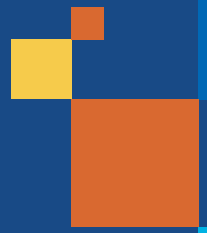
AI in Financial Services

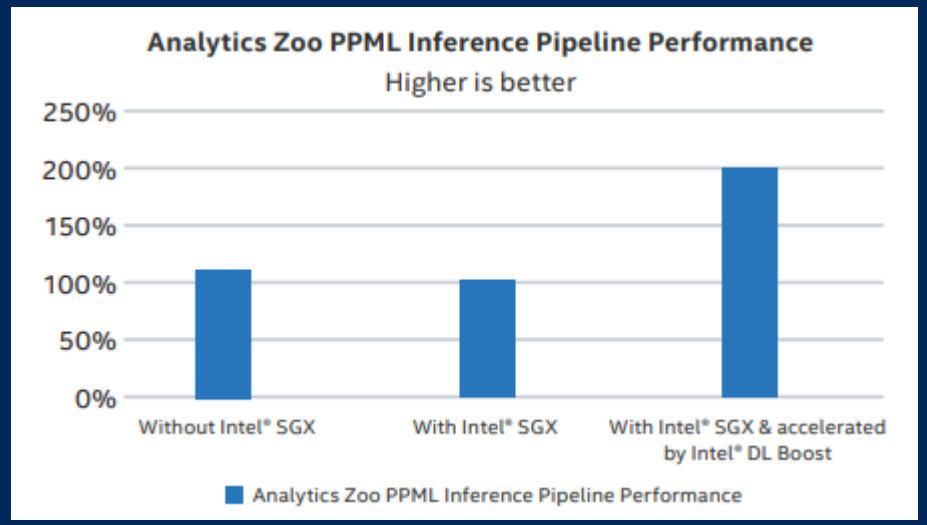
# 3rd Gen Xeon Proof Points

June 2022



intel®

# Privacy Preserving Machine Learning


[Link to Deeper Content]

## Privacy Preserving Machine Learning (PPML) Powered By 3rd Generation Xeon Scalable Processors

Ant Group has collaborated with Intel to build a PPML platform on top of Intel Software Guard Extension (Intel SGX) and Occulum (Ant Group's Multi Process OS)

**KEY CHALLENGE**

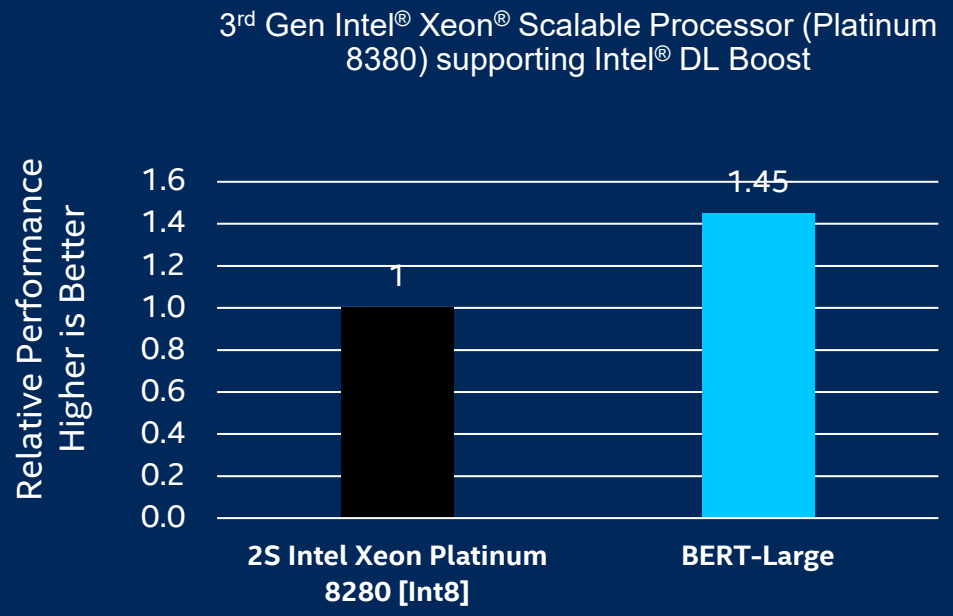
- Financial Institutions often benefit from sharing and analyzing data to predict consumer trends or manage compliance processes.
- Highly regulated industries like financial services are often unable to collaborate since breaches of consumer data can result in fines from regulating industry bodies and lack of consumer trust.

**Technology**

- Privacy preserving machine learning helps to address these risks by protecting the privacy of sensitive user data. This capability enables FI's to share and analyze data without jeopardizing confidentiality of sensitive information.
- Intel and Ant Group worked together to build a more secure, end-to-end and distributed inference service pipeline based on an Analytics Zoo PPML platform.
- Upon application of Intel DL Boost with INT8, the Intel SGX protected inference pipeline demonstrated a 2x increase in throughput.

See https://www.intel.com/content/www/us/en/financial-services-it/preserve-privacy-with-ml-paper.html for more information

# Natural Language Processing


[Link to Deeper Content]

## Real Time Inference Performance Gains using TensorFlow for Natural Language Processing

**Key Challenge**

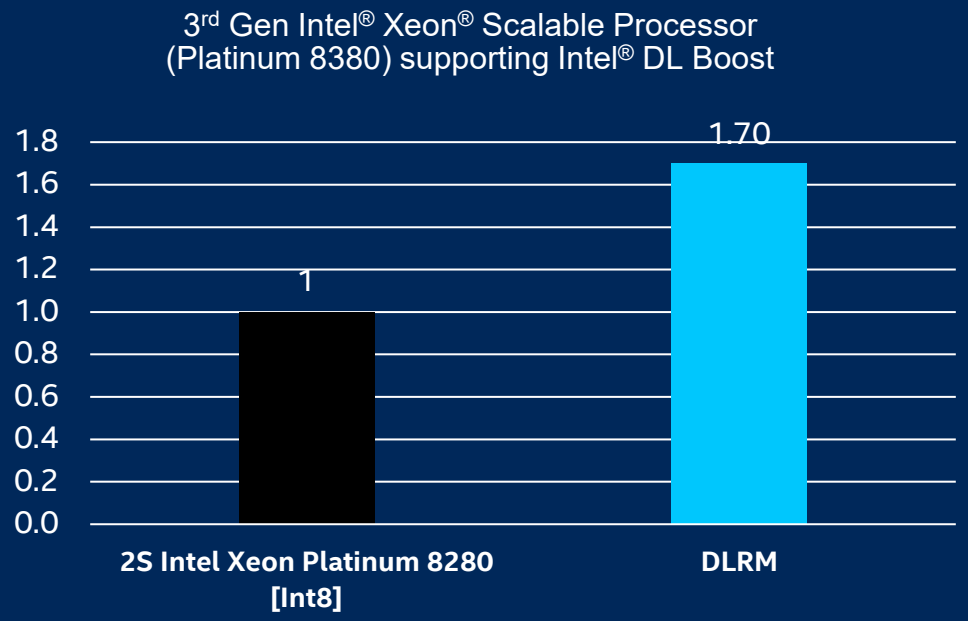
- Customers are seeking more convenient ways to engage with Financial Institutions. This has created a competitive market landscape that has been amplified by the impact of technology on the industry.
- Natural Language Processing is one technique that financial institutions are using to differentiate themselves by generating new revenue streams, reducing costs or managing compliance.

**Technology**

- Deploying digital capabilities such as roboadvisors can help generate new revenue streams or reduce costs. For example, banks are using roboadvisors to interact with customers in different languages globally or automate the retail banking experience.
- These roboadvisors use natural language processing to communicate with humans by processing and analyze large amounts of natural language data. The benchmark, BERT Large, is a key metric to evaluate NLP performance of a technology configuration.
- 3rd Gen Intel® Xeon® Scalable Processors delivered 1.45x better performance than the previous generation on BERT-large.

See https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/?r=1677269083 for more information

# Recommendation Engine


[Link to Deeper Content]

See slide 10 for more information.

## Batch Inference Performance Gains with optimized PyTorch for Recommendation Engine

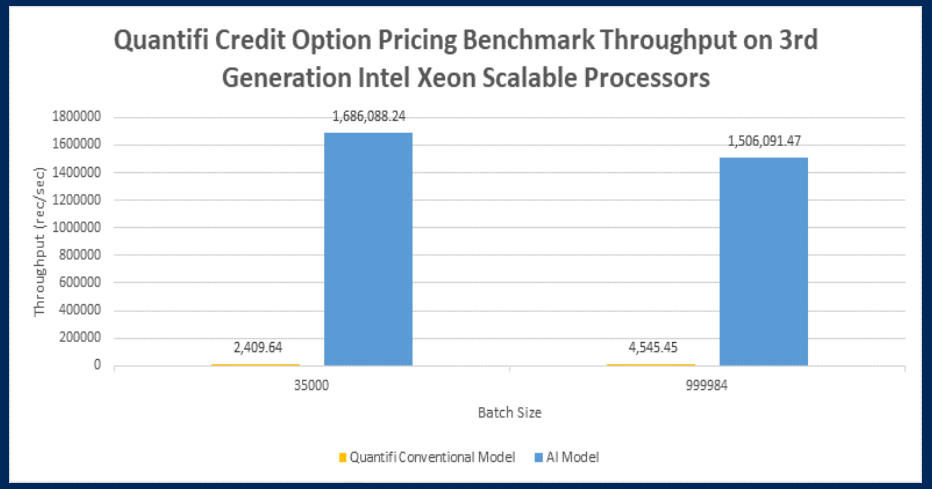
**Key Challenge**

- Customers are seeking more convenient ways to engage with Financial Institutions. This has created a competitive market landscape that has been amplified by the impact of technology on the industry.
- Recommendation Engines are one technique that financial institutions are using to differentiate themselves by generating new revenue streams, reducing costs or managing compliance.

**Technology**

- Financial Institutions are using RE's to improve customer satisfaction by offering hyper personalized products and services. For example, the banking industry is using recommendation engines to offer loans that are tailored to an individual based on key factors (e.g. credit score, interest rates).
- These capabilities require recommendation engines which is the ability for a computer to filter data to recommend the most relevant items to a particular user or customer. DLRM is a popular benchmark that evaluates recommendation engine performance based on a technology configuration.
- 3rd Gen Intel® Xeon® Scalable Processors delivered 1.7x better performance than the previous generation on DLRM.

# Risk Management


## Intel and Quantifi Accelerate Derivative Valuations Using AI on 3rd Generation Xeon Scalable Processors

Intel and Quantifi collaborated to enable financial institutions to calculate fair value pricing in real time without specialized hardware.

**Key Challenge**

- Following the global financial crisis in 2008, global regulators established new compliance measures that require financial institutions to calculate the valuation and risk associated with their investment positions.
- Portfolio managers and traders that use over the counter derivatives often lack an accurate real time view of their positions since there isn't always a market price available. This requires traders to accurately calculate their theoretical value which can be expensive and require significant machine time.

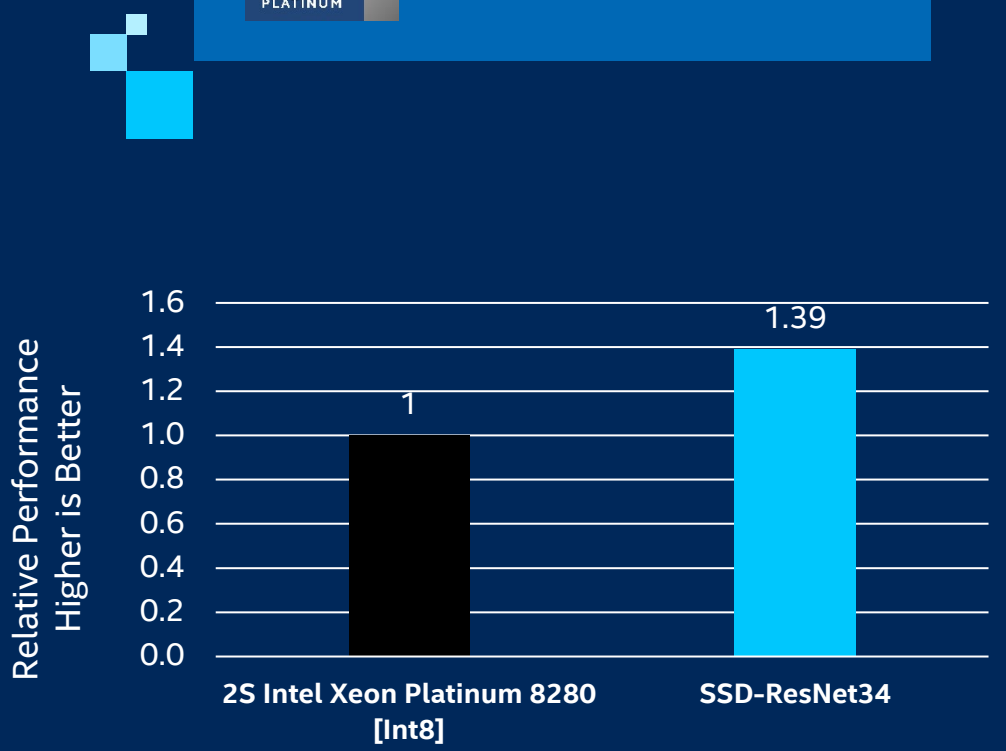
**Technology**

- Artificial intelligence enables FI's to calculate risk in 'real time' by using historical data to predict theoretical future outcomes.
- Quantifi developed an Artificial Neural Network (ANN) to approximate a known derivative valuation using inputs such as market data and interest rates.
- The Credit Option Pricing Benchmark measured on 3rd Gen Intel Xeon Scalable System demonstrated an approximated 700x increase in inference throughput using the AI model compared to a conventional model.

Link to Deeper Content

See https://www.intel.com/content/www/us/en/financial-services-it/derivative-valuations-faster-with-ai-paper.html for more information

# Object Detection




## Real Time Inference Performance Gains using TensorFlow for Image Recognition

**Key Challenge**

- Customers are seeking more convenient ways to engage with Financial Institutions. This has created a competitive market landscape that has been amplified by the impact of technology on the industry.
- Object Detection is one technique that financial institutions are using to differentiate themselves by generating new revenue streams, reducing costs or managing compliance.

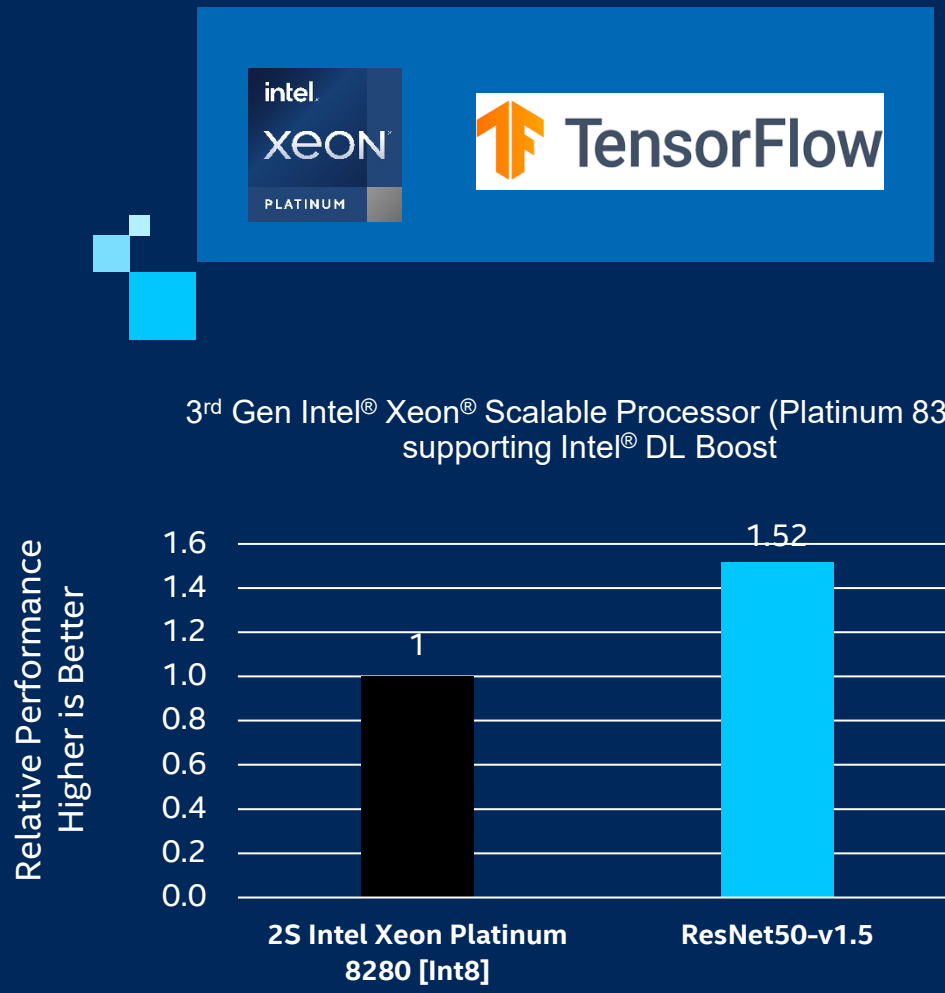
**Technology**

- Financial Institutions are using object detection to enforce compliance measures. For example, capital markets firms are using object detection to monitor trade floors to ensure that traders are conducting business legally.
- Trade floor surveillance uses object detection to identify objects of a certain class with an image. The benchmark, SSD-ResNet34, is a key metric to evaluate object detection performance of a technology configuration.
- 3rd Gen Intel® Xeon® Scalable Processors delivered 1.39x better performance than the previous generation on SSD-ResNet34.

Link to Deeper Content

See https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/?r=1677269083 for more information

# Image Classification




[Link to Deeper Content]

## Real Time Inference Performance Gains using TensorFlow for Image Classification

**Key Challenge**

- Customers are seeking more convenient ways to engage with Financial Institutions. This has created a competitive market landscape that has been amplified by the impact of technology on the industry.
- Image Classification is one technique that financial institutions are using to differentiate themselves by generating new revenue streams, reducing costs or managing compliance.

**Technology**

- Financial Institutions are using image classification to reduce costs by automating processes. For example, the insurance industry is using image classification to evaluate claim liability. Customers can take a picture of an accident that requires insurance and artificial intelligence will determine claim processing details (e.g. payout, liability, etc.)
- These capabilities require image classification which is the ability for a computer to identify people or objects using a camera and artificial intelligence software. ResNet50 is a popular benchmark that evaluates image classification performance based on a technology configuration.
- 3rd Gen Intel® Xeon® Scalable Processors delivered 1.52x better performance than previous generations on ResNet50.

See https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/?r=1677269083 for more information


intel

# Notices and Disclaimers

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.

Your costs and results may vary.

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

# Configuration Details

Slide 2: https://www.intel.com/content/www/us/en/financial-services-it/preserve-privacy-with-ml-paper.html

Slide 4: Platinum 8380: 1-node, 2x Intel Xeon Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X261 , HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, INT8, RN50 BS=128, DLRM BS=16, INT8. Platinum 8280: 1-node, 2x Intel Xeon Platinum 8280 processor on Wolf Pass with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-48-generic, 1x Samsung_SSD_860, Intel SSDPE2KX040T8, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, INT8, RN50 BS=128 , DLRM BS= 16 , Software : PyTorch 1.5.0 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed with intel-extension-for-pytorch branch icx (container- intel/image-recognition:pytorch-1.5.0-rc3-icx-a37fb5e8-resnet50-int8, container- intel/intel/recommendation:pytorch-1.5.0-rc3-icx-a37fb5e8-dlrm-int8, ), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart), test by Intel on 3/12/2021. Results may vary.

Slide 5: https://www.intel.com/content/www/us/en/financial-services-it/derivative-valuations-faster-with-ai-paper.html

Slide 3, 6-7: https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/?r=1677269083