Technical Research Report

PROWESS

Accelerate Compute-Intensive Financial Workloads

Improvements in 4th Generation AMD EPYC™ processors over previous generations accelerate common workloads in the financial industry, according to third-party testing.

Executive Summary

Financial firms depend more than ever on computationally intensive workloads to compete and succeed. While many facets of IT hardware and infrastructure can affect application performance, Prowess Consulting examined the influence of the newest-generation system improvements to fuel application speed. To do this, we tested similar Dell™ PowerEdge™ R6525 and Dell PowerEdge R6625 platforms powered by 2nd and 4th Generation AMD EPYC™ processors, respectively.[1]

Our results showed marked improvements to benchmark performance and throughput for the PowerEdge R6625 server running the 4th Generation AMD EPYC processor. Our analysis also considered the qualitative performance and security features provided by the Dell™ platforms powered by AMD EPYC processors for the needs of financial firms.

Quantitative results

Dell™ PowerEdge™ R6625 servers powered by 4th Generation AMD EPYC™ processors have

up to **3.6x** greater performance

for artificial intelligence (AI) workflows than PowerEdge R6525 servers powered by 2nd Generation AMD EPYC processors.[1]

Industry Landscape

The financial industry has always been computationally intense. Proprietary applications, artificial intelligence (AI), deep learning (DL), and machine learning (ML) have become integral tools 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

Computational needs continue to increase in all of these corners of the financial industry. However, application performance can be driven by so many different variables that it can be difficult to identify where to start, particularly in an era when IT departments are routinely told to do more with less.

Moreover, the increasing reliance on more and higher-performing servers for finance 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 industries 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 therefore they need robust security features that extend into the hardware and even into server supply chains.

Putting System-wide Improvements to the Test

Server performance is multifaceted, with components ranging from memory to storage all contributing. Prowess Consulting decided to investigate the extent to which system-wide performance improvements—including upgrades to processor, memory, and PCIe® performance—can help financial firms keep up with their increasing computational needs. We did this by comparing two generations of AMD EPYC processors running on similar models of PowerEdge R6525 and PowerEdge R6625 servers. Details about both configurations are given in Table 1.

Table 1 | System configurations used for performance testing

	Dell™ PowerEdge™ R6525 (2nd Generation AMD EPYC™ processor)	Dell™ PowerEdge™ R6625 (4th Generation AMD EPYC™ processor)
Processor	2 x 64-core AMD EPYC 7702 processor	2 x 64-core AMD EPYC 9554 processor
Processor base frequency	2.0 GHz (max 3.9 GHz)	3.1 GHz (max 4.4 GHz)
Total cores/threads	128/256	128/256
Memory	512 GB (balanced configuration is 16 DIMMs)	768 GB (balanced configuration is 24 DIMMs)

We performed our testing of both platforms on bare-metal servers only, with Red Hat® Enterprise Linux® 8.7 installed (that is, with no virtualization layer). In addition, we selected configurations that would represent typical deployments, rather than maxing out the systems.

Measuring Performance

To measure performance for the computationally intensive workloads that are typical in the financial industry, we used a benchmark similar to TPCx-AI from TPC. The benchmark measures the performance of an end-to-end AI and data science platform, which emulates the behavior of financial AI solutions in current production environments.

The benchmark used in this study is designed to emulate the behavior of representative industry AI solutions that are present in production data centers and cloud environments. It assesses AI performance through several use cases. Use cases refer to single problems solved by the DL and ML data science pipeline in TPCx-AI. The pipeline is agnostic of any specific framework or syntax and can be implemented in many ways.

For financial firms, this means that this benchmark can give an idea of relative performance between different hardware systems across the entire DL/ML pipeline across different phases in the pipeline that include data generation, data management, training, scoring, and serving. This is crucial information, as different systems might perform different pipeline phases relatively better, but it is the overall DL/ML performance for the entire pipeline that matters most in the real world.

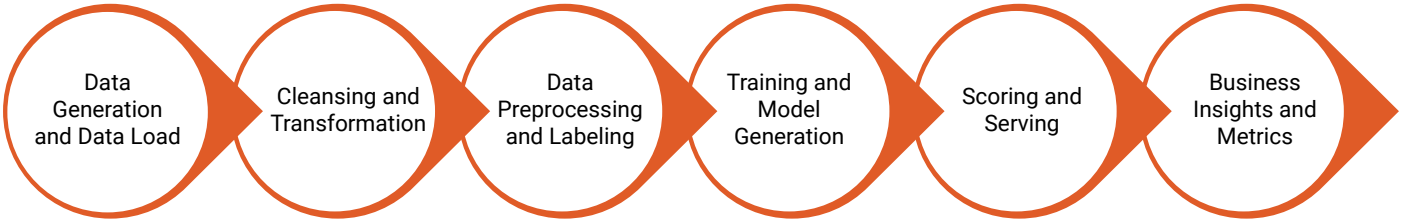

Figure 1 | The DL and ML data science pipeline used in this benchmark

Prowess Consulting engineers used a dataset of 10 GB (scale factor 10) and ran the benchmark three times on both servers. The median results show the PowerEdge R6625 server powered by 4th Generation AMD EPYC 9554 processors having up to 3.6x greater performance than the PowerEdge R6525 server powered by 2nd Generation AMD EPYC 7702 processors.

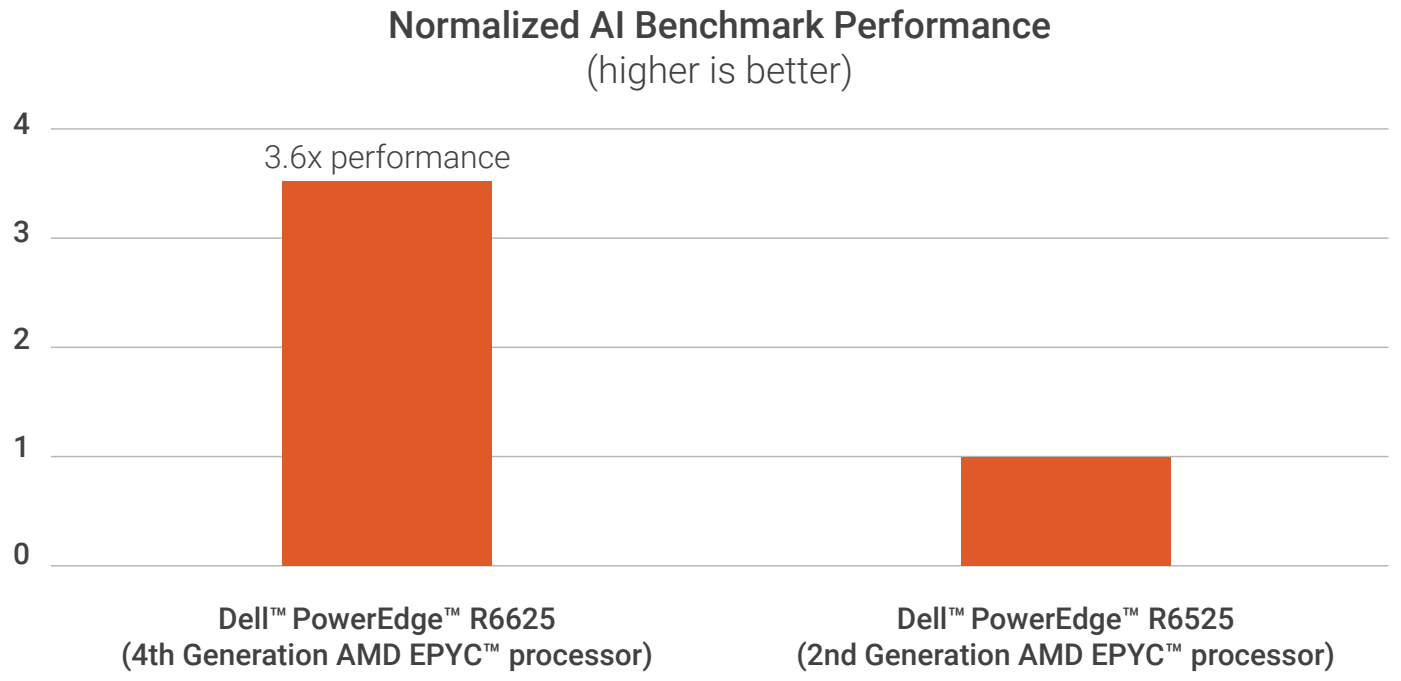

Figure 2 | Normalized AI benchmark results between a Dell™ PowerEdge™ R6625 server powered by 4th Generation AMD EPYC™ processors and a Dell PowerEdge R6525 server powered by 2nd Generation AMD EPYC processors

This benchmark measures both throughput and performance for the entire AI pipeline, and these benchmark results are suggestive of overall improvement to the AI performance. Isolating and comparing 4th Generation AMD EPYC processor performance to that of 2nd Generation AMD EPYC processors helps verify the suggestion of overall improvement to AI pipeline performance. Comparing Standard Performance Evaluation Corporation (SPEC) benchmark performance of 4th Generation AMD EPYC processors to 2nd Generation AMD EPYC processors shows a 2.1x improvement to both integer and floating-point computational performance.[2] The remainder of the 3.6x greater AI-pipeline performance that PowerEdge R6625 servers powered by 4th Generation AMD EPYC 9554 processors exhibited over PowerEdge R6525 servers powered by 2nd Generation AMD EPYC 7702 processors means that the PowerEdge R6625 servers also help accelerate the entire AI pipeline from end to end. This is particularly relevant for financial firms that can have multiple, long pipelines to address all their AI needs.

4th Generation AMD EPYC Processor Differentiators from 2nd Generation AMD EPYC Processors

Beyond generation-on-generation improvements to performance, we found that 4th Generation AMD EPYC processors provide additional benefits over 2nd Generation AMD EPYC processors for computationally intensive workloads, particularly for AI inferencing. The following sections will describe architectural improvements we found that benefit AI workloads.

AMD Zen Deep Neural Network (ZenDNN)

The ZenDNN library enables DL application and framework developers to improve DL inference performance on AMD processors. It features highly optimized programmatic components for AMD processors and targets a variety of workloads, including computer vision, natural language processing, and recommender systems. The latest version of the library is enabled, tuned, and optimized for inference on 4th Generation AMD EPYC processors.

Register Bandwidth Optimization

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 accelerates computationally intensive operations such as risk-management calculation and AI inferencing.

4th Generation AMD EPYC processors have double the intake floating point of 2nd Generation AMD EPYC processors. An appealing aspect of accelerating performance through increased register bandwidth for floating-point numbers is that it is completely transparent to existing applications and AI models. Thus, financial organizations using 4th Generation AMD EPYC processors can see immediate benefits to the performance of their current applications without requiring any recoding or recompiling.

INT8 Inferencing

4th Generation AMD EPYC processors can accelerate AI inferencing by being able to use AI models quantized to use 8-bit integers instead of floating-point numbers. Moving to 8-bit integer (INT8) inferencing can help with both memory management and processing time. For memory management, a single 32-bit floating-point number takes up the same memory footprint as four 8-bit integer numbers. To accelerate processing time, the increased bandwidth in 4th Generation AMD EPYC processors can perform the same operation on multiple 8-bit integers in a single processor clock cycle.

AMD® Technologies for Confidential Computing in Finance

In addition to generation-on-generation improvements to performance and register size, 4th Generation AMD EPYC processors—and AMD EPYC processors in general—provide numerous security features that are especially important for financial firms to help secure server memory, virtual machines (VMs), and the server-boot process.

AMD® Secure Memory Encryption (AMD® SME)

System memory can represent a vulnerability for financial 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 personally identifiable information (PII) data at risk. This risk can be particularly acute in case of attacks on the hardware itself that seek to recover information from system memory, such as from computers that are physically accessed in or taken from a bank branch office, for example.

AMD helps address this vulnerability by transparently encrypting system memory for operating systems. AMD does this with AMD SME®, which enables financial firms to protect their data in system memory without having to make any changes to their applications.

AMD® Secure Encrypted Virtualization (AMD® SEV)

VMs play an increasingly important role in financial services for efficiency and cost. However, they can also pose a new avenue of attack on financial firms. For example, in 2021, researchers at Symantec published evidence that ransomware attackers had started using VMs to help prevent discovery of their malware after encryption had begun.[3] One way to help limit the threats that target VMs is to isolate guest operating systems and hypervisors from one another.

AMD provides hardened defense for VMs in two ways. At the most general level, AMD SEV® helps ensure that the respective pages in system memory are encrypted so that VMs and hosts cannot directly access each other's data in memory. At a deeper level, 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 and can help defend against threats such as Meltdown and Spectre.

AMD® 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. AMD® 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.

Hardware-Based Security Enhancements from Dell Technologies

Our research investigated what security enhancements PowerEdge R6625 servers—and Dell™ servers more generally—provide to help protect financial firms' infrastructure against threats to firmware, the boot process, and the server-assembly process.[4] These features from Dell Technologies cover the core aspects of the U.S. National Institute of Standards and Technology (NIST) Cybersecurity Framework: Protect, Detect and Recover.

Firmware Protection

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.

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 operating system (OS)—and thus the software-based security running on that OS—has even started. To head off these attacks, the root of trust enabled by iDRAC provides server processors with a read-only encryption key that verifies that the BIOS drivers are legitimate. Cryptographically verified trusted booting like this helps meet NIST recommendations for BIOS protection for servers and BIOS-integrity measurement.

Secure Supply Chains

The most fundamental attack vector on financial organizations' servers is during manufacturing and shipping. This is the stage at which hardware and firmware components can be altered in ways IT organizations cannot detect. Dell Technologies defends against this avenue of attack through the Dell™ Secured Component Verification (SCV) tool by working to ensure that there is no tampering with products or insertion of counterfeit components before shipping products to customers. Dell Technologies manufacturing controls span supplier selection, sourcing, production processes, and governance through auditing and testing. Material inspections during production also help identify components that are mismarked, that deviate from normal performance parameters, or that contain an incorrect electronic identifier.

Upgrading Hardware Yields Important Results for Finance

Financial firms have traditionally been at the vanguard of IT, and continuing trends in the industry are keeping them there. On the one hand, the increasing computational needs of financial applications dictate continued investments in computational power for IT organizations in finance. On the other hand, the financial sector's continued prominence as targets for attack, in addition to the high level of regulatory oversight of the industry, forces financial firms to constantly think about security for their IT infrastructures.

Testing by Prowess Consulting demonstrates that just upgrading from servers powered by 2nd Generation AMD EPYC processors to servers powered by 4th Generation AMD EPYC processors can provide substantial benefits to computational performance. Coupled with this generation-on-generation performance boost are other performance and security features built into AMD EPYC processors and PowerEdge servers. Clearly, not only does processor and hardware selection matter for financial performance, but the choice of server vendor can also carry ramifications for server security.

[1] To test processor performance, Prowess Consulting used a benchmark similar to TPCx-AI from the Transaction Processing Performance Council (TPC®).
[2] 4th Generation AMD EPYC 9554 processors in PowerEdge R6625 servers achieved a SPECrate 2017 Integer score of 1300 and a SPECrate 2017 Floating Point score of 410; 2nd Generation AMD EPYC 7702 processors in PowerEdge R6525 servers achieved a SPECrate 2017 Integer score of 605 and a SPECrate 2017 Floating Point score of 192. Standard Performance Evaluation Corporation. "SPEC CPU2017 Results." https://www.spec.org/cpu2017/results/.
[3] Symantec. "Ransomware: Growing Number of Attackers Using Virtual Machines." June 2021. https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/ransomware-virtual-machines.
[4] For more information, see Dell Technologies. "Technical White Paper: Cyber Resilient Security in Dell EMC PowerEdge Servers." December 2020. https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/cyber-resilient-security-with-poweredge-servers.pdf.

0523/220062