Technical Research Study

PROWESS

Evolving Needs for On-Premises Servers in Healthcare

Healthcare organizations need infrastructure that can securely meet the performance requirements of electronic health record (EHR) systems and artificial intelligence (AI) workloads, both now and into the future.

Executive Summary

Healthcare organizations have unique infrastructure needs due to their diverse workloads and regulatory requirements. For example, organizations typically run database workloads to support electronic health records (EHRs). They also need to support hundreds or thousands of employees accessing confidential data from their PCs, and they face unique emerging artificial intelligence (AI) workloads that require specialized performance. Across all these scenarios, security is paramount for healthcare organizations trying to protect confidential medical and billing records in the face of constant and growing cyberattacks.

With this research note, Prowess Consulting aims to help healthcare organizations plan for future hardware investments by making sure that the servers they consider can meet their 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 also helping ensure security and privacy for users and data.

Industry Landscape

Healthcare organizations rely heavily on on-premises infrastructure to support critical services. There are several reasons for keeping servers and data on-site instead of in a public or hybrid cloud. For example, it can be easier for organizations to meet Health Insurance Portability and Accountability Act (HIPAA) requirements and other regulatory requirements when they have tighter physical control of their hardware and software. Organizations also might realize performance or cost advantages by optimizing server configurations and density on-site.

Those in charge of planning and purchasing on-premises infrastructure need to address several key challenges and considerations. For example, what servers are best suited to hosting the Epic® Systems software and databases that manage electronic medical and billing records and that track patient health information? What systems are ideal for hosting employee virtual desktop infrastructure (VDI) sessions with the required levels of performance and security? If the organization wants to implement AI and machine learning (ML) workloads that help improve diagnostics, treatment, and billing, what infrastructure will properly support the inferencing needs of those AI workloads? And finally, how can the organization help ensure strong levels of infrastructure security and privacy for medical records, billing information, and other sensitive data?

This paper briefly examines these topics under three broad areas: performance considerations, AI considerations, and security requirements.

Performance Considerations

Healthcare organizations typically rely on Epic Systems or similar EHR solutions. In fact, 45 percent of the US population has its medical records in an Epic electronic healthcare system.[1]

Due to regulatory requirements and the need for privacy and security, healthcare organizations often deploy and manage their employees' operating systems and software via VDI sessions. VDI allows doctors and other healthcare workers to access patient records from any device while keeping all compute and storage inside the data center. But this model relies on servers that can support large numbers of VDI sessions with performance and security. Organizations need to ensure that the underlying hardware they employ is built with technologies that support and accelerate performance for VDI. Budget limitations can also weigh on purchasing decisions for servers supporting VDI. Administrators strive to reduce total cost of ownership (TCO) by maximizing consolidation of VDI sessions on each server. And admins need to ensure strong levels of security with isolation of individual VDI sessions.

VDI Performance Metrics

For maximizing VDI performance and consolidation, several key metrics come into play:

- High core counts
- High-frequency CPUs
- High memory bandwidth
- PCIe® 4.0 to support modern NVM Express® (NVMe®) solid-state drives (SSDs) and graphics processing units (GPUs)
- Support for physical GPUs that can provide a strong allocation of cores/memory to virtual GPUs (vGPUs) to better support virtual clients

Servers powered by recent-generation processors generally offer greater VDI performance and consolidation compared to legacy systems, which can lead to a lower TCO. For example, a recent study demonstrated a gain of 20 percent more VDI users and 11 percent lower cost per VDI user for a Dell™ PowerEdge™ R7515 server powered by a 3rd Generation AMD EPYC™ 75F3 processor, compared to the same server with a 2nd Generation AMD EPYC processor.[2]

Login VSI™ benchmarking tests provide a widely recognized method for evaluating and comparing VDI performance between systems. Login VSI simulates typical workloads, and it then measures latency as the number of active VDI sessions is increased. This continues until no additional users can be added without degrading the user experience.

Leading hardware and CPU vendors often provide Login VSI numbers that can help organizations with their purchasing decisions. For example, AMD published a technical data sheet touting that its AMD EPYC 7763 processor enables up to 2.1x more VDI sessions on Login VSI, compared to an Intel® Xeon® Gold 6258R processor.[3] Similarly, AMD claims the ability to get up to 46 percent more "knowledge worker" desktop sessions per core with the AMD EPYC 7543 processor versus the competition.[4]

Database Performance Metrics

Database performance is also critical for generating scheduled and ad-hoc reports related to patient care, contact information, public health data, and more. Many healthcare organizations rely on the Epic® Clarity® database, which runs on Oracle® Database or Microsoft® SQL Server®. These databases can be many terabytes—or even exabytes—in size, depending on the size of the organization. Additionally, organizations will often have both test/dev and production environments operating in tandem.

When evaluating systems for database performance, healthcare organizations should also consider the types of transactions that will dominate in general usage. For many clinics and hospitals, that will equate to frequent, small reads and occasional small writes as healthcare and billing records are accessed throughout the day. Large writes are typically batch processed in off hours, when performance is less likely to be a factor.

The key considerations for systems providing database support are similar to the ones called out for VDI: higher core counts for parallel processing, high-frequency CPUs for supporting relational databases like SQL Server, fast memory bus, and fast PCIe interconnect performance, such as with PCIe 4.0.

One way to evaluate systems for database performance in advance of a purchase is to consult TPC® benchmark results online. A quick review of published performance results shows AMD® processor–powered Dell™ servers taking four of the top five spots for TPC Express Benchmark V (TPCx-V)—a virtualization benchmark for database workloads.[5] In addition, Dell servers occupy four of the five top spots for performance measured with TPC-H—a decision-support benchmark built on a suite of business-oriented ad-hoc queries and concurrent data modifications.[6]

AI Considerations

AI workloads are growing in popularity with healthcare organizations, and for good reasons. AI provides opportunities to streamline processes, improve the accuracy of medical billing and diagnostics, and potentially reduce imaging diagnostics backlogs.

Examples of AI in Healthcare

AI is frequently used to help in imaging analysis. Organizations can rely on AI inferencing to correlate past diagnostics, medical procedures, lab results, medical history, allergies, and other patient data into a summary for radiologists and cardiologists.[7] This capability can provide context for recent radiological images that can help physicians improve accuracy for diagnostics.

AI can also be used to analyze unstructured healthcare records and biomedical data to help providers make faster, more accurate, and tailored treatment plans for patients.[7] And there are several other emerging use cases for AI in healthcare, including forecasting kidney disease, assessing quality of care and outcomes in cancer treatment, or helping predict a disease at an earlier stage, before it becomes life threatening.[7]

Regardless of the specific use case, healthcare AI workloads all require processors designed to deliver optimum inferencing performance. That's because—unlike training, which is often GPU-intensive—inferencing is an AI operational phase in which CPUs can play a key role.

Register Bandwidth Optimization for Inferencing

AI inferencing performance is driven by how much calculating the processor register can do. The register is the location within a processor that holds the instructions, storage addresses, and individual numeric data necessary for computation.

One way to speed up AI inferencing, then, is to increase the amount of numeric data that a processor can calculate in a clock cycle (in other words, the bandwidth of the processor). Increasing the quantity of floating-point numbers on which processors can calculate automatically speeds 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.

One advantage to this approach is that it is completely transparent to existing AI models. Therefore, if a healthcare organization wanted to speed up inferencing for its diagnostics imaging AI models, it could achieve immediate results by using newer servers with processors built to accommodate more floating-point numbers in their registers.

Another way to accelerate inferencing is to quantize AI models to use 8-bit integers (INT8) instead of floating-point numbers. However, this method of acceleration does require some reworking of existing AI models. Using INT8 can substantially increase inferencing speed while incurring minimal loss to accuracy.[8] 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.

If a healthcare organization wanted to speed up inferencing for its diagnostics imaging AI models, it could achieve immediate results by using newer servers with processors built to accommodate more floating-point numbers in their registers.

PCIe® 4.0

A third way to speed up AI workloads is to use inference accelerators, which are specialized processors specifically designed for AI inferencing. To make full use of these accelerators requires high bandwidth for the connection between the accelerator and the server processor. Organizations can increase this bandwidth by ensuring their servers include PCIe 4.0 interconnect protocols. PCIe 4.0 provides 16 gigatransfers per second (GT/s), compared to 8 GT/s in PCIe 3.0, which can translate to faster data transfers to inference accelerators.

Examples of server processors that support PCIe 4.0 include 2nd Generation AMD EPYC processors (and later) and 3rd Generation Intel Xeon Scalable processors. A healthcare organization seeking to accelerate the performance of its AI models that forecast kidney disease, for example, could use AI inference accelerators along with PCIe 4.0 interconnects between processors and accelerators.

Security Requirements

There is no shortage of cyberattacks on healthcare systems. According to cybersecurity firm Sophos, an astonishing 34 percent of healthcare organizations were hit by ransomware in 2020.[9] And according to a study by the CyberPeace Institute, in the period between June 2020 and September 2021, more than 10 million healthcare records were stolen. These records included social security numbers, patient medical records, financial data, HIV test results, and the private details of medical donors.[10]

Healthcare organizations can strengthen their security defenses by employing software-based tools to help with threat detection, response, and recovery. But to fully embrace cybersecurity best practices established by the National Institute of Standards and Technology (NIST) Cybersecurity Framework, healthcare organizations need to take a more proactive approach. Modern security features, such as confidential computing, go beyond software-based techniques by extending protections into the hardware layer and even into the supply chains of hardware suppliers.

To fully embrace cybersecurity best practices established by the National Institute of Standards and Technology (NIST) Cybersecurity Framework, healthcare organizations need to take a more proactive approach. Modern security features, such as confidential computing, go beyond software-based techniques by extending protections into the hardware layer and even into the supply chains of hardware suppliers.

Confidential Computing and Hardware-Enhanced Security

Confidential computing is an emerging initiative that focuses on improving isolation for sensitive data payloads, securing data while at rest, in motion, and in use. Storage and network encryption play key roles in protecting health data, but confidential computing also relies heavily on hardware-based memory protections and additional technologies for securing servers during the boot process.

Memory Encryption

System memory can represent a vulnerability for data. While data might be encrypted at rest in storage and in motion across the network, it typically needs to be decrypted into memory for use by applications. This can put sensitive healthcare information at risk, particularly from attacks focused on accessing system memory from physical devices that are lost or stolen.

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.

Developers of healthcare applications can go one step further, modifying applications to use a trusted-execution environment (TEE), such as with Intel® Software Guard Extensions (Intel® SGX), to prevent data or code in an application from being altered.

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.[11] Because VMs play an increasingly important role in healthcare organizations for efficiency and cost reasons, protecting or isolating VMs is critical to safeguarding healthcare data.

One way to help limit threats to 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 can ensure 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. This technique 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.

Silicon Root of Trust and Secure Boot

Firmware is an attractive attack vector because it provides a way to compromise servers while they are booting, before software-based malware defenses even have a chance to start running. To head off these attacks, server processors require a read-only encryption key that validates that the BIOS or Unified Extensible Firmware Interface (UEFI) drivers are legitimate. This type of cryptographic verification 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®. This encryption key must be burned into the silicon during the manufacturing process. Examples of this include the root of trust enabled by Integrated Dell™ Remote Access Controller (iDRAC) and HPE® Project Aurora.

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 iDRAC can verify the integrity and authenticity of the BIOS image when a server is powered on, which can help protect systems from attacks that manipulate BIOS as a way to alter firmware. In addition, dynamic system lockdown, such as that provided by iDRAC, can even help prevent 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 compromised firmware and configuration settings from one server to another, which can lead to additional vulnerabilities on the other servers.

Secure Supply Chains

Another vector of attack is one that often goes unnoticed by healthcare IT admins: manufacturing supply chains. During manufacturing and shipping, hardware and firmware components can be altered in ways customers can't 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, that deviate from normal performance parameters, or that contain an incorrect electronic identifier. Healthcare organizations can better protect themselves by seeking out vendors, such as Dell Technologies, that provide Secured Component Verification (SCV) on top of a secured supply chain.

Accelerate AI and EHR Workloads with Confidence

Healthcare organizations have several reasons for deploying servers on premises, from regulatory and privacy requirements to improved on-site performance to cost and management considerations. Before purchasing servers, several factors need to be carefully considered. These include providing high performance for databases and AI inferencing workloads and strengthening protections from sophisticated and growing malware.

Systems with high core counts and core frequency, high memory bandwidth, and PCIe 4.0 can significantly boost performance. Specialized processor instruction sets and other inferencing accelerators are crucial to meeting healthcare AI needs. And strong hardware-enhanced protections in conjunction with supply-chain integrity can help reduce risk to healthcare organizations.

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

[1] Johns Hopkins Medicine. "Epic at Johns Hopkins Medicine." Accessed April 20, 2022. **www.hopkinsmedicine.org/epic/why_epic/**.

[2] Principled Technologies. "Support more VDI users with a Dell EMC PowerEdge R7515 server powered by an AMD EPYC 75F3 processor." Commissioned by Dell Technologies. March 2021. **www.delltechnologies.com/asset/en-us/products/servers/industry-market/support-more-vdi-users-with-a-dell-emc-poweredge-r7515.pdf**.

[3] AMD. MLN-004. Login VSI™ Pro v4.1.40.1 comparison based on AMD internal testing as of 02/01/2021 measuring the maximum "knowledge worker" desktop sessions within VSI Baseline +1,000 ms response time using VMware ESXi™ 7.0u1 and VMware Horizon® 8 on a server using 2 x AMD EPYC™ 7763 processors versus a server with 2 x Intel® Xeon® Gold 6258R processors for ~112 percent more max [~2.1x the] performance. Results may vary. **www.amd.com/en/claims/epyc#faq-MLN-004**.

[4] AMD. MLN-005. Login VSI™ Pro v4.1.40.1 comparison based on AMD internal testing as of 02/01/2021 measuring the maximum "knowledge worker" desktop sessions within VSI Baseline +1,000 ms response time divided by the number of cores using VMware ESXi™ 7.0u1 and VMware Horizon® 8 on a server using 2 x AMD EPYC™ 7543 processors versus a server using 2 x Intel® Xeon® Gold 6258R processors for ~46 percent more [~1.5x the] performance. Results may vary. **www.amd.com/en/claims/epyc#faq-MLN-005**.

[5] TPC. "TPCx-V Top Performance Results." Results shown as of April 29, 2022. **www.tpc.org/tpcx-v/results/tpcxv_perf_results5.asp**.

[6] TPC. "TPC-H V3 All Results." Results shown as of April 29, 2022. **www.tpc.org/tpch/results/tpch_results5.asp**.

[7] MobiHealthNews. "Contributed: Top 10 Use Cases for AI in Healthcare." July 2021. **www.mobihealthnews.com/news/contributed-top-10-use-cases-ai-healthcare**.

[8] 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." Accessed April 27, 2022. **https://docs.openvino.ai/latest/openvino_docs_performance_int8_vs_fp32.html**.

[9] Sophos. "The State of Ransomware in Healthcare 2021." May 2021. **https://assets.sophos.com/X24WTUEQ/at/s49k3zrbsj8x9hwbm9nkhzxh/sophos-state-of-ransomware-in-healthcare-2021-wp.pdf**.

[10] CyberPeace Institute. "If healthcare doesn't strengthen its cybersecurity, it could soon be in critical condition." November 2021. **www.weforum.org/agenda/2021/11/healthcare-cybersecurity**.

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