White Paper

Financial Services
Derivatives Pricing

Intel and Quantifi Accelerate Derivative Valuations by 700x Using AI on Intel Processors

Intel and Quantifi have demonstrated that accurate, real-time pricing for a portfolio of derivatives can be generated locally or in the cloud using AI technology running on 3rd Generation Intel® Xeon® Scalable processors—no special hardware required

Table of Contents

Executive Summary

Portfolio managers and traders that use over the counter (OTC) derivatives often lack an accurate real-time view of the valuations and risk of their derivative positions, especially when trading exotic derivatives. Unlike liquid securities or exchange traded products, there is not always a market price available for OTC derivatives. These products therefore need to be valued according to models that accurately calculate their theoretical fair value. Obtaining real-time risk metrics for a portfolio of derivatives has been challenging, as the commonly used valuation techniques for these products are computationally expensive and require significant machine time. Portfolio valuations and risk calculations typically require overnight runs in a data center or the cloud.

This whitepaper reports the successful use of Artificial Neural Network models (ANNs) by Quantifi to model and deliver real-time pricing with an accuracy considered equivalent to conventional approaches such as numerical integration and Monte Carlo methods, which will be referred to as the conventional model in this whitepaper.

Comparative testing shows that both the conventional and Quantifi ANN models exhibit a less than 0.01% deviation when compared to the theoretical fair value for the derivative.[1] A deviation of less than 0.01% lies well within the bid-offer spread for credit options, which is the economically relevant comparison for a credit option trader.[2] The ANN model is also orders-of-magnitude faster and can deliver real-time valuations. These accuracy and order-of-magnitude performance improvements are consistent with those observed when replacing Monte Carlo methods with ANNs in scientific fields such as High Energy Physics (HEP).[3]

The Intel benchmarks show that the recently launched 3rd Generation Intel Xeon Scalable processors run the Quantifi ANN model 1.56x faster than previous 2nd Generation Intel Xeon Scalable processors—in part due to their greater core count. Overall, these new processors can generate accurate valuations 700x faster than the conventional model, which is sufficient to provide real-time results for common valuation workloads.

To establish that traders can receive fair value pricing in real-time without requiring specialized computational hardware, Quantifi partnered with Intel to evaluate the performance of their AI technology on CPU-based servers. As will be discussed in this whitepaper, the Intel benchmarks show that the recently launched 3rd Generation Intel Xeon Scalable processors, run the Quantifi ANN model 1.56x faster than previous 2nd Generation Intel Xeon Scalable processors[4]—in part due to their greater core count. Overall, these new processors can generate accurate valuations 700x faster than the conventional model,[5] which paves the path toward real-time results for common valuation workloads.[6]

> *The Intel benchmarks show that the recently launched* **3rd Generation Intel Xeon Scalable processors** *run the Quantifi ANN model 1.56x faster than previous 2nd Generation Intel Xeon Scalable processors—in part due to their greater core count. Overall, these new processors can generate accurate valuations 700x faster than the conventional model, which paves the path toward real-time results for common valuation workloads.*

Orders-of-magnitude faster performance means that financial institutions can receive an accurate real-time view of their portfolio valuations without requiring specialized hardware. Instead, traders can quickly evaluate the risk in their derivative positions utilizing servers running 3rd Generation Intel Xeon Scalable processors in their local data centers, private cloud, or externally via a public cloud provider.

A New Paradigm: Managing OTC Derivative Risk with Real-time Pricing

This research into using AI as a replacement for conventional methods is transformative as it changes how risk is managed. Currently, traders often have to rely on risk metrics that are calculated in an overnight batch calculation based on the positions and market data obtained at the end of the previous day. These batch calculations can take anywhere between minutes and hours depending on the complexity of the portfolio and available hardware. During the day, traders then selectively request ad hoc risk calculations or try to approximate their current risk based on yesterday risks and changes in market data and positions. The Quantifi AI model introduces a new paradigm for financial management where portfolio managers can see a comprehensive set of real-time risk metrics on their whole portfolio.

> *"The switch to AI technology has the potential to transform the industry."*
>
> —Sebastian Hahn, AI and ML Lead at Quantifi

These accelerated derivative valuations are also important for electronic market making desks that need to automate trading decisions in fractions of seconds and monitor risks in real-time.

Most risk metrics that traders are interested in, such as scenarios and sensitivities, can be thought of as valuing derivatives with different sets of hypothetical market data. Real-time valuations will therefore accelerate risk calculations across a wide range of metrics.

While this whitepaper focuses on valuations, it should be realized that the results reported here have much wider implications for risk management. An example of a risk metric that is particularly computationally demanding are XVA calculations for banks, which are usually performed via Monte Carlo simulation. Simulating a portfolio of 40,000 trades across 80 time steps on 2,000 paths, for example, produces more than 6 billion valuations as discussed in the Quantifi whitepaper How to Accelerate XVA Performance. As the number of valuations increases to these magnitudes fast models become of utmost importance.

What are the Conventional Methods for Valuing Derivatives?

Derivatives are valued using no-arbitrage pricing theory. The price of a derivative should be equal to the value of a replicating portfolio of simpler assets. When devising a valuation model based on no-arbitrage pricing theory, researchers usually need to make assumptions about the stochastic processes that drive the price of these simpler assets. Numerical integration and Monte Carlo simulations are frequently used to solve the resulting valuation problem. This approach has produced a wide range of sophisticated and incredibly useful valuation models.

How can AI be used to Accelerate Valuations?

Quantifi has taken an approach in developing the ANN model that is agnostic towards the theoretical foundations of the valuation function. As such, it is entirely consistent with mainstream no-arbitrage pricing theory. This approach asks the question: given a valuation function that can compute the fair value of a derivative based on a range of inputs, how can one speed up the computation?

To answer this question, an Artificial Neural Network (ANN) was trained to approximate a known valuation function using data from one of Quantifi's proprietary models that uses numerical integration to compute derivative prices. The ANN was then trained until it was able to value derivatives with an accuracy that is economically equivalent to the conventional model.

This approach differs from previous studies in the literature that have evaluated the potential of machine learning to speed up derivative valuations for equity products such as equity basket options. At this time, very little research seems to have focused on using ANNs to speed derivative valuations on credit derivatives. In this research Quantifi focused on credit index options

Transforming a conventional model into an artificial neural network

It has been known since 1987 that ANNs can approximate very complex multi-dimensional functions[7] by essentially fitting a multidimensional surface to the training data.[8] While the training process is computationally expensive and requires large amounts of data to describe all the important points of inflection on this surface accurately for a complex problem, the inference operation tends to be very fast as the trained ANN simply needs to calculate a single point on the surface for one set of input values.[9]

Recent work by CERN in 2018 demonstrated that orders of magnitude speedups are possible when using trained ANN models to replace Monte Carlo methods in HEP models.[10] The emergence of powerful hardware has further reduced the cost of training and broadened the applicability of ANNs in a variety of scientific and commercial fields for a wide range of extremely complex applications including computer vision, natural language processing and speech recognition. Quantifi has leveraged these AI and computer technology advances to benefit their customers.

In order to train the ANN to approximate a known derivative valuation function, the Quantifi team generated a supervised learning dataset for ANN training that maps input values such as market data and contractual information of the derivative to the desired output – in this case the fair value of the derivative. Specific inputs provided to the ANN to value credit index options include the index level, interest rates, the volatility, the number of names that have already defaulted, the strike price and maturity of the option among others.

Quantifi has the ability to generate arbitrarily large training sets through the use of their proprietary valuation model for credit options which is able to determine a fair value for each input vector. This means Quantifi can simulate a realistic range of input values over which the ANN has to be accurate, and calculate prices based on the known valuation function, to generate an appropriate training data set. A variety of statistical distributions were utilized to generate the results in this whitepaper including uniform, normal, and beta distributions which were based on a 10 million example training set.

A standard ANN training workflow was used to train the ANN. Training was stopped after the ANN model achieved a 0.01% deviation when compared to the theoretical fair value for the derivative. The end result is a trained ANN that represents a function f(x) = y, that maps contract and market data (x) to prices (y). The accuracy of this function can then be evaluated on synthetic and real market data.

Internally, Quantifi determined that the accuracy of the trained ANN model is equivalent to the conventional model —meaning they both exhibit a deviation less than 0.01% compared to the theoretical fair value. This was deemed sufficient as a deviation of less than 0.01% lies well within the bid-offer spread for credit options, which is the economically relevant comparison for a credit option trader.

Quantifi partnered with Intel to determine the usefulness of this trained ANN model for traders in the field, who typically use institutional data centers and cloud based computing resources.

Intel Benchmarks: Runtime Performance of the Trained ANN Model

To evaluate the efficacy of the Quantifi software utilizing the trained ANN model in the data center, Intel ran a series of benchmarks to:

1. Quantify the performance improvement of the AI model over the legacy model on Intel Xeon Scalable processors
2. Compare the performance between the 2nd and 3rd Generation Intel Xeon Scalable processors
3. Profile the code to see the impact of Intel AVX-512 vector instructions

Overall, the benchmark results clearly show that traders can use the Quantifi software with the AI model to get approximately a 700x speedup to effectively receive their results in real-time on CPU-based servers running in institutional, private cloud, and public data centers. An accelerator is not required.

Production Potential Using the Quantifi Compute Engine Architecture

All of Quantifi's proprietary models that lie at the foundation of its trading, risk and analytics solutions are exposed via programming APIs. Quantifi's Python API can be used to combine Quantifi's proprietary models with standard data science and machine learning libraries such as Pandas and TensorFlow, which was used in this research. The production use will fit into the Quantifi Compute Engine architecture illustrated in Figure 1 below.

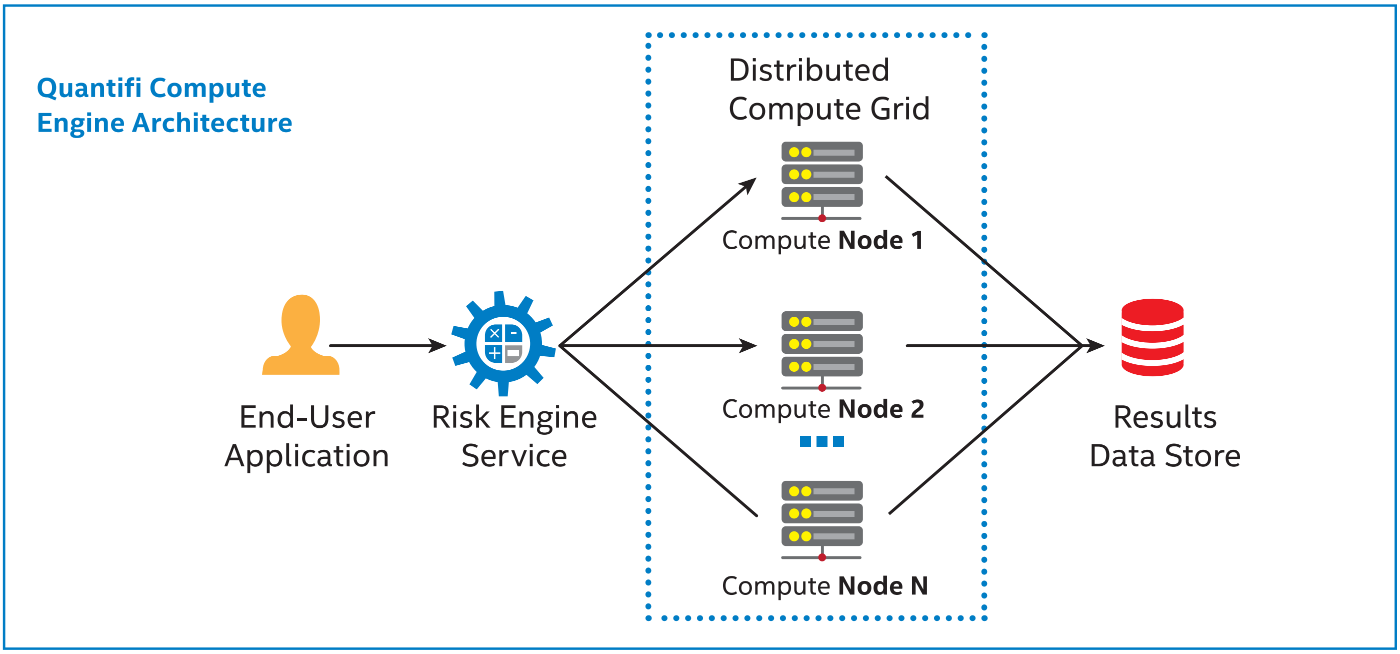

Figure 1: The Quantifi compute engine architecture. (Source Quantifi)

System choice and configuration information

To generate comparative performance data, Intel utilized two servers: one running the recently announced 3rd Gen Intel Xeon Scalable processors[11] and one running 2nd Gen Intel Xeon Scalable processors.

Both the 2nd and 3rd Gen Intel Xeon Scalable processors have been designed by Intel to deliver high performance on floating-point and AI workloads in the data center and the cloud. For this reason, both processor Generations are able to run both the conventional Quantifi and trained ANN models with high performance. This choice of processor architectures also reflect the performance traders should see in the field when running on Intel-based servers in institutional and cloud data centers.

The configuration details are shown below:

- A 2nd Gen Intel Xeon Scalable processor-based server: This server utilizes two 2.7 GHz Intel Xeon Platinum 8280 processors (28 cores/processor) with a total of 56 cores (112 threads), 384 GB memory, Intel TensorFlow 2.4.0 and Python 3.7.9
- A 3rd Gen Intel Xeon Scalable processor-based server: This server utilizes two 2.3 GHz Intel Xeon Platinum 8380 processors (40 cores/processor) with a total of 80 cores (160 threads), 512 GB memory, Intel TensorFlow 2.4.0 and Python 3.7.9

The AI Model Delivers a 700x Increase in Throughout on the Quantifi Credit Option Pricing Benchmark

To compare the performance of the new AI model against the conventional model, Intel utilized the Quantifi Credit Option Pricing Benchmark. The performance results on the 3rd Gen Intel Xeon Scalable system shown in Figure 2 below. They demonstrate an approximate 700x increase in inference throughput using the AI model compared to the Quantifi conventional model.

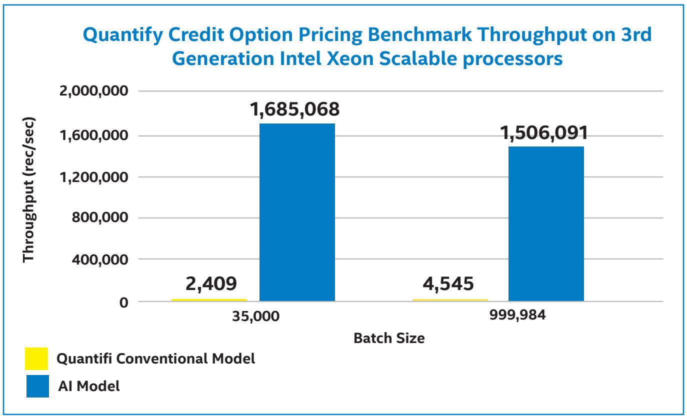

Figure 2: Measured performance of the Quantifi conventional model vs AI model on a 3rd Generation Intel Xeon Scalable processor. Higher is better. (Source Intel)

The throughput of the AI model was measured using one instance of the Quantifi software pinned to each of the two sockets present in the system. The sum of the throughput of the two instances gives the total throughput of the AI model on the system. The benchmark reported throughput in terms of number of records per second. The same process was followed for batch sizes of 35,000 and 999,984 positions.

These two batch sizes were chosen to reflect two use cases:

- A "typical portfolio" of a small bank or a large hedge fund may be around 35,000 positions. To value every position once as a snapshot in time and know what the portfolio is worth in this moment, one would need to calculate 35,000 valuations.
- There are some other calculations where the models might be used that require a much higher number of valuations. This includes valuation adjustments (XVA), which involve pricing the same portfolio at multiple time steps and multiple possible future paths. For those calculations the number of valuations is millions or even billions. For the benchmark for this whitepaper, a value of 1M was chosen.

Improved AI performance on 3rd Generation Intel Xeon Scalable processors

Figure 3 below reports the 1.56x performance improvement that the 3rd Generation Intel Xeon Scalable processors was able to deliver when running the Quantifi AI model as compared to the 2nd Generation processor.

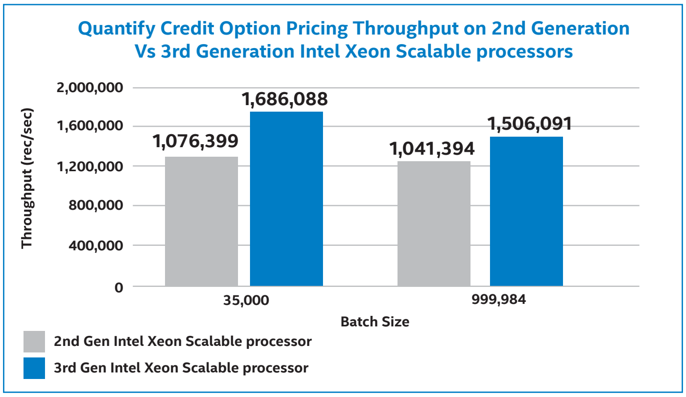

Figure 3: Intel 3rd Generation Intel Xeon Scalable processors deliver a 1.56x increase in the runtime performance of the Quantifi AI model (Source Intel)

For this benchmark, one instance of the Quantifi AI model is pinned to each of two Intel Xeon Platinum 8380 processors and the performance is compared when the same model is pinned to each of the two Intel Xeon Platinum 8280 processors. The greater number of cores on the 3rd Gen Intel Xeon Scalable processors mean they can deliver higher floating-point performance. However the ratio of cores (80/56 or 1.46) is not sufficient to account for all of the 1.56x performance increase, which indicates that other architectural improvements (such as increased cache sizes and memory bandwidth) in the Intel Xeon Platinum 8380 processors are also working to speed this AI workload.

Instruction profiling shows heavy AVX-512 utilization

As expected, the Quantifi ANN model is floating-point intensive. Figure 4 reports the runtime instruction profile of the Quantifi code. This profile clearly shows that the

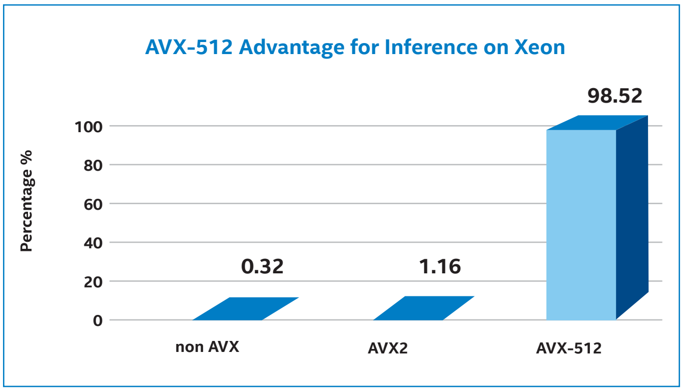

Figure 4: Runtime instruction profile of the Quantifi model when utilizing the new AI model (Source Intel)

ANN model is floating-point intensive with over 98% of the runtime spent performing Intel AVX-512 floating-point operations. Both the 2nd and 3rd Gen Intel Xeon Scalable processor families provide dual per core Intel AVX-512 vector units to deliver high performance on such floating-point intensive workloads. The benchmark shows that the Quantifi software is able to benefit from the greater number of cores on the 3rd Gen Intel Xeon Scalable processors.

> *"The current 3rd Generation Intel Xeon Scalable processors delivers the performance needed to produce real-time results using the Quantifi ANN model in the data center and cloud without requiring an accelerator."*
>
> —Mahesh Bhat, Principal Engineer at Intel

About Quantifi

Quantifi is a provider of risk, analytics and trading solutions. Our award-winning suite of integrated pre and post-trade solutions allow market participants to better value, trade and risk manage their exposures and respond more effectively to changing market conditions. Quantifi is trusted by the world's most sophisticated financial institutions, including five of the six largest global banks, two of the three largest asset managers, leading hedge funds, insurance companies, pension funds, and other financial institutions across 40 countries. Renowned for our client focus, depth of experience, and commitment to innovation, Quantifi is consistently first-to-market with intuitive, award-winning solutions.

Conclusions

This whitepaper reports the successful use of ANNs by Quantifi to model and deliver real-time pricing with an accuracy that exhibits a less than 0.01% deviation when compared to the theoretical fair value for the derivative. This real-time performance was obtained without the use of accelerators, which means portfolio managers and traders can get real-time reports from CPU-based servers running locally or in the public cloud.

[1] Source Quantifi

[2] Source Quantifi.

[3] Using ANNs to replace Monte Carlo methods has recently become an accepted practice by the scientific community. CERN, for example, is using ANNs to realize orders-of-magnitude speedups in some of their higher energy particle simulations https://www.hpcwire.com/2018/08/14/cern-incorporates-ai-into-physics-based-simulations/

[4] Results reported in Figure 3, 1.56x higher performance on gen 3 Intel Xeon Scalable Processors: Gen-3: 1-node, 2x 3rd Gen Intel Xeon Platinum 8380 on Intel Reference Platform (Coyote Pass) with 512 GB (16 slots / 32 GB / 3200) total memory, BIOS: SE5C6200.86B.0022.D08.2103221623, HT on, Turbo on, with CentOS Linux Version 8, 4.18.0-240.15.1.el8_3.x86_64 1x 480GB SSD boot drive, 2x3.2T P4610 data, 1Gbps NIC, Quantifi Credit Option Pricing AI Inference 1.0, Intel TensorFlow 2.4.0 and Python 3.7.9, test by Intel on 03/04/2021. Baseline: 1-node, 2x 2nd Gen Intel Xeon Platinum 8280 on Intel Reference Platform (Wolf Pass) with 384 GB (12 slots / 32 GB / 2933) total memory, BIOS: SE5C620.86B.0D.01.0395.022720191340, HT on, Turbo on, with Red Hat Enterprise Linux Server release 7.9 (Maipo), 3.10.0-1160.15.2.el7.x86_64 , 1x 480GB SSD boot drive, 2x3.2T P4610 data, 1Gbps NIC, Quantifi Credit Option Pricing AI Inference 1.0, Intel Tensor Flow 2.4.0 and Python 3.7.9, test by Intel on 03/04/2021.

[5] Results reported in Figure 2, 700x higher AI performance, gen 3 Intel Xeon Scalable Processors: Gen-3: 1-node, 2x 3rd Gen Intel Xeon Platinum 8380 on Intel Reference Platform (Coyote Pass) with 512 GB (16 slots / 32 GB / 3200) total memory, BIOS: SE5C6200.86B.0022.D08.2103221623, HT on, Turbo on, CentOS Linux Version 8, 4.18.0-240.15.1.el8_3.x86_64, 1x 480GB SSD boot drive, 2x3.2T P4610 data, 1Gbps NIC, Quantifi Credit Option Pricing AI Inference 1.0, Intel TensorFlow 2.4.0 and Python 3.7.9, test by Intel on 03/04/2021.

[6] Source: Quantifi.

[7] https://www.osti.gov/biblio/5470451-nonlinear-signal-processing-using-neural-networks-prediction-system-modelling

[8] https://papers.nips.cc/paper/1987/file/093f65e080a295f8076b1c5722a46aa2-Paper.pdf

[9] https://www.osti.gov/biblio/5470451-nonlinear-signal-processing-using-neural-networks-prediction-system-modelling

[10] https://openlab.cern/sites/openlab.web.cern.ch/files/2018-06/Vallecorsa_poster.pdf

[11] https://www.intel.com/content/www/us/en/newsroom/news/3rd-gen-xeon-scalable-processors.html#gs.11y76w

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.