functions that are easily integrated with the user's code. The Intel® FPGA Acceleration Libraries include over 350 functions, featuring math, linear algebra, statistics, random numbers, date utilities and options pricing. There are also full-blown accelerators available along with pre-compiled FPGA binary files. These include Black and Scholes, Binomial pricing, Quadrature Integration, Partial Differential Equation solving and Monte Carlo.

InAccel PAC cluster manager

In cases that multiple applications or processes need to utilize the PAC-based accelerators, and the application needs to be deployed in multiple servers, InAccel® Coral manager is used to abstract away the available resources and provide a simple API for software developers. InAccel Coral manager abstracts away the available resources in a cluster of PAC cards, making it easier than ever to deploy one or more applications targeting multiple FPGAs.

InAccel's manager is used to schedule, dispatch and manage the functions that need to be accelerated. It performs the load balancing among the available resources in the PAC cluster and is also used for the management and configuration of the cards based on the functions that are offloaded.

Software developers can simply call the functions that need to be accelerated without worrying on the scheduling of the functions to the available resources or the contention of the resources from multiple applications.

InAccel's orchestrator is fully compatible with Intel® FAL, Quantitative Finance accelerators and the Intel® PAC cards.

By using InAccel framework, software developers can benefit from the performance offered by PAC accelerators for financial applications without any additional overhead. Also, it allows the seamless sharing of the resources from multiple applications.

InAccel's unique repository of FPGA precompiled bitstreams using Intel® library functions allows software developers to utilize the financial accelerators on the PAC cards in the same way as any software library. InAccel's repository includes all the required functions provided by Intel as ready-to-use host software libraries.

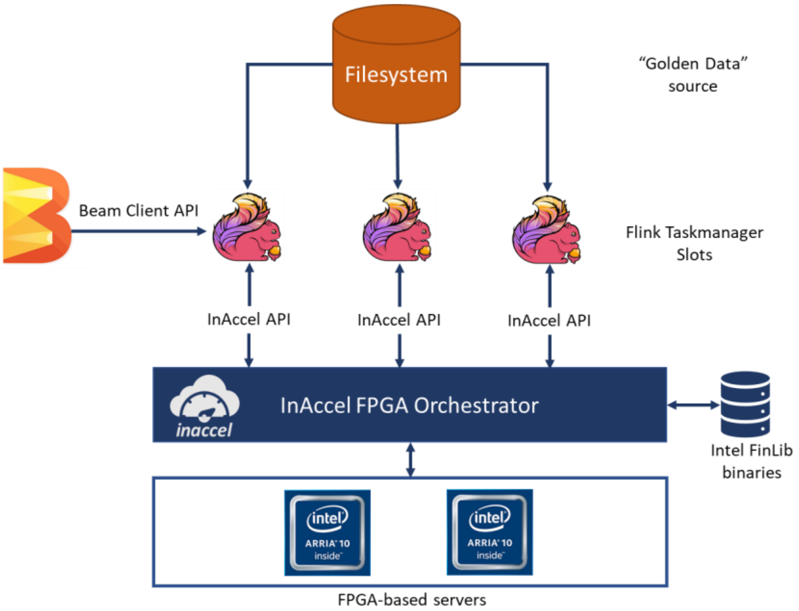


Figure 1. Flumaion testbench for accelerating quantitative finance with Intel® FPGA Acceleration Libraries on PAC-based server using InAccel orchestrator

Testing on Quantitative finance

A typical analytics workflow involves reading data from a golden source, running the calculations on a grid cluster and then persisting the results. Data storage resources are usually separated from the compute grid to allow to horizontally scale and at different rates. Thus, data serialization and its distribution are an important consideration.

Apache Beam is an open source API based on technology that has been evolved from Google's Map Reduce. Beam is designed for Big Data processing across large distributed grids. The Beam API is used to describe the workflow pipeline which processes the pricing requests using Apache Flink server as the implementation.

The test itself involves reading ten text files each containing 3.4 million rows of equity option trades in JSon format.

The pipeline then calls C++ which reads and transforms the JSon into a matrices type that the FPGA understands. The FPGA kernel is then called with the built matrix. The output is then serialized to disk and read back by Beam.

As noted previously, the InAccel framework abstracts the FPGA resources by providing scheduling and task management. The client merely makes calls through the InAccel API without considering what underlying FPGA resources are available. When multiple clients make API requests in parallel, InAccel internally buffers these calls and schedules them onto the FPGA(s).

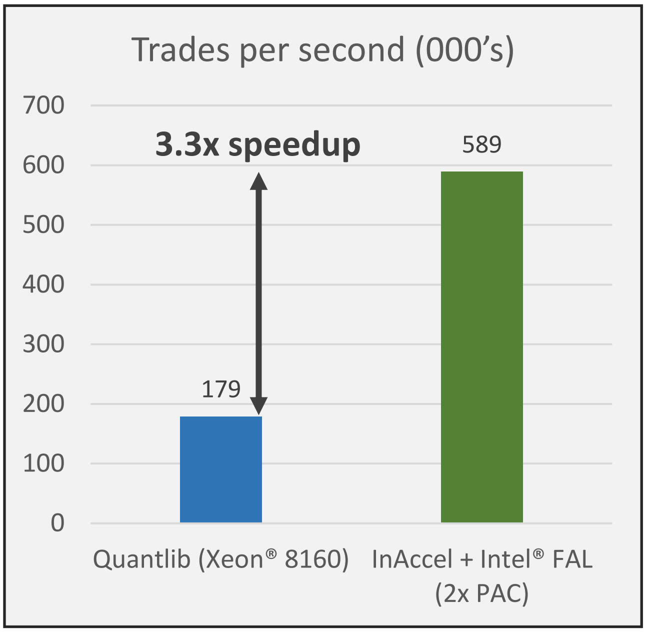
This greatly simplifies parallelization as illustrated in the above diagram. Multiple Flink Task-manager slots can call the InAccel API as work is allocated to them without the need to block calls to the FPGA.

> *"Inaccel framework gives you a great abstraction layer resulting in less code. Also it has monitoring tools and other language bindings like Python, etc. making it very easy to integrate with our applications"*
>
> *"…this is where InAccel's framework helps immensely: we don't need to code MT C++, rather we call the FPGA from multiple workers and we let the Inaccel framework schedule the FPGA workloads as fast as possible."*
>
> Fred Tsang, Flumaion

When Flink processes are placed in Docker containers and managed by tools like Kubernetes, one creates a grid that can automatically scale by the data throughput or CPU usage.

Using Intel® FPGA Acceleration Libraries, Intel® PACs and InAccel's orchestrator Flumaion can benefit from the significant speedup FPGAs can offer and the seamless integration with higher-level programming frameworks.

In a simple case where, Black & Scholes with Greeks has been deployed, Intel® FPGA Acceleration Libraries and the InAccel framework can provide as much as 3.3X speedup compared to dual Intel® Xeon® Platinum 8160M processors (24 cores each) running Quantilib.



InAccel helps companies' speedup their applications, with zero code changes using efficiently state-of-the-art accelerators. InAccel provides a unique technology that allows the easy deployment, management, scaling and virtualization of FPGA-based accelerators. InAccel's FPGA orchestrator allows instant deployment and scaling of accelerator for widely-used applications like quantitative finance, big data analytics and machine learning.

Learn more : https://inaccel.com

Flumaion

Flumaion provide expertise in Risk Analytics including mathematical modelling and coding (C++) to high performance grids for Risk and P&L valuations, including real time metrics. We are experienced in Java, Python, C++ and Javascript. Flumaion has expertise in Electronic trading (pre and post trade) as well as Distributed Ledger Technology including International Swaps and Derivatives Association (ISDA) Common Domain Model (CDM) definitions, design, implementation and integration. Flumaion are experts in Big Data Technology in the Cloud (Google BigQuery) as well as on-premise (Hadoop clusters).

Learn more: https://www.flumaion.com/

Intel® technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure.
Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors.

The configuration used for both the Quantlib results and the Intel® FAL/Inaccel solution comprises Dual Intel® Xeon® Platinum 8160M processors, 384 GB RAM at 2666MHz, Optane™ P4800x 375GB storage and 2 Intel® PAC with Intel Arria® 10 GX FPGA cards.