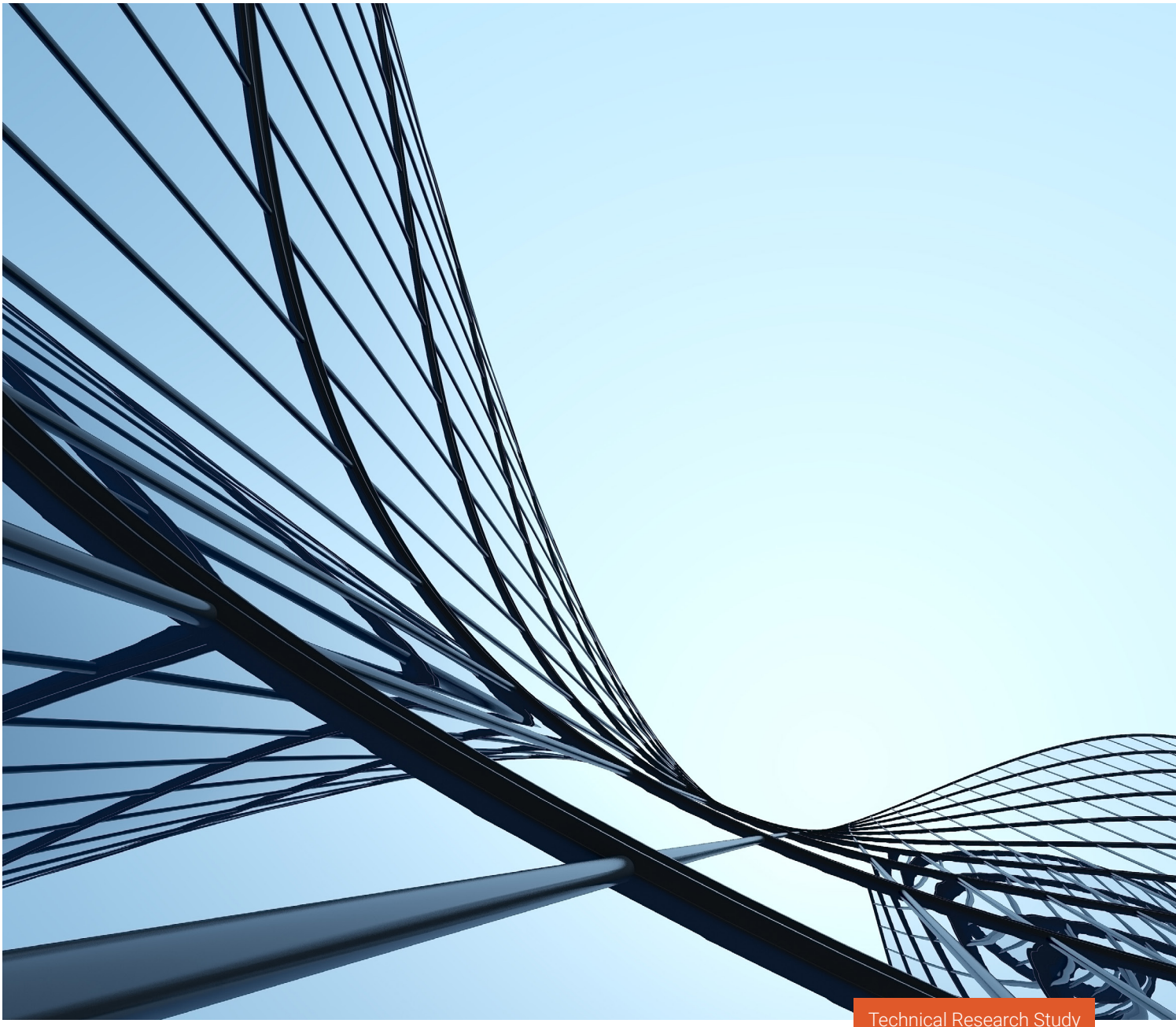

Technical Research Study

PROWESS

World-Record Performance for AI and ML

Prowess research draws links from the benchmark strength of Dell™ servers to real-world performance for training and inferencing on artificial intelligence (AI) and machine learning (ML) models.

Executive Summary

Organizations from a wide array of industries depend on artificial intelligence (AI) and machine learning (ML) to forecast sales, segment their customer bases, identify risks, manage complex supply networks, optimize costs, and improve efficiencies. However they are applied, all AI and ML use cases depend on compute performance (in addition to the speed and quantity of memory and the bandwidth of interconnects and networking). And due to the nature of the data companies use for activities like predictive customer analytics or fraud detection, security is often just as critical.

Because of the performance needs of AI and ML, the infrastructure from compute, local memory, network bandwidth, and data storage to support these workloads can represent a significant investment, which drives the need for rigorous evaluation before purchase. Industry-standard benchmarks can be good for this—and world records can be even better—if they are evaluated in the right way.

Companies need to rapidly ingest, store, and process their data. These benchmarks can provide insights into how quickly data can be collected, processed, and accessed once stored. In order to investigate the relationship between high benchmark performance and potential business value in the real world, Prowess Consulting dug deeper into what strong showings in industry benchmarks can mean for businesses deploying world-record servers. Because of its outsized market share and the number of world records Dell Technologies holds in AI and ML scenarios, represented by industry-recognized benchmarks, Prowess specifically looked at Dell™ PowerEdge™ servers.

Among the newest-generation Dell PowerEdge servers, we identified three that have recently set world records in a benchmark specifically designed to provide real-world performance data on AI and ML workloads:

- **Dell PowerEdge R6625 server (world record, TPC Express Benchmark™ AI [TPCx-AI])**
- **Dell PowerEdge R7615 server (world record, TPCx-AI)**
- **Dell PowerEdge R7515 server (world record, TPC Express Benchmark™ IoT [TPCx-IoT])**

Augmenting performance optimizations from Dell Technologies for AI and ML workloads, we also found that AMD EPYC™ processors and Broadcom® network cards can help drive big-data and analytics performance both for benchmarks and real-world applications. For example, support for the bfloat16 numeric format in 4th Gen AMD EPYC processors enables running larger AI models with bigger datasets, while support of INT8 numbers can accelerate inferencing on AI models. In addition, dual NVM Express® (NVMe®) Dell™ PowerEdge™ RAID Controller (Dell™ PERC) cards and support for PCIe® 4.0 with Broadcom network cards enable dual-port 100 gigabit Ethernet (GbE) network interface controllers (NICs) and help eliminate bandwidth constraints and further accelerate AI and ML workload performance.

This study covers the following topics:

Industry Landscape: AI and ML

AI and ML have come to fill a central role in business operations for organizations in a variety of industries. Whether it's a bank monitoring for fraud, a retailer projecting sales, a hospital striving for more accurate diagnoses, or a mid-size manufacturer implementing predictive maintenance on its assembly lines, organizations of all types and sizes rely on AI to tease out patterns that might be invisible to people. AI models in general—and deep learning (DL) models in particular—are generally more accurate when there is more data available to train them. This appetite for data drives the need for larger, more capable storage, faster networking, and more performant servers to find more value in data—data which must also be kept secure.

Businesses often rely on on-premises servers rather than cloud implementations for a variety of reasons. For AI and ML, these reasons often center on data gravity and lower latency for using AI models. It is often faster and easier to bring AI training functions closer to the data rather than bear the cost and time of moving large amounts of data to centralized compute. Working with data close to where it resides can also reduce the latency for training AI, which can speed up the process. Moreover, regulatory requirements and data-sovereignty laws can also be compelling reasons to keep data on premises, depending on an organization's industry and location. In all cases, performance is a core requirement for businesses when dealing with their data and tapping the analytical value that data contains through AI and ML.

The performance demands of workloads like analytics mean that the data infrastructure must be tuned to meet service-level agreements (SLAs). The interplay of processor, memory size, network bandwidth, and storage subsystems is critical. One prominent tool for comparing server performance for this interplay is benchmark results. Because benchmarks produce numeric results, comparisons between competing systems can feel straightforward.

Precisely because benchmarks produce clear and seemingly objective results, however, understanding what they measure—and thus what they actually say about server platforms—is crucial. Organizations that ignore the nuance of these benchmarks and blindly chase the top benchmark performers can wind up disappointed when their return on investment (ROI) fails to meet their expectations.

Prowess Research Methodology

In order to investigate the relationship between high benchmark performance and potential business value in the real world, Prowess Consulting dug deeper into what strong showings in industry benchmarks can mean for businesses deploying world-record servers. To simplify our investigation and focus on how individual benchmarks can provide insights into performance in particular facets of AI and ML, we specifically looked at Dell PowerEdge servers. We did this both because of the large market share for Dell Technologies servers and because of the number of world records Dell Technologies holds across a variety of AI and ML benchmarks.

A single benchmark world record is impressive, but what stands out for AI and ML workloads is that Dell platforms achieve world records across multiple benchmarks. Each benchmark can be viewed as a piece of the workload puzzle, and the achievement of multiple world records provides good insight into how Dell platforms will operate in real-world environments.

We looked at benchmark results to help determine which platforms offered the best performance for different aspects of AI and ML workloads. Our research focused on Dell™ rack-mount servers. Dell Technologies has the largest market share of servers worldwide (17.2 percent),[1] and Dell PowerEdge servers are popular workhorse servers, built for standard to medium-heavy workload needs. Specifically, for this study we examined 1U (Dell PowerEdge R6625 server) and 2U (Dell PowerEdge R7615 and PowerEdge R7515 server) rack-mount platforms.

When examining the benchmark results, it is essential to view them through the lens of the most important performance factors. For AI and ML workloads, these include:

- Performance
- Price/performance

Dell Technologies has optimized its PowerEdge platform, based upon AMD® processors, for AI and ML solutions. AI and ML workloads can have specific needs, and PowerEdge servers enable a high degree of flexibility for customers to run their unique workloads according to their specific requirements. In addition, Dell Technologies has optimized performance for its PowerEdge platforms by incorporating newer Dell PERC cards and Broadcom NVMe network adapters that bring significant bandwidth improvements to the table.

AI and ML Benchmarks

Industry-recognized benchmarks can provide insights into common uses of a server platform, and they can help inform customers on whether that platform will meet the needs of the workloads the customer is running. For this study, we specifically looked at the benchmarks most directly applicable to AI and ML workloads.

Note: MLPerf™ Inference Benchmark Suite is commonly used to measure ML performance. Prowess opted not to use MLPerf for our analysis because that benchmark focuses on graphics processing unit (GPU) performance, whereas our study focused on investigating cost-effective CPU performance for AI inferencing.

TPC Express Benchmark™ AI (TPCx-AI)	Measures end-to-end performance of industry-representative AI and ML workloads
TPC Express Benchmark™ IoT (TPCx-IoT)	Measures performance, price-performance, and availability for systems that ingest massive amounts of data from large numbers of devices

TPC Express Benchmark™ (TPCx-AI)

TPCx-AI measures the performance of an end-to-end ML or data science platform. The benchmark is designed to emulate the behavior of representative industry AI solutions that are present in production data centers and cloud environments.

TPCx-AI assesses AI performance through a number of 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. Use cases in TPCx-AI generally include data generation, data management, training, scoring, and serving phases.

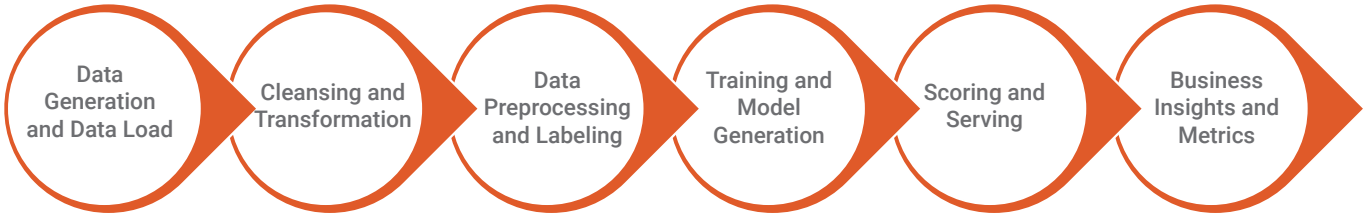

Figure 1. The TPC Express Benchmark™ AI (TPCx-AI) DL and ML data science pipeline

Use cases in TPCx-AI include customer segmentation, customer-conversation transcription, sales forecasting, spam detection, price prediction, classification, and fraud detection.

Dell Technologies has four world records for **TPCx-AI** running 4th Gen AMD EPYC processors:[2]

- Dell PowerEdge R7615 server (at scale factor 3)
- Dell PowerEdge R7615 server (at scale factor 10)
- Dell PowerEdge R6625 server (at scale factor 30)
- Dell PowerEdge R6625 server (at scale factor 100)

PowerEdge R7615 servers achieved AI pipeline throughput of 408.36 GB/second at scale factor 3 (3 GB of data) and 425.31 GB/second at scale factor 10 (10 GB of data), with a price-performance ratio of $118.56/GB/second at scale factor 3.[2]

The PowerEdge R6625 server achieved 365.59 GB/second at scale factor 30, while a cluster of four PowerEdge R6625 servers reached 868.49 GB/second at scale factor 100, with price-performance ratios of $196.38/GB/second and $356.56/GB/second, respectively.[2]

These servers are powered by 4th Gen AMD EPYC 9374F, EPYC 9174F, and EPYC 9354 processors. In addition to having a higher core and thread count than the preceding generation, 4th Gen AMD EPYC processors also provide DDR5 memory support for more memory, PCIe Gen 5 support for higher data throughput within servers and server clusters, support for the latest Advanced Vector Extensions 512 (AVX-512) for faster data throughput in the processor register, and support for efficient numeric types (such as bfloat16 and INT8) to speed up model training and inferencing. All of these features helped the Dell servers at every stage of the data science pipeline in TPCx-AI.

Internet of Things (IoT): TPC Express Benchmark™ IoT (TPCx-IoT)

The Internet of Things (IoT) represents a large source of data for AI and ML, particularly for real-time inferencing by AI models, such as those performing quality-control checks on manufacturing lines. IoT also represents an essential source of training data for AI models. This applies to initial training or new models, but also to retraining existing models that have drifted.

Though not part of the data science pipeline encapsulated in the TPCx-AI benchmark, continuous training of AI models in production is essential. Over time, the composition of data in production environments such as retail, finance, or healthcare can change in subtle ways. This can cause the accuracy of AI models to steadily decrease. Efficiently capturing updated training data in order to correct the drift of AI models is a crucial part of any AI or ML workflow.

What sets IoT apart from other kinds of unstructured data is that IoT devices often generate their data from their environment, such as smart speakers listening to commands or industrial drones collecting agricultural land data. IoT data thus needs to be efficiently captured and used for AI and ML workloads because of its sheer volume, which is typically greater than that of other data sources, such as retail transactions or healthcare records.

The TPCx-IoT benchmark enables direct comparison of different software and hardware solutions for IoT gateways. Because they are positioned between edge architecture and the back-end data center, gateway systems perform functions such as data aggregation for real-time AI and ML. The TPCx-IoT benchmark was specifically designed to provide verifiable performance, price-performance, and availability metrics for commercially available systems that typically ingest massive amounts of data from large numbers of devices.

Dell Technologies holds a world record for **TPCx-IoT** for Dell PowerEdge 7515 servers running 3rd Gen AMD EPYC processors. The cluster achieved throughput of 1,617,545,000 records/second at a price/performance ratio of $329.75/million records/sec.[3]

In addition to the processing power provided by the 3rd Gen AMD EPYC 75F3 processor, our analysis of the benchmark results also indicates that the performance of the Dell servers was boosted by the Broadcom 25 GbE network cards used in the cluster. All of these features helped the Dell servers attain their performance in the TPCx-IoT benchmark, with the Dell Technologies results being especially relevant for organizations that use Cloudera®—as the Dell Technologies record-holder did—to store large amounts of data on the edge for analytics and AI.

Behind the Performance Results

Because Dell Technologies has optimized its PowerEdge platform based upon AMD processors, it has achieved numerous world records in benchmarks measuring AI and ML performance. In an on-premises implementation, 3rd Gen AMD EPYC processors already offered strong performance, performance per watt, and performance per CPU dollar. In the cloud, AMD EPYC systems on chip (SoCs) power high-performance computing (HPC)-optimized infrastructure-as-a-service (IaaS) instances for many cloud service providers (CSPs) including Amazon Web Services® (AWS®), Microsoft® Azure®, Google Cloud Platform™, and others.

The 4th Gen AMD EPYC processors used in Dell's world-record server configurations for TPCx-AI provide performance gains that can be traced to several platform improvements over the previous-generation platform, including:

- 50 percent increase in core count,[4] increased thread count, and higher frequencies, which can directly increase processing performance.
- 12 DIMMs/socket (up from 8), which allows organizations to significantly increase available memory. This translates to processing larger datasets faster, particularly for in-memory analytics such as those performed by Apache Spark™.
- DDR5 memory support for faster access to data.
- AVX-512 support, which enables 4th Gen AMD EPYC processors to complete more simultaneous calculations in their registers.
- Greater L2 cache, doubled from 512 KiB to 1 MiB per core, which also accelerates operations in memory.
- PCIe Gen 5 support, which enables faster interconnects to move more data with lower latency.
- Enhancements specifically for AI and ML workloads, including support for the bfloat16 numeric type to accelerate the training of AI models and support for INT8 inferencing to increase the performance of already trained models in production.

Overall, 4th Gen AMD EPYC processors operate more efficiently than their predecessors. The Standard Performance Evaluation Corporation's SPEC CPU® 2017 Floating Point Rate results show a gain in performance of 121 percent in tests run on a system powered by 4th Gen AMD EPYC processors, compared to a system powered by 3rd Gen AMD EPYC processors.[5] The SPEC CPU 2017 Integer Rates results showed gains of 102 percent.[6] These processor performance results are reflected in the world-record benchmark results achieved by several of the PowerEdge platforms we examined.

The number of cores in these processors increased by 50 percent, compared to the previous generation, which also boosts performance. At the same time, published specifications from AMD show an increase in maximum default power consumption of only 42 percent, from 280-watt thermal design power (TDP) to 400-watt maximum TDP.[7] When compared to the SPEC performance results above, these power numbers show the capability for servers built on 4th Gen AMD EPYC processors to provide up to a 55 percent power-performance benefit for businesses running AI and ML workloads.[8]

The TPCx-AI benchmark, which measures AI and ML performance, also reflects the performance brought to the table by the Broadcom network adapters and Dell PERC cards. Organizations deploying AI and ML workloads require RAID controllers for redundancies to meet internal or regulatory requirements. The PowerEdge R6625 and PowerEdge R7615 servers that set the TPCx-AI world records were outfitted with Dell PERC cards with fast NVMe RAID support.

AMD® Hardware-Based Security

For all of the workloads evaluated in this research study, security considerations are critical. 3rd Gen AMD EPYC™ and 4th Gen AMD EPYC processors can provide hardware-based security for AI and ML workloads. AMD® Secure Memory Encryption (AMD® SME) encrypts system memory to protect data in use. AMD® Secure Encrypted Virtualization (AMD® SEV) protects running virtual machines (VMs) so that they are encrypted and isolated from each other and the host-system hypervisor. AMD® Secure Encrypted Virtualization-Encrypted State (AMD® SEV-ES) encrypts the CPU register contents of stopped VMs to help protect the data stored in them. And AMD® Secure Boot helps protect servers during the boot process, providing defenses against rootkits, bootkits, and firmware while servers are most vulnerable.

Network cards from Broadcom speed up the flow of data for AI/ML workloads. Support for PCIe Gen 5 in both 4th Gen AMD EPYC processors and Broadcom NICs allows the use of 100 GbE Broadcom network adapters built on the Open Compute Project (OCP) NIC 3.0 form factor. These modern designs reflect a rapid shift in the industry toward 100 GbE adapters built on a more efficient form factor and enabled by PCIe 4.0 and PCIe 5.0. In addition, support for PCIe 4.0 and PCIe 5.0 can provide performance numbers from a single NIC that are on par with dual 100 Gbps NICs. The OCP NIC 3.0 specification enables server manufacturers like Dell Technologies to use more compact designs that can support high-performance adapters with advanced hardware-acceleration capabilities to further speed up AI and ML workloads.[9]

The Dell™ PowerEdge™ RAID Controller (Dell™ PERC) Protects Data and Boosts Storage Performance

Modern PCIe® Gen 4 RAID interfaces work with high-bandwidth NVM Express® (NVMe®) solid-state drives (SSDs) to significantly boost storage performance. Dual Dell PowerEdge RAID Controller 11 and 12 (PERC 11 and PERC 12) cards and NVMe adapters with both PCIe Gen 4 host and PCIe Gen 4 storage interfaces can help remove bandwidth and latency constraints.

Conclusion

Benchmark results in general (and world-record results in particular) are about more than bragging rights for server manufacturers. Interpreted correctly, best-in-industry results in benchmarks can offer insights as to how servers could perform in real-world use cases. Because of the Dell Technologies market share and the number of world records the company holds, its PowerEdge servers provide a natural opportunity to examine how benchmark results can map to performance benefits for organizations in production. While no mapping of benchmark performance (world record or otherwise) is 1:1, our investigation shows that the performance of Dell PowerEdge servers in an industry-recognized benchmark indicates strong AI and ML performance across several use cases in a variety of industries. Moreover, the number of world records held by Dell Technologies across different benchmarks demonstrates how the company has developed platforms that take advantage of the individual components' strengths in order to deliver real value to its customers for a variety of workloads.

Appendix A: Benchmark Performance Links

- TPCx-AI top performance results: **www.tpc.org/tpcx-ai/results/tpcxai_perf_results5.asp**
- TPCx-IoT V2 top performance results: **www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2**

Appendix B: Dell Technologies System-Specification Links

- Dell PowerEdge server specification sheets: **www.dell.com/en-us/dt/servers/poweredge-rack-servers.htm**

[1] History-Computer. "The 10 Largest Server Companies In The World, And What They Do." September 2022. **https://history-computer.com/largest-server-companies-in-the-world-and-what-they-do/**.

[2] TPC. "TPCx-AI Top Performance Results." Accessed November 1, 2022. **www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2**.

[3] TPC. "TPCx-IoT V2 Top Performance Results." Accessed November 1, 2022. **www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2**.

[4] Tom's Hardware. "Zen 4 Madness: AMD EPYC Genoa With 96 Cores, 12-Channel DDR5 Memory, and AVX-512." August 2021. **www.tomshardware.com/news/zen4-madness-amd-epyc-genoa-with-96-cores-12-channel-ddr5-memory-and-avx-512**.

[5] Up to 121 percent higher SPEC® Floating Point performance comparing top-bin 4th Gen AMD EPYC™ processors with top-bin 3rd Gen AMD EPYC processors based on SPEC Floating Point rate score of 1,410 achieved on a Dell™ PowerEdge™ R7625 server powered by AMD EPYC 9654 processors, compared to a score of 636 achieved on a Dell PowerEdge R7525 server powered by AMD EPYC 7763 processors. Scores accessed as of November 10, 2022. See Standard Performance Evaluation Corporation benchmark results. **http://spec.org/benchmarks.html**.

[6] Up to 102 percent higher SPEC® Integer Rate performance comparing top-bin 4th Gen AMD EPYC™ processors with top-bin 3rd Gen AMD EPYC processors based on SPEC Integer rate score of 1,660 achieved on a Dell™ PowerEdge™ R7625 server powered by AMD EPYC 9654 processors, compared to a score of 821 achieved on a Dell PowerEdge R7525 server powered by AMD EPYC 7763 processors. Scores accessed as of November 10, 2022. See Standard Performance Evaluation Corporation benchmark results. **http://spec.org/benchmarks.html**.

[7] AMD. AMD EPYC 7003 Series processors specifications webpage. **www.amd.com/en/processors/epyc-7003-series**.

[8] 55 percent CPU performance per watt improvement calculated using the SPEC® Floating Point score of 1,410 achieved on a Dell™ PowerEdge™ R7625 server powered by AMD EPYC 9654 processors with a processor cTDP of 400 watts, compared to a score of 636 achieved on a Dell PowerEdge R7525 server powered by AMD EPYC 7763 processors with a processor cTDP of 280 watts.

[9] Broadcom. NetXtreme E-Series OCP NIC 3.0 Ethernet Adapters Product Brief. 2021. **https://docs.broadcom.com/doc/12395120**.