Transforming Quantitative Finance with Optimized AI Software and the Latest Intel® Xeon® Scalable Processors

Access real-time valuations and scenario analyses by running Riskfuel's groundbreaking AI software on 5th Gen Intel® Xeon® Scalable processors

Solution Benefits

- **Faster, more accurate deep-learning models** with deep neural networks.
- **Reduced run time** by using built-in hardware acceleration.
- **Extremely low computation cost** by reducing the number of servers necessary to run calculations.
- **Real-time view of valuations and risk** for a more accurate, agile trading strategy.
- **Lower operational carbon footprint** by consolidating workloads on fewer servers.

Riskfuel

Executive Summary

Financial institutions perform quantitative finance to better understand pricing trends and measure risk. Deep learning techniques have helped transform the quantitative finance field by enabling far more data to be analyzed than was possible using traditional mathematical models. However, even with deep learning, calculations can take too long to complete to keep pace with volatile markets.

All that changed when AI software vendor Riskfuel and Intel decided to collaborate. Riskfuel software harnesses AI to produce fast versions of the proprietary software that banks and insurance companies use to value and risk-manage their portfolios of financial instruments. **Customers experienced a 13x improvement in throughput (valuations per second) when using Riskfuel.**[1]

Application acceleration is even higher when customers combine Riskfuel with 4th or 5th Generation Intel® Xeon® Scalable processors. These processors offer Intel® Advanced Matrix Extensions (Intel® AMX), which is a built-in accelerator that improves the performance of deep-learning training and inference on the CPU. **Refreshing the hardware to 5th Gen Intel Xeon Scalable processors resulted in up to an additional 28% more valuations per second compared to 4th Gen Intel Xeon Scalable processors**[2] **and surpassed NVIDIA A100 GPU performance for batch sizes that are less than 1,000.**[3]

Running Riskfuel on Intel® processors with Intel AMX, customers can accelerate their deep-learning models, increase model accuracy, and reduce run time while cutting computation costs to virtually zero.[4] At long last, traders can perform on-demand recalculation of portfolio values and gain an up-to-the-second view of risk without purchasing expensive, alternative compute architectures.

Business Challenge

Quantitative finance involves the pricing of derivative securities, such as options, as well as risk management for a trading portfolio. For example, a trading company uses Finite Differences or Monte Carlo simulations to assess the risk associated with a particular trade. Historically, these simulations—which are extremely compute-intensive—are run only periodically because they take hours to complete and consume substantial compute resources, driving up compute costs.

However, relying on a simulation that is hours (or perhaps even days) out of sync with the current market reality can result in less-than-ideal investment decisions. Quickly changing market factors may have increased the risk associated with a particular transaction during the time it takes to run the simulation, requiring an adjustment to trading strategy. But with an out-of-date simulation, a trader may not make that adjustment. An additional concern for financial institutions is regional regulations that call for an increased frequency of risk assessment. Complying with such regulations could drive up operational costs.

If only there were a way to run quantitative finance calculations in real time.

Solution Value: Optimized Software Empowers Real-time Valuations and Risk Analysis

Riskfuel and Intel engineers optimized Riskfuel's AI software to take advantage of Intel AMX. Therefore, running Riskfuel on servers equipped with a recent generation of Intel® Xeon® Scalable processors provides customers with a trading advantage. Plus, finance workloads can run lightning-fast on the same hardware that handles other enterprise workloads, which helps decrease total cost of ownership and can potentially reduce a trading company's operational carbon footprint.

Running Riskfuel on Intel® architecture enables traders to calculate valuations and risk sensitivities in real time throughout the trading day. If something in the market changes, they can immediately adapt. Even as the sun sets, they will have the latest data.

Riskfuel uses machine learning to quickly make a fast copy of the in-house model. From the outside, the Riskfuel model looks like the in-house model because it takes in the same input parameters and produces the same results. Internally, however, the complex simulation has been replaced by very fast deep neural network (DNN) inferencing. The calculations are performed in a small fraction of the time at a low compute cost. What once used to take a data center full of servers all night long to calculate can now be achieved with a handful of servers in minutes.[5] When customers used Riskfuel, they achieved up to a 13x increase in valuations per second in their simulations compared to not using Riskfuel.[6]


13x Increase in valuations per second generated in simulations running on Riskfuel[1]

Further testing revealed that after Riskfuel software was optimized to take advantage of Intel AMX, 5th Gen Intel Xeon Scalable processors produced up to 28% more valuations per second compared to 4th Gen Intel Xeon Scalable processors (see Figure 1).[7] Furthermore, 5th Gen Intel Xeon Scalable processor performance substantially surpassed an NVIDIA A100 GPU's performance for workloads with a batch size of less than 1,000.[8]

Latency Can Be More Important than Throughput

If a trader can afford to wait nearly a half-second or more for results, large batch sizes—such as 5,000—are probably acceptable. However, sometimes it's more valuable to have computational results in milliseconds, rather than tenths of a second. In these real-time valuation scenarios, a batch size of 1,000 or less is realistic. For example, consider a scenario where a trader has 50 trades with exposure to a certain company. If that company's stock price is extremely volatile, the trader needs constant revaluation to lower risk. In another scenario, 4th and 5th Gen Intel Xeon Scalable processors can deliver prices every millisecond. According to our testing, this kind of latency simply isn't possible with the NVIDIA A100 GPU.[9]
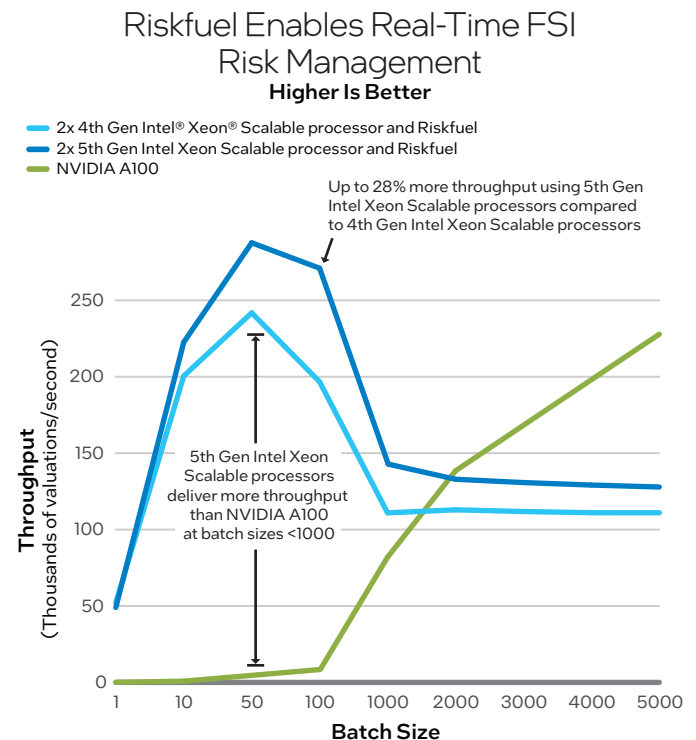
Figure 1. Riskfuel running on recent generations of Intel® Xeon® Scalable processors enables real-time risk management—with nearly identical or even better performance than provided by the NVIDIA A100 GPU.[8]
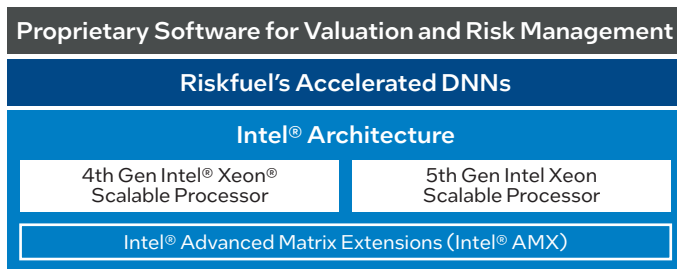
Solution Architecture: Riskfuel on Intel® Xeon® Scalable Processors

Riskfuel models learn an interpolation between millions of data points using a DNN. The Riskfuel solution generates these data points by repeatedly running the target model on different input combinations covering the domain of approximation. This strategy of learning the relationship between inputs and outputs makes it extremely fast.

Intel AMX is a dedicated hardware block on the Intel Xeon Scalable processor core that helps optimize and accelerate deep-learning training and inferencing workloads that rely on matrix math. Its architecture supports BF16 (training and inference) and INT8 (inference only) data types. Two main components of Intel AMX are responsible for its acceleration capabilities:

- **Tiles.** These consist of eight two-dimensional, 1 KB registers that store large chunks of data.
- **Tile Matrix Multiplication (TMUL).** TMUL is an accelerator engine attached to the tiles that performs matrix-multiply computations for AI.

These two components work together to enable Intel AMX to store more data in each core and compute larger matrices in a single operation.[10] Riskfuel and Intel engineers worked together to optimize Riskfuel software to take advantage of Intel AMX (see Figure 2) to further speed up financial model performance.

Proprietary Software for Valuation and Risk Management	
Riskfuel's Accelerated DNNs	
Intel® Architecture	
4th Gen Intel® Xeon® Scalable Processor	5th Gen Intel Xeon Scalable Processor
Intel® Advanced Matrix Extensions (Intel® AMX)	

Figure 2. Intel® AMX accelerates Riskfuel's AI DNNs, enabling proprietary valuation and risk management models to run faster.[11]

> "We were able to unlock our software and deliver better customer outcomes by taking advantage of Intel® AMX, which is great for AI. Working with Intel enabled us to revolutionize the trading market."
>
> **– Philippe Chatigny, Senior Machine Learning Research Engineer, Riskfuel Analytics**

Conclusion

Are you ready to perform lightning-fast "what if" analyses and get your portfolio's current, real-time valuation? Running Riskfuel, which has been optimized for Intel AMX, enables traders to adjust their strategy and portfolios in real time to improve business outcomes—without investing in expensive, alternative compute architectures. The up-to-date data provided by Riskfuel on Intel architecture can help deliver new business opportunities and significantly reduce operating costs. Use the resources below to learn more about refreshing your compute resources to 4th and 5th Gen Intel Xeon Scalable processors and put Riskfuel to work in your trading business today.

Learn More

You may find the following resources helpful:

- 5th Generation Intel® Xeon® Scalable processors
- 4th Generation Intel® Xeon® Scalable processors
- Intel® Advanced Matrix Extensions
- Riskfuel home page

Find the solution that is right for your organization. Contact your Intel representative or visit **intel.com/FSI**.

Solution Provided By:

Riskfuel | intel.

[1] See [P15] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/5th-generation-intel-xeon-scalable-processors/

[2] See [P16] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/5th-generation-intel-xeon-scalable-processors/

[3] See [P14] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/5th-generation-intel-xeon-scalable-processors/

[4] Riskfuel, January 2020, "1,000,000x faster models."

[5] Riskfuel, https://riskfuel.com/slash-your-compute-costs/

[6] See endnote 1.

[7] See endnote 2.

[8] See endnote 2.

[9] See endnote 2.

[10] Intel, https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/what-is-intel-amx.html.

[11] See endnote 2.