Technical Research Study

PROWESS

Changing Hardware Needs in the Evolving Financial-Industry Landscape

Finance is a broad industry with ubiquitous needs for fast artificial intelligence (AI) inferencing and robust security.

Executive Summary

Finance is at the forefront of artificial intelligence (AI), with the industry employing AI across its breadth. Because financial organizations might use AI models hundreds or even thousands of times per day, they rely on AI inferencing as a way to quickly use AI models for decision making. An equally ubiquitous need in the financial industry is for hardware-based security built into servers and the server-manufacturing process. After all, financial services is a highly targeted industry. With this research note, Prowess Consulting aims to help you start your planning for future hardware investments by helping you make sure that the servers you consider can meet your security and AI-performance needs.

This document is part of an ongoing series of research notes exploring innovative technologies that can help forward-thinking businesses expand their offerings and reduce costs, while helping ensure security and privacy for users and data.

Industry Landscape

The financial industry was one of the first to broadly embrace machine learning (ML) and AI. Everyday use of AI has spread across the financial industry to become an integral tool in nearly every part of finance:

- Risk assessment and management
- Credit decision making
- Underwriting
- Advisory services and personalized banking
- Customer service
- Trading
- Fraud detection and prevention

A common theme among these diverse AI use cases in finance is the need for fast AI inferencing. ML and AI are divided into two principal phases: training (model creation) and inferencing (model use). During training, developers create a model by feeding it data so that it can start to make decisions and predictions, such as: "Is this a fraudulent transaction?" or "What is the maximum amount of risk this loan can generate?" When inferencing, the model then makes its decisions or predictions using new data that it has never seen before.

While training AI models often takes the limelight in public perception, inferencing on AI models—actually using those models to make predictions—is driving a shift from model-building to model-using. Calculating predictions from existing AI models is not nearly as computationally intensive as training new models. Depending on the complexity of a model, that model can take hours, days, or even weeks to train and tune. By contrast, an AI model in finance, once trained, might be used thousands of times per day—which means fractions of seconds count. In addition, inferencing must be consistently fast, and it must work in a variety of IT environments, such as branch offices or edge locations.

The increasing reliance on more and higher performing servers for AI underscores the need for strong hardware-based security to protect sensitive data and comply with regulations. The financial industry is one of the most heavily regulated around the world. Regulatory restrictions can limit moving customer data to the cloud, necessitating a strong on-premises footprint for at least part of financial firms' IT infrastructures. What's more, financial firms represent prime targets for malevolent actors, and they thus need robust security features that extend into the hardware and even into server supply chains. Let's take a closer look at each of these areas of concern in turn.

AI Inferencing

In contrast to training AI models, inferencing is an AI operational phase in which CPUs can play a key role. This role can involve both computing the output of AI models on operational data and operating as a hub for specialized inference accelerators.

Register Bandwidth Optimization for Inferencing

AI inference is just calculation, and so its performance is driven by how much calculating the processor register can do. The processor register is the location within a processor that holds the instructions, storage addresses, and individual numeric data necessary for computation. Increasing the number of computations that the processor register can compute in a single clock cycle is at the heart of accelerating AI inferencing.

One means of speeding up AI inferencing is by increasing the amount of numeric data that a processor can calculate in a clock cycle (that is, the bandwidth of the processor). Increases in the amount of floating-point numbers on which processors can calculate automatically speed up inferencing because AI models are generally based on decimal numbers. The generation-on-generation doubling of the intake floating point in 3rd Generation AMD EPYC™ processors is an example of this.

An appealing aspect of accelerating AI inferencing through increased register bandwidth for floating-point numbers is that it is completely transparent to existing AI models. Thus, if the risk-management office at a brokerage or a large bank wanted to speed up inference for its risk-assessment AI models, using newer servers with processors that can accommodate more floating-point numbers in their registers would produce immediate results.

INT8 Inferencing

A second means of accelerating inferencing is to quantize AI models to use 8-bit integers instead of floating-point numbers, although this method of acceleration requires some reworking of existing AI models. Moving to 8-bit integer (INT8) inferencing aids with memory management (a single 32-bit floating-point number takes up the same memory footprint as four 8-bit integer numbers) and processing time (the increased bandwidth in current-generation processors can perform the same operation on multiple 8-bit integers in a single processor clock cycle). Using INT8 can substantially increase inferencing speed while incurring minimal loss to accuracy.[1] Examples of processors increasing their bandwidth for INT8 include the increased INT8 bandwidth in 3rd Generation AMD EPYC processors and the specialized Intel® Advanced Vector Extensions 512 (Intel® AVX-512) Vector Neural Network Instructions (VNNI) instruction set.

> Using 8-bit integer (INT8) inferencing helps with AI memory management and processing time and can substantially increase inferencing speed while incurring minimal loss to accuracy.[1]

AI models trained using floating-point numbers cannot automatically use INT8 inferencing. Instead, these models must first be quantized to use 8-bit integers. Thus, a bank seeking to speed up its AI-driven customer-service applications would have to first quantize the models that power that app in order to harness the faster performance provided by INT8-based inference. Frameworks and toolkits for quantization include TensorFlow™, OpenVINO™, and AMD ROCm™.

PCIe® 4.0

A third way to speed up inferencing is to use inference accelerators. These accelerators are specialized processors designed for AI inferencing. A common processor-to-processor interconnect protocol used to connect accelerators to server processors is PCIe®.

The greatly increased bandwidth in PCIe 4.0 (16 gigatransfers per second [GT/s], compared to 8 GT/s in PCIe 3.0) speeds up data transfers to inference accelerators. Examples of server processors that support PCIe 4.0 include 2nd Generation AMD EPYC processors and 3rd Generation Intel® Xeon® Scalable processors. An insurance firm seeking to accelerate the performance of its underwriting AI models, for example, could use AI inference accelerators, but it would want to be careful to use interconnects based on PCIe 4.0 between processors and accelerators.

Performance Benchmarking

The workloads run by financial organizations—AI and others—are highly sensitive to changes in performance. To help uniformly measure the performance of workloads central to financial institutions, the industry organized the Securities Technology Analysis Center (STAC) Benchmark Council.

The STAC Benchmark Council is composed of more than 400 financial institutions and more than 50 vendor organizations that coordinate on addressing financial-services technical challenges and developing technology benchmark standards for financial organizations. Benchmark research at STAC ranges from general-purpose logistics such as data distribution and time synching to specialized tasks such as financial-algorithm back-testing. To see a complete list of these research areas, view recent performance results, and learn how STAC categorizes performance metrics for capital-market workloads, visit the STAC Research Domains page: stacresearch.com/domains.

Security

Financial firms often find themselves in the crosshairs of bad actors, whether criminals employing ransomware or state-sponsored hackers deploying malware. Because of the prominence of the industry as a target, organizations within finance universally need security features that extend into their hardware and even into the supply chains of their hardware suppliers. And while it's important for financial organizations—like other elements of critical infrastructure—to employ all aspects of cybersecurity best practices such as the National Institute of Standards and Technology (NIST) Cybersecurity Framework, security features predicated on confidential computing and hardware-enhanced security help most with protecting identified assets (in contrast to software tools best employed for detection, response, and recovery).

Confidential Computing

Confidential computing is an emerging initiative in finance and other industries that focuses on improving isolation for sensitive data payloads and securing data while at rest, in motion, and in use. While storage and network encryption are important in finance (as in other industries), much of this industry push also comes in the form of hardware-based memory protections; securing servers while they are booting is also a major component.

Memory Encryption

Hardware system memory can represent a vulnerability for data. While data might be encrypted at rest in storage and in motion across the network, it often subsequently needs to be decrypted into memory for use by applications. This can put sensitive information (such as the personally identifiable information [PII] of financial organizations' clients) at risk, particularly to attacks on the hardware itself that seek to recover information from system memory, such as from computers physically accessed in or taken from a bank branch office.

One method for addressing this vulnerability is to transparently encrypt system memory for operating systems. Both AMD and Intel have security features for this, with AMD® Secure Memory Encryption (AMD® SME) on AMD EPYC processors and Intel® Total Memory Encryption (Intel® TME) on 3rd Generation Intel Xeon Scalable processors. Further application isolation can be achieved at the cost of having to modify applications to use a trusted-execution environment (TEE), such as with Intel® Software Guard Extensions (Intel® SGX), to prevent data or code in the application from being altered. Thus, the wealth-management department of a bank seeking to better secure customer data could simply deploy computers powered by processors that used AMD SME or Intel TME, but it would have to modify its applications if it wanted to house proprietary models in an Intel SGX memory enclave, for example.

> Symantec researchers revealed that ransomware attackers have started using virtual machines (VMs) to help prevent discovery of their malware after encryption has begun.[2] Such attacks could be particularly costly for smaller banks and credit unions that might lack the security resources of larger financial organizations.

Encrypted Virtualization

In 2021, researchers at Symantec published evidence that ransomware attackers had started using virtual machines (VMs) to help prevent discovery of their malware after encryption had begun.[2] While such an evolution in tactics would be a problem for any organization, it could be particularly costly for smaller banks and credit unions that might lack the security resources of larger financial organizations.

Because VMs play an increasingly important role in financial services for efficiency and cost, protecting or isolating VMs is critical to protect financial data. A way to help limit threats like those illustrated by Symantec, which seek to compromise host servers through VMs, is to isolate guest operating systems and hypervisors from one another. An example of hardware-based technology that does this is AMD® Secure Encrypted Virtualization (AMD® SEV), which ensures that the respective pages in system memory are encrypted so that VMs and hosts cannot directly access each other's data in memory. Indeed, when using VMs, AMD SEV can achieve much of the same security for memory that TEEs such as Intel SGX put in place, but without having to modify applications running inside the VMs. AMD® Secure Encrypted Virtualization-Encrypted State (AMD® SEV-ES) goes further and encrypts all CPU register contents when a VM stops running. Doing so helps prevent leakage of information in CPU registers to components such as the hypervisor. AMD SEV-ES can even detect malicious modifications to a CPU register's state.

Secure Boot

Firmware-level remote attacks are a growing threat across the financial industry. Firmware is an attractive attack vector because it provides a means to compromise servers while they are booting, before software-based malware defenses even have a chance to start running. Secure boot helps defend against these threats by extending the silicon root of trust. Doing so helps protect the system by establishing an unbroken chain of trust from the silicon root of trust burned into the silicon itself to the BIOS. Unified Extensible Firmware Interface (UEFI) secure boot similarly helps continue the chain of trust from the system BIOS to the operating system (OS) bootloader. Examples of this technology include AMD® Secure Boot and Intel® Boot Guard.

Firmware can be further protected by additional hardware-based technologies. For example, BIOS live scanning in Integrated Dell™ Remote Access Controller (iDRAC) can verify the integrity and authenticity of the BIOS image when a server is powered on, which can help guard against firmware attacks engineered by compromising the BIOS. In addition, dynamic system lockdown such as that provided by iDRAC prevents system access using administrator privileges from altering firmware while the lockdown is in place. Locking down firmware in this manner also helps prevent unintentional migration of firmware and configuration settings from one server to another, which can pose a security vulnerability for other servers.

Hardware-Enhanced Security

Cyber-resilient architecture refers to robust layering of security through server hardware and following best practices such as NIST guidelines to prevent attacks, quickly identify attacks that do occur, and build an organization's capacity to recover quickly. Among other practices and technologies, cyber-resilient architecture encompasses both hardware-based hardening of the server boot process in order to secure server firmware and secured supply chains for server manufacturing.

Silicon Root of Trust

Firmware attacks can be a particularly pernicious threat for financial organizations. This is because an attack vectored on firmware can implant malware before the OS—and thus the software-based security running on that OS—has even started. To head off these attacks, server processors require a read-only encryption key that validates that the BIOS or UEFI drivers are legitimate. Cryptographically verified trusted booting such as this helps meets NIST recommendations for BIOS protection for servers and BIOS-integrity measurement; it also undergirds software-based security features such as secure boot in Windows Server®. The encryption key for security features such as these must be burned into the silicon during the manufacturing process. Examples of this include the root of trust enabled by iDRAC and HPE® Project Aurora.

Secure Supply Chains

The most fundamental attack vector on financial organizations' servers is during manufacturing and shipping, when hardware and firmware components can be altered in ways customers cannot detect. The only way to defend against these attacks is for server vendors to work to ensure that there is no tampering with products or insertion of counterfeit components before shipping products to customers. To do this, original equipment manufacturer (OEM) controls must span supplier selection, sourcing, production processes, and governance through auditing and testing. Material inspections during production can help identify components that are mismarked, deviate from normal performance parameters, or contain an incorrect electronic identifier. A prime example of these safeguards is Secure Component Verification on top of a secured supply chain, as offered by Dell Technologies.

Conclusion

Because of its unique regulatory and privacy requirements, the financial industry has a greater-than-usual need for on-premises servers. Nuances in server selection can impact two ubiquitous necessities in finance: AI inferencing and security. Specialized processor instruction sets and other inferencing accelerators are crucial to meeting financial AI needs. Security for finance must extend to server hardware and even server supply chains.

Start your planning for future hardware investments by making sure that the servers you consider can meet your security and AI-performance needs.

[1] Kim, et al. "Performance Evaluation of INT8 Quantized Inference on Mobile GPUs." *Institute of Electrical and Electronics Engineers (IEEE)*. December 2021. **https://ieeexplore.ieee.org/document/9638444**. Additional detail of inferencing speed-up and accuracy loss on Intel® processors is available at: OpenVINO. "Model Accuracy for INT8 and FP32 Precision." **https://docs.openvino.ai/latest/openvino_docs_performance_int8_vs_fp32.html**.

[2] Symantec. "Ransomware: Growing Number of Attackers Using Virtual Machines." June 2021. **https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/ransomware-virtual-machines**.