White Paper

Financial Services

How to Accelerate XVA Performance

As banks look to reduce, mitigate, and optimize XVA and other capital charges, they are making an investment in XVA capabilities in an attempt to solve the computational challenge of simulating a full universe of risk factors.

quantifi

Table of Contents

Introduction

In the post-crisis world, an increasing number of banks have set up a centralized XVA desk. With the introduction of new regulations to ensure that banks are adequately capitalized, it has become common practice to include certain costs in the pricing of OTC derivatives that, in many cases, had previously been ignored. To assist in the pricing for the cost of dealing with a counterparty in a derivative transcription, the markets have developed various metrics including CVA, DVA, FVA, CoIVA, KVA, and MVA—collectively known as XVAs.

One of the key challenges of XVAs is that adjustments need to be calculated on a portfolio basis rather than trade by trade. This requires dealing with a large number of computations and orders of magnitude more calculations for accurate results. The calculation of XVAs is highly complex, combining the intricacies of derivative pricing with the computational challenges of simulating a full universe of risk factors.

With the introduction of new regulations to ensure that banks are adequately capitalized, it has become common practice to include certain costs in the pricing of OTC derivatives that, in many cases, had previously been ignored.

Another key challenge is how to efficiently calculate XVA sensitivities. While sensitivities have always been an important component of XVA desk risk management, the FRTB-CVA framework published by the Basel Committee in 2015 has made managing regulatory capital a priority for banks globally. This has further driven the demand for calculation of sensitivities. Banks that are unable to calculate CVA capital charge using the sensitivity-based FRTB approach will have to use the rather punitive formula-based basic approach.

Why Is It Important to Make XVA Calculations Faster?

XVAs are simulation-based calculations. The market standard is to use Monte Carlo (MC) simulations, with each covering thousands of paths across a large number of future time steps.

For example, an MC simulation of 2000 paths across 78 time steps for a portfolio of 40,000 trades requires up to 6.24 billion calculations that on average produce over 10GB of compressed result data. Sensitivities, stress tests, and attribution calculations can increase the number of calculations by an order of magnitude. XVA-related calculations are by far the most computationally resource-intensive for a bank. Hence banks are looking for quantitative as well as technology-based solutions designed to optimize performance.

What Is XVA?

X-value adjustment (XVA) is a generic term referring collectively to a number of different valuation adjustments in relation to derivative instruments held by banks. XVA is a generic acronym where "X" is related by a letter such as "C" for credit, "D" for debt, "F" for funding, "K" for capital, "M" for margin, and so on, and VA stands for valuation adjustment.

Credit value adjustment (CVA) is the difference between the risk-free portfolio value and the portfolio value that takes into account the possibility of a counterparty's default. In other words, CVA is the marketing value of counterparty credit risk.

While CVA reflects the marketing value of counterparty credit risk, additional valuation adjustments for debit (DVA), funding (FVA), capital (KVA), and margin (MVA) represent some of the other valuation adjustments that need to be considered when valuing derivatives.

Why is XVA Important?

Fair value accounting rules require banks to adjust derivative book valuations by credit and debt value adjustments. Moreover, market-making bank trading desks need to price derivatives taking into account all the XVA adjustments to ensure profitability.

Typically, larger banks will have dedicated XVA desks for managing XVA. These desks rely on measures such as XVA sensitivities, stress tests, and attribution to monitor and hedge XVA risks.

Calculating XVA

The demand for higher performance has highlighted the need to get the most out of the latest generation of software. A distributed architecture that supports the heavy demands of big data provides a number of benefits when dealing with large, complex portfolios. The main benefits include scalability, reliability, and resilience. However, the use of distributed computing for calculating XVA also presents a number of challenges, mainly in regard to I/O performance and CPU processing.

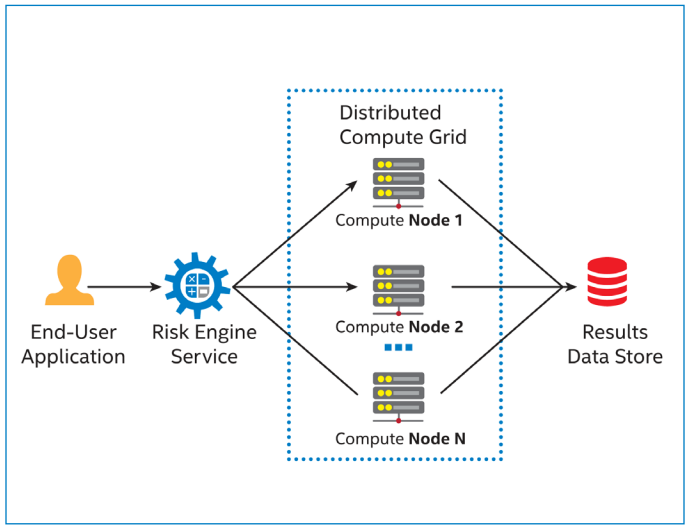
While distributing the workload increases the calculation performance, it comes at a cost of transporting and persisting results to the data store. Furthermore, to compute the results the calculations reference data is loaded from the data store. Measuring XVA is a highly complex process that requires having to save and manipulate large amounts of data. Having access to large-scale distribution and big data technology to minimize I/O is important. XVA calculations are also time-critical, which requires a high-performance CPU to handle the workload.

A number of factors influence the speed of XVA computations, including the size of the portfolio, the amount of market data, and the configuration of the Monte Carlo simulations themselves. Collectively these increase the number of calculations that need to be performed. Simulation configuration increases the complexity of each individual calculation, which means each calculation takes longer to complete and generate results.

The number of simulated market variables, the amount of time steps to simulate, and the number of simulation paths are the configurations that have the greatest impact on performance. Simulations are configured to have a number of market variables that evolve over time. The number of time steps are the number of steps into the future that are simulated to evolve the market variables, whereas the number of simulation paths is the number of times the market variables are simulated over time. Naturally, as any of these are increased, the calculations become more complex. The most expensive factor in the performance of these simulations is the number of simulation paths.

All these factors result in XVA calculations being very resource-intensive, even when using a distributed computing architecture. On a set-up with an average number of simulation paths and an average portfolio size, a full set of XVA calculations (simulations, aggregation, and sensitivities) can take well over 30 minutes to complete and generate over 200Gb of result data to be persisted.

Quantifi is built on a modern microservices architecture using a distributed computation system backed by a data store. XVA calculations are set up to take advantage of the distributed computation system, as it allows results to be calculated quickly in parallel. The distributed architecture also has the benefit of being able to scale up to meet increased workloads.



Quantifi is built on a modern microservices architecture using a distributed computation system backed by a data store.

CPU vs. I/O

For the Quantifi XVA use case, performance analysis can be broken down into two main components: CPU performance

(time spent on calculations) and I/O performance (time taken to transfer and store the data involved in the calculations). In an ideal scenario, Quantifi XVA calculations would be limited almost entirely by CPU performance, since the majority of the workload is due to the large set of computations.

Inherent to the performance of Quantifi XVA is a push-and-pull relationship between the CPU and I/O. As calculations complete faster, the time spent on I/O to transfer and save the results increases. This challenge is made more prevalent using a distributed architecture, as adding more computing resource also increases the amount of results to be transferred and stored at the same time. Similarly, if calculations become more complex, then the size of the results to store and transfer will increase and affect how quickly I/O can complete.

To determine the impact on Quantifi XVA calculations, a sample environment was created. A dummy portfolio was used to profile the performance of the system with a set of XVA calculations. While the calculations completed, system-wide performance metrics were collected to serve as a set of baseline numbers to be examined and compared with those of subsequent tests.

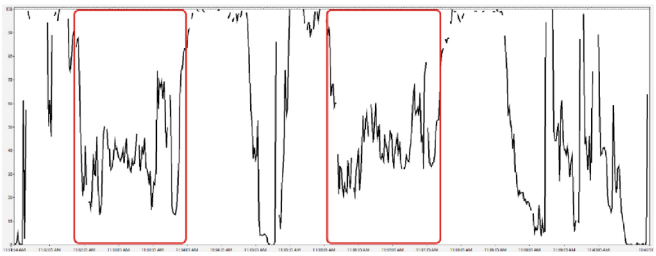


Diagram 1: % CPU time used with Intel Xeon Platinum 8180 processor persisting results to disk.

Diagram 1 demonstrates the existence of the I/O bottlenecks during the XVA calculations. The circled sections outline where the CPU should be performing calculations but is instead at a lower utilization as they wait to persist result data to the disk on the data store. The circled sections account for 38 percent of the overall duration.[1]

Diagram 1 highlights the percentage of CPU time used across the distributed computation system. This metric is used to determine the times when the CPU is running calculations versus when it is idle. Ideally, CPU performance would reach 100 percent and remain at that level for the duration of the calculations. The troughs in the middle of the graph illustrate the presence of I/O, causing the CPU to go into an idle state.

The initial analysis indicated that the first step to improve performance was to minimize the time spend on I/O. From the baseline performance statistics and the accompanying graphs, it appears that XVA calculations are being stopped because of the I/O required to transfer and store the results.

To compare performance, another set of calculations was completed, with the results persisted to memory. The expectation was that by eliminating the step of persisting the results back to the data store, the time spent on I/O between calculations should all but disappear.

Runtime with persistence (mm:ss)	Runtime with persistence (mm:ss)	% Decrease
9:01	6:59	23

Table 1: XVA Caluculation durations with and without persisting results to disk

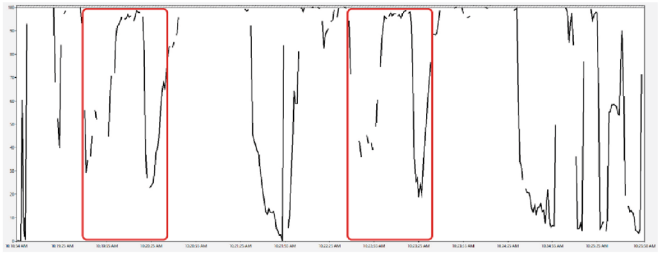


Diagram 2: CPU tiime used without persisting results to disk.

Diagram 2 illustrates the best-case scenario I/O performance using the existing hardware. The circled regions in Diagram 2 highlight that the CPU performance dips are much smaller than those in Diagram 1. As shown in Diagram 2, the CPU utilization is much better in the regions affected by I/O in Diagram 1. The 2nd diagram's circled regions are also 38 percent shorter than they were in Diagram 1.

Comparing the durations in Table 1 with and without persisting, the results highlight that approximately 23 percent of the total duration is spent on the I/O required to save the results. This shows that the I/O related to persisting result data can be improved. The first step was to improve the write I/O performance.

Intel Recommendations and Test Results

For the next test, in an effort to improve the write I/O performance and reconcile the difference between the durations collected with and without persisting to the data store, the CPU on the data store was upgraded to the 2nd Generation Intel® Xeon® Scalable processor.

The 2nd Gen Intel Xeon Scalable processors boast a number of features designed to improve I/O performance, including expanded I/O through 48 lanes of PCIe 3.0 bandwidth, integrated Intel Quick Assist Technology (Intel QAT), and access to the Intel Intelligent Storage Acceleration Library (Intel ISA-L). The expanded I/O bandwidth could directly improve the I/O performance out of the box, as more lanes allow more results to be written to the data store in parallel. Integrated Intel QAT promises efficient, enhanced data transport capabilities. Intel ISA-L is a feature that will become more useful as more software adopts it to optimize and improve operation performance.

2nd Gen Intel Xeon Scalable processors also support Intel Optane™ persistent memory, an innovative storage technology in which specialized memory provides a faster alternative to traditional disk storage. As a first step, Quantifi replaced the existing 1st Gen Intel Xeon Scalable processor

Intel Xeon Platinum 8180 processor runtime (mm:ss)	Intel Xeon Platinum 8260L processor CPU runtime (mm:ss)	% Decrease
9:01	7:11	20

Table 2: XVA calculation durations on Intel Xeon Platinum 8180 processor vs Intel Xeon Platinum 8260L processor. [2]

Intel Xeon Platinum 8260L processor runtime with persistence (mm:ss)	Intel Xeon Platinum 8260L processor runtime without persistence (mm:ss)	% Decrease
7:11	6:59	2

Table 3: XVA Calculation durations on Intel Xeon Platinum 8260L processor without persisting results to disk.

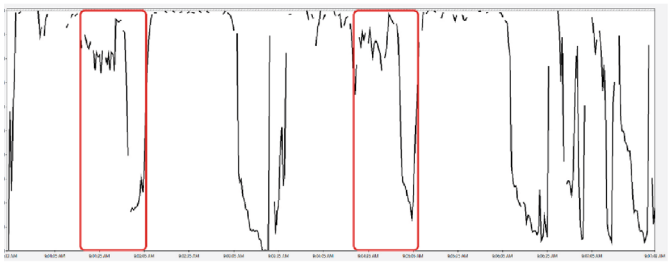


Diagram 3: CPU time used with Intel Xeon Platinum 8260L processor persisting results to disk.

The conclusion to draw from **Diagram 3** is that is closely resembles Diagram 2—the circled areas are a similar size. This indicates that the change in CPU has a demonstrable impact on I/O performance. Switching to the 2nd Gen Intel Xeon Scalable processor and persisting results to the disk yielded similar performance when compared with using a 1st Gen Intel Xeon Scalable processor and persisting results to memory. The circled regions in Diagram 3 are also 40 percent smaller than those in Diagram 1.

on the data store but kept the storage technology constant and collected the same metrics as the previous test.

Comparing this test to the previous test where results were persisted to memory shows that the change in hardware alone improved the I/O performance of writes considerably. In terms of total duration, the gap in performance between a set of calculations with persistence and a set of calculations without persistence has dramatically reduced using a 2nd Gen Intel Xeon Scalable processor. The change in CPU resulted in a better write performance, so repeating the test without persisting results with the 2nd Gen Intel Xeon Scalable processors was a useful step to determine whether the I/O could be further improved.

Intel Xeon Platinum 8260L processor runtime with persistence (mm:ss)	Intel Xeon Platinum 8260L processor runtime without persistence (mm:ss)	% Decrease
7:11	6:37	8

Table 4: Total duration on Intel Xeon Platinum 8260L processor with and without persisting results to disk.

Table 4 demonstrates that the change from 1st Generation to 2nd Generation Xeon Scalable processor reduced the I/O time for calculations from 23 percent to 8 percent. This is a significant boost to I/O performance from just a CPU change. Quantifi conducted another test, this time utilizing Intel Optane persistent memory (Intel Optane PMem) for the data store. The goal was to determine whether Intel Optane PMem could further enhance the I/O performance using the 2nd Gen Intel Xeon Scalable processor given that the read and write speeds are considerably faster than that of the non-volatile memory express (NVMe) drive previously used.

The first results using Intel Optane PMem were not very promising in terms of overall runtime. The runtimes between NVME and Intel Optane PMem were very similar, and across all samples fell within the margin of error for elapsed time. As such, the conclusion from this set of results was that performance remained consistent for 2nd Gen Intel Xeon Scalable processors with NVMe versus 2nd Gen Intel Xeon Scalable processors with Intel Optane PMem.

A more positive outcome was the dramatically improved throughput of the I/O, having switched from NVMe to Intel Optane PMem. The runtimes were similar from NVMe to Intel Optane PMem, which is likely due to the result data not being large enough to max out the NVMe write speeds for long. In a more expensive environment using more computing resources or a larger number of simulation paths, there could be a greater benefit from using Intel Optane PMem.

Intel Xeon Platinum 8260L processor + NVMe (bytes/sec)	Intel Xeon Platinum 8260L processor + Intel Optane PMem (bytes/sec)	% Decrease
567,351,788	1,377,171,632	143

Table 5: Maximum recorded write pressure on Intel Xeon Platinum 8260L processor using the NVMe and Intel Optane PMem.[3]

The table above, comparing the maximum recorded disk writes in bytes per second, shows Intel Optane PMem writes data over 130 percent faster than NVMe. Intel claimed that if the size of the results grew larger than NVMEs max throughput, then the Intel Optane PMem calculation durations would stay consistent while the NVMe durations

Number of simulation paths	Average disk writes (bytes/sec)	Duration (mm:ss)	% increase average disk writes from 2000 paths	% increase duration from 2000 paths
2000	73,053,585	7:09	-	-
3000	83,965,349	9:17	13	23
4000	95,406,641	11:10	23	36
5000	108,578,910	12:54	33	45

Table 6: Average I/O throughput comparison using variable amount of simulation paths.

would become larger. This highlights that Intel Optane PMem provides for greater system scalability. When combined with the other I/O performance benefits seed with the 2nd Gen Intel Xeon Scalable processors, these new CPUs appear to be very attractive in the long term for large-scale XVA calculation workloads.

Additional testing was carried out to determine the effects of running with a greater amount of computing resources and an increased number of simulation paths on the systems. Metrics collected on a consistent hardware set-up demonstrate the effect of increasing the number of simulation paths by increments of 1000. As the number of simulation paths increase, both the average disk writes and duration jump by a sizeable margin.

In a similar fashion, increasing the amount of computing resources for these calculations also has a direct impact on the average disk writes. Raising the amount of distributed computing resources ups the number of writers persisting results to the data store. The calculations also generally complete faster using more compute cores up to the point where the I/O becomes the limiting factor.

Number of compute cores	Average disk writes (bytes/sec)	% increase between average disk writes from 176 cores
176	77,651,854	-
248	85,761,755	10
320	93,116,821	20

Table 7: I/O throughput comparison using variable amount of compute node cores.

The above table demonstrates that the average I/O pressure increases as we continue to add compute nodes. Extending more distributed computing resources to the system causes increased stress on I/O. Combined with how additional computing resources enable calculations with more simulations to complete faster, it is evident that the I/O pressure will continue to scale up with just both of these variables increasing. The I/O can be scaled up by using other techniques, too, such as increasing the size of the portfolio. If all these factors were to scale up, eventually the average throughput would surpass the maximum on NVMe and Intel Optane PMem will become the most performant storage option.

Conclusions

The various tests carried out in this white paper demonstrate that leveraging Intel's newer hardware can accelerate the performance of large-scale XVA workloads by increasing performance of the CPU and improving the efficiency of I/O.

This makes upgrading to Intel's newer generation processor much more compelling, as it provides the ability to scale with portfolio size, computing resources, and calculation complexity.

The test revealed that the use of Intel Optane persistent memory over more traditional storage offers greater scalability if the XVA workload becomes more expensive. Further, the number of simulations, the amount of computing resources, and portfolio size all stress the system's I/O performance when writing to the data store. This makes upgrading to Intel's latest newer generation processor much more compelling, as it provides the ability to scale with portfolio size, computing resources, and calculation complexity.

quantifi

Quantifi is a provider of risk, analytics and trading solutions. The company's 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 client focus, depth of experience, and commit to innovation, Quantifi is consistently first-to-market with intuitive, award-winning solutions.

1 Quantifi worked on this project from Oct 2019 through Jan 2020 and did various tests using different system configurations. Baseline Database Server Configuration: Intel Xeon Platinum 8180 processor, 56 physical cores (112 logical cores), Frequency: 2.5 GHz, HT & Turbo on, 288GB RAM (12x16GB and 12X8GB) PC4-21300 ECC Registered 1.2 Volts DDR4 DIMM, M, Samsung SSD 970 Pro 1 TB.

2 Quantifi's benchmark using 2nd Generation Intel Xeon Scalable Processor: Intel Xeon Platinum 82600L, 48 physical cores (96 logical cores) Frequency 2.4 GHz, HT & Turbo on 192 GB RAM, 12x16 GB 2933MHz PC4-23400 ECC Registered 1.2 Volts DDR4 DIMM, 1 x 480 GB Intel DC S4600 Series SATA 6 Gb/s 2.5" SSD TLC, 1 x Aspeed AST2500 On-Board, 1 x Intel® Server Board S2600WFTR, 1 x Intel® Ethernet Connection X722 10 Gbps Dual-Port On-board, 1 x Intel Remote Management Module 4 Lite, 2 x 1100 Watts Hot-Swappable 110/220v Samsung SSD 970 Pro 1 TB. In order to test the scalability, Quantifi initially used 176 cores on Quantifi's compute grid and then gradually scaled it up to 248 cores and then 320 cores. These are computation resources to generate more and more data in parallel to test the performance of Intel 2nd generation Xeon Scalable processor and Intel Optane persistent memory. Performance test results are present in the paper.

3 Quantifi benchmark using 2nd Gen Intel Xeon Scalable Processor and Intel Optane persistent memory: Intel Xeon Platinum 8260L, 48 physical cores (96 logical cores), Frequency: 2.4 GHz, HT & Turbo on 192 GB RAM, 12 x 16 GB 2933MHz PC4-23400 ECC Registered 1.2 Volts DDR4 DIMM, 12 x 128 GB AEP DIMM, 1 x 480GB Intel DC S4600 Series SATA 6Gb/s 2.5" SSD TLC, 1 x Aspeed AST2500 On-Board, 1 x Intel® Server Board S2600WFTR, 1 x Intel® Ethernet Connection X722 10Gbps Dual-Port On-board, 1 x Intel Remote Management Module 4 Lite, 2 x 1100 Watts Hot-Swappable 110/220v Samsung SSD 970 Pro 1 TB, Intel Optane persistent memory. In order to test the scalability, Quantifi initially used 176 cores on Quantifi's compute grid and then gradually scaled it up to 248 cores and then 320 cores. These are computation resources to generate more and more data in parallel to test the performance of Intel 2nd generation Xeon Scalable processor and Intel Optane persistent memory. Performance test results are present in the paper.

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.

Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

© 2021 Intel Corporation Printed in USA

082021/RJMJ/CP/PDF ♻ Please Recycle 348167-001US