Solution Brief

Public Clouds
Amazon EC2

Accelerate Cloud Work in AWS Intel Instances

Amazon Web Services (AWS) cloud instances with 3rd Gen Intel® Xeon® Scalable processors outperform instances with AWS Graviton, especially with Intel acceleration.

Executive summary

- AWS C6i, R6i, and M6i instances with Intel outperform C6g, R6g, and M6g instances with Graviton2 on common workloads
- Built-in accelerators in C6i, R6i, and M6i widen the performance advantage over Graviton2 significantly—even outperforming Graviton3
- C6i, R6i, and M6i deliver better performance with accelerators at a fraction of the cost of Graviton2

C6i R6i M6i

The case for better performance in the cloud

We live in an age of impatience. By some estimates, more than half of all mobile searches are abandoned when a page takes more than three seconds to load. Sluggish performance can also lower a website's search engine results page (SERP) ranking.[1]

To achieve higher SERP rankings for their websites, many Fortune 500 companies use the WordPress content-management system (CMS) because of its search engine optimization (SEO) and support for numerous SEO plug-ins.[2] However, all these plug-ins and scripts running in the background can have the opposite of the desired effect, resulting in slow-loading web pages.

With such high expectations from their customers online, cloud users likewise demand that their hosting providers find ways to accelerate website performance in the back end to meet these demanding service levels. The ability to process search queries with blazing speed is why NGINX has overtaken Apache as the leading web server stack for hosting high-traffic websites.[3]

The demand for fast data delivery keeps going up, along with the volume of data needing to be managed, accessed, and analyzed. Shortening response time is also a challenge for organizations that rely on artificial intelligence (AI) to analyze enormous datasets.

These content-management, web server, and AI operations are conducted millions of times a day,[4] often using virtual data centers in the AWS cloud, the world's largest cloud platform.[5] Some AWS customers use instances with Intel to boost their performance right now. However, others are leaving those benefits on the table because meaningful data that would help to inform decisions is missing.

Intel went after that data by testing performance as any AWS user would experience it. This approach is helpful for analyzing typical workloads in noisy environments with many variables, such as on-demand instances in the cloud. Results are measured as actual operation rates (not scores), such as transactions per second, images processed per second, or connections per second.

WordPress, NGINX, and AI workloads perform better in AWS instances with Intel processors than in AWS Graviton2 instances. Workload performance improves substantially when built-in Intel acceleration technologies and software are enabled.

Defining acceleration in the cloud

Intel acceleration is a more efficient way to achieve higher performance in the public cloud than increasing virtual CPU (vCPU) count, moving into a higher-priced specialized instance, or re-platforming to a different architecture. Intel Xeon processors have a growing number of workload accelerators built into each generation that are integrated into software applications. Ultimately, the result is better workload performance and better price performance (the amount of work that can be accomplished per dollar at an hourly rate) in Intel instances.

Finding the best business value

Business success demands the ability to get to market quickly and deliver consistently as you grow. In the past, the well-worn path to growth was to move to a higher-frequency or higher-core-count CPU. But today's workloads and computing architectures are more intricate. Throwing more muscle at them might not produce the intended outcome, and there are hundreds of different instances to choose from in the cloud. How do you know which offers the best business value for your needs?

A common starting point for many cloud users is to look for the highest performance at the lowest hourly rate. However, while rates are easy to compare before you move, how your applications and code will perform is not.

In looking at performance, you might get bogged down in trades with significant downsides. Purchase more instances in pursuit of performance, and you might end up wasting budget on excess capacity. Sign up for a lower-priced option today, and you might find it costs more tomorrow, when it locks you into a vendor or an architecture that fails to meet your future technical needs. Re-platform with the goal of lowering costs, and you could discover that you have lost flexibility and performance.

Instead, start with applications, service levels, and results that matter to you to select the right hardware platform for your business. Cloud instances with Intel might not top your list based on traditional measures like core counts, chip speeds, and raw computing power, but they offer superior performance on real-world benchmarks and better outcomes in the long run.

Use active benchmarking for a more accurate picture

The days of core counts and frequencies as measures of performance are gone. Instead, highlight target workloads and get indicators of real-world results in the cloud. Select benchmarks that match the computing footprint of your critical workloads,[6] optimize them with accelerators, and calculate cost from published rates. Testing described here does just that to demonstrate compute and cost outcomes in typical scenarios. Take the guesswork out of getting performance and price performance where you need it most.

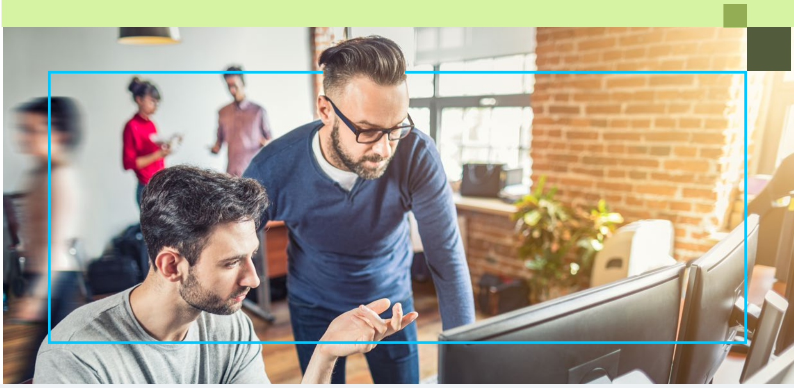
Online publishing and content management

Most companies have a website to manage. Web content–management systems are used to build websites and deliver content to site visitors. WordPress is a widely used content-management system.[7]

We tested a standard WordPress workload that simulates an end-to-end multi-tier content-management system in five popular sizes of two generally available instances, C6g with Graviton2 and C6i with Intel. First we ran the workload on Graviton2. Then we ran the workload on Intel before and after enabling acceleration technologies.

Figure 1 shows that C6i with Intel outperforms C6g with Graviton2 across all instance sizes by up to 1.42x, and up to 1.50x with Intel acceleration.[8]

Application:	WordPress
Metric:	Transactions per second (TPS)
Instance type:	Compute Optimized
Instances:	C6i with 3rd Gen Intel Xeon processors and C6g with AWS Graviton2 processors
Workload acceleration:	Intel® Crypto Acceleration



Up to 1.20x better price performance on WordPress TPS with acceleration in C6i than C6g[9]

C6i outperformed C6g instances on WordPress TPS in all vCPU sizes tested. Across the board, the best performance was achieved with accelerators enabled.

Encryption and decryption of security-enabled web pages in WordPress demand a great deal of computing power. Workloads like WordPress are complemented by Intel Crypto Acceleration, which speeds up cryptographic operations and the overall performance of the software.

Acceleration also delivered up to 1.20x better performance per dollar in C6i instances than C6g.[9]

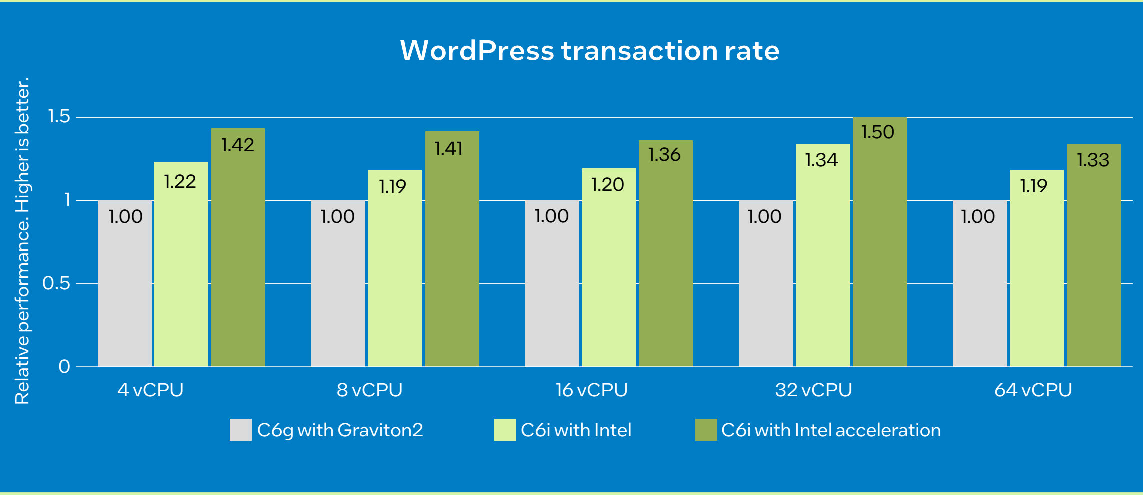


Figure 1. AWS instances with Intel outperform instances with AWS Graviton2 in WordPress TPS measurements. Intel Crypto Acceleration increases the performance advantage.[8]

Tune WordPress to get optimal performance from Intel Crypto Acceleration and avoid frustrating your customers with page bounces or slow content loading.[10]

Web servers

Web server software is at the heart of the web. These applications accept HTTP and HTTPS requests and deliver content in response. NGINX is the most widely used web server, followed by Apache and Cloudflare.[11] Its popularity comes from its design, which focuses on performance optimization. NGINX also provides advanced load balancing, web serving, and reverse-proxy serving.

For this test, we compared CPS from a standard NGINX workload in popular sizes of generally available instances. We ran the workload in R6g with Graviton2 and then in R6i before and after enabling acceleration software.

As shown in Figure 2, R6i with Intel consistently delivered better NGINX performance than R6g by up to 2.93x.[12] With Intel acceleration, NGINX performance in R6i was up to 8.45x better than R6g.[13]

Application:	NGINX
Metric:	Connections per second (CPS)
Instance type:	Memory Optimized
Instances:	R6i with 3rd Gen Intel Xeon processors and R6g with AWS Graviton2 processors
Workload acceleration:	Intel® QuickAssist Technology Engine for OpenSSL

Up to 6.77x better price performance on NGINX CPS in R6i with acceleration than in R6g[14]

These results show how dramatically enabling Intel QuickAssist Technology (Intel® QAT) Engine software with 3rd Gen Intel Xeon processors improves performance. The Intel QAT Engine software uses hardware capabilities for cryptographic and compression algorithms to accelerate web server processes and serve encrypted web connections. This software and hardware combination improves website user experience by loading pages more quickly.

Differences in cost are equally striking. R6i instances with Intel deliver up to 2.35x better price performance in NGINX than R6g.[15] With workload acceleration, R6i delivers up to 6.77x better price performance than R6g.[14]

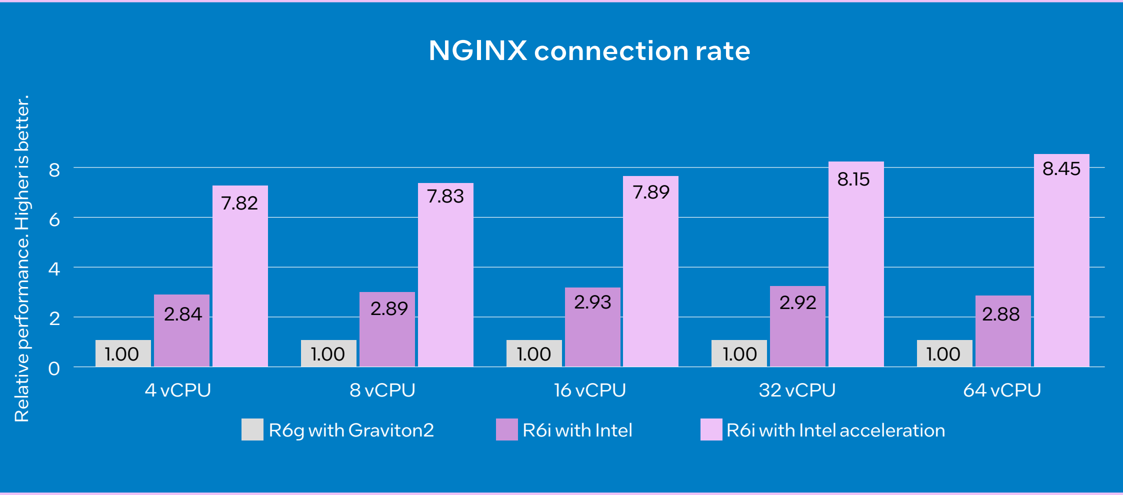


Figure 2. AWS instances with Intel outperform instances with AWS Graviton2 in NGINX CPS. Configuring OpenSSL RSA2K handshakes with Intel QAT Engine software increases the performance advantage.[12,13]

Get better results from NGINX on Intel processors in the cloud or on-premises with optimizations for your workloads.[16]

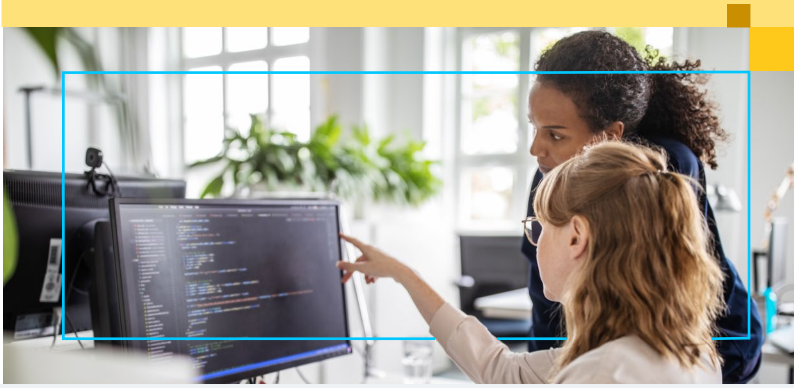
AI inferencing

Inferencing is the stage of AI where value is realized. By carefully selecting instances for optimal inference performance, cloud users can deliver outcomes from AI more effectively. Advantages for AI users are built into AWS instances with Intel.

ResNet-50 is widely used for image processing and model training in TensorFlow. We used this standard workload for Intel testing and in a sponsored test event. The accelerator is Intel® DL Boost, which speeds up inferencing and training operations in compute-intensive workloads.[17] The workload ran in the 16 vCPU size of M6g with Graviton2 and M6i with Intel, both AWS instances in general availability.

Figure 3 shows the results. M6i with Intel outperformed M6g in ResNet-50 by 3.84x.[18] With acceleration enabled, M6i performance increased to 8.0x better than M6g.[19]

Application:	ResNet-50
Metric:	Images processed per second
Instance type:	General Purpose
Instances:	M6i with 3rd Gen Intel Xeon processors and M6g with AWS Graviton2 processors
Workload acceleration:	Intel® Deep Learning Boost (Intel® DL Boost)



Up to 6.4x better price performance on ResNet-50 image processing in M6i with acceleration than in M6g[20]

Workloads running in M6i benefit from high-throughput, low-latency performance for AI with Intel® DL Boost, and developers can choose to use INT8 to improve object detection, image recognition and classification, and natural language processing performance.[21] These tests were conducted using single-precision floating-point format (FP32) due to lack of support for INT8 processing on ARM.[22] FP32, bfloat16, and INT8 data formats are available to AWS cloud users on Intel.

In addition to getting better performance and more choices, cloud users running ResNet-50 with workload acceleration in M6i get up to 6.4x better price performance than M6g.[20]

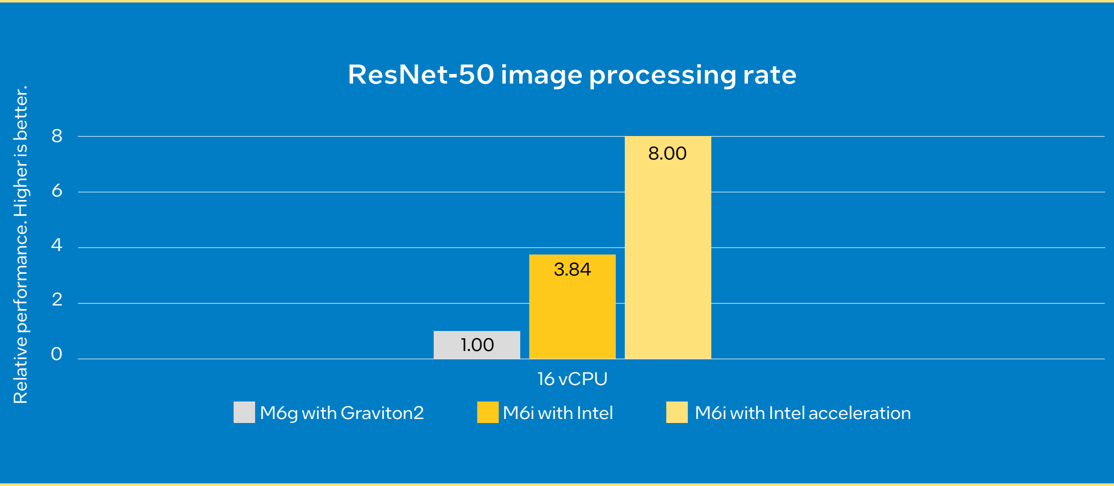


Figure 3. AWS instances with Intel outperform instances with AWS Graviton2 in ResNet-50 inference throughput. Intel DL Boost increases the performance advantage.[18,19]

High-performing inference applications help you realize revenue from applications sooner. Get the benefits of faster performance and better price performance for AI inferencing workloads in Intel cloud instances.

What about the future?

We can compare instances based on the latest processors from AWS and Intel that are widely available in Amazon EC2 today. Instances designated "6g" have AWS Graviton2 processors, and instances designated "7g" have AWS Graviton3 processors. Instances designated "6i" have 3rd Gen Intel Xeon processors, and instances designated "7i" have 4th Gen Intel Xeon processors.

At the time of writing, Graviton3 processors are available in one instance type and two geographic regions. Instances with 4th Gen Intel Xeon processors are available in private preview.[23] Data from 6g, 7g, and 6i helps inform immediate instance selection decisions and anticipate the future.

Acceleration makes the difference

Testing shows that current-generation Intel instances with workload acceleration outperform next-generation Graviton instances in active benchmark testing. Standard NGINX workloads with Intel QAT Engine software optimization deliver up to 2.67x better performance in C6i than on Graviton3 in C7g with 2.28x better performance per dollar.[24]

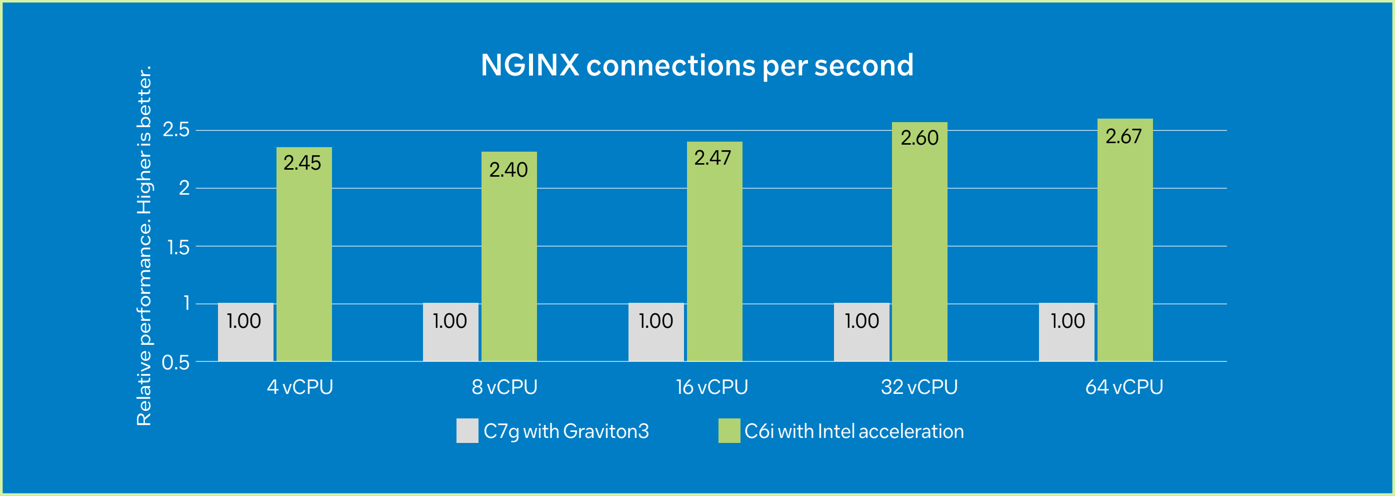


Figure 4. Configuring OpenSSL RSA2K handshakes with Intel QAT Engine software delivers up to 2.67x faster connections in C6i than C7g.[12,13]

$ Up to **2.28x** **better price performance on NGINX CPS in C6i with acceleration than in C7g**[24]

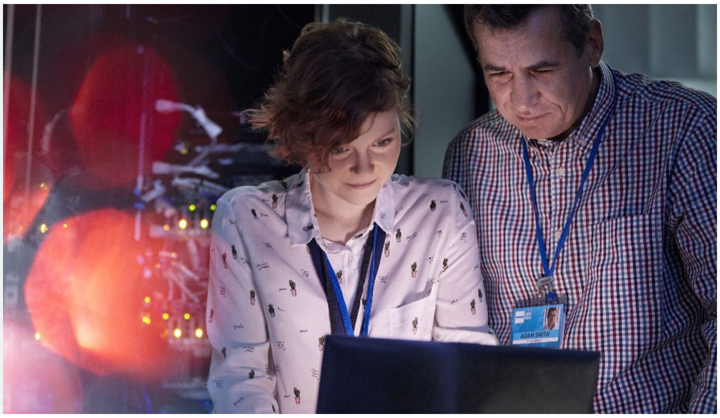
Test data shows better performance per dollar calculated by rate. Beyond those calculations, enabling Intel acceleration can lower overall spend by improving CPU utilization. Better CPU utilization means that you can run your workload at a specific performance level without adding more resources. Using less costs less.

Along with performance and cost data, consider flexibility and level of effort. Workloads optimized for Intel architecture can move and run anywhere, but workloads optimized for AWS Graviton2 and AWS Graviton3 processors are limited to AWS. Also, AWS Graviton3 processors have Neoverse V1 cores, which are not fully compatible with Neoverse N1 cores in Graviton2. Moving from Graviton2 processors to Graviton3 processors requires rewriting, recompiling, and revalidating your code from the bottom of the stack to the top.

Extraordinary effort is not necessary when moving from one AWS Intel instance to another. Instead, move your workloads easily among cloud instances with Intel Xeon processors, take advantage of generational improvements to build out your code base, and grow your business. Placing your workloads on Intel in the cloud or on-premises enables you to take advantage of a large and growing collection of accelerators into the future.

Conclusion

> "The world of CPU performance is changing. We used to be able to simply know the rough IPC of an architecture, multiply by cores and clock speed, and have some sense of what it can do. In 2022, that is going to start changing more and more. … Deciphering CPU performance going forward is going to be a lot more personal and will require a lot more thought."
>
> — Patrick Kennedy, ServeTheHome.com[19]



Predicting performance is not simple and demand for services is immense. For example, an estimated 455 million websites run on WordPress—in other words, 35 percent of all active websites. These websites generate more than two million posts per day.[4] WordPress.com estimates the number of websites powered by its CMS to be even higher, at 42 percent.[7] To add pressure to cloud management decisions, customers give you less than three seconds to deliver online content[1] before they move on to your competitors.

AWS offers more than 500 different choices with different instance types, processors, vCPU sizes, and geographies. How do you determine what combination delivers the best value for your business? Physical core count and lab-measured clock speeds are poor predictors of cloud performance and cost when cloud service is calculated by time to process workloads on virtual resources. Instead, look to data from real-world scenarios. Cloud content-management, web server, and AI operations workload testing delivers more meaningful data than traditional benchmarking.

Testing with common applications, available workload optimizations, and popular instances demonstrates that users in AWS 6i instances with workload accelerators enabled get better results than AWS Graviton 6g and 7g instances. Intel accelerators also extend your data center strategy. Intel accelerators are available in AWS instances, other public clouds, and on-premises data centers.

intel. XEON

Learn more

Take the work out of optimizing your cloud workloads.

Comparing Graviton and Intel on AWS?

Visit us for performance data, resources, and expert help.

[1] Chrome Developers. "Speed is now a landing page factor for Google Search and Ads." July 2018. https://developer.chrome.com/blog/search-ads-speed/.

[2] Virtual. "8 Reasons Why Fortune 500 Companies Use the WordPress CMS." https://vtldesign.com/digital-marketing/social-media/8-reasons-fortune-500-companies-use-wordpress-cms/.

[3] Kinsta. "What Is Nginx? A Basic Look at What It Is and How It Works." January 2022. https://kinsta.com/knowledgebase/what-is-nginx/.

[4] Digital.com. "15 Fascinating Facts About WordPress." January 2023. https://digital.com/best-web-hosting/wordpress/statistics/.

[5] Dgtl Infra. "Top 10 Cloud Service Providers Globally in 2022." January 2023. https://dgtlinfra.com/top-10-cloud-service-providers-2022/.

[6] Intel. "Active Benchmarking for Better Performance Predictions." May 2022. intel.com/content/www/us/en/content-details/752333/active-benchmarking-for-better-performance-predictions.html.

[7] WordPress claims to power more than 42% of the web. Source: WordPress. About Us web page. Accessed January 2023. https://wordpress.com/about/.

[8] See intel.com/3gen-xeon-config [131] and [132] for configuration and test information.

[9] See intel.com/3gen-xeon-config [132] for configuration and test information. AWS C6i with Intel acceleration delivers up to 1.20x better price performance and delivers up to 1.50x better performance on WordPress TPS than C6g with AWS Graviton2. Pricing information from US-East-1c as of January 5, 2023. https://aws.amazon.com/ec2/pricing/on-demand/.

[10] See intel.com/content/www/us/en/developer/articles/guide/wordpress-tuning-guide-on-xeon-systems and https://artifacthub.io/packages/helm/bitnami/wordpress-intel for more information on optimizing workloads in WordPress.

[11] W3Techs. "Usage of web servers broken down by ranking." Accessed October 13, 2022. https://w3techs.com/technologies/cross/web_server/ranking.

[12] See intel.com/3gen-xeon-config [137] for configuration and test information.

[13] See intel.com/3gen-xeon-config [136] for configuration and test information.

[14] See intel.com/3gen-xeon-config [136] for configuration and test information. AWS R6i instances with Intel acceleration technologies deliver up to 6.77x better price performance and deliver up to 8.45x better performance on NGINX CPS than R6g instances with AWS Graviton2 processors. Pricing information from US-West-2b as of January 5, 2023. https://aws.amazon.com/ec2/pricing/on-demand/.

[15] See intel.com/3gen-xeon-config [137] for configuration and test information. AWS R6i instances deliver up to 2.35x better price performance and deliver up to 2.93x better performance on NGINX CPS than R6g instances with AWS Graviton2 processors. Pricing information from US-West-2b as of January 5, 2023. https://aws.amazon.com/ec2/pricing/on-demand/.

[16] For more information about Intel QAT Engine, see https://github.com/intel/QAT_Engine/. To take advantage of Intel QAT Engine acceleration, see: Intel. "Building Software Acceleration Features in the Intel® QuickAssist Technology (Intel® QAT) Engine for OpenSSL* 1.1.1." intel.com/content/www/us/en/developer/articles/guide/building-software-acceleration-features-in-the-intel-qat-engine-for-openssl.html. To tune NGINX with OpenSSL and Intel QAT Engine, see intel.com/content/www/us/en/developer/articles/guide/nginx-https-with-qat-tuning-guide.html.

[17] See intel.com/content/www/us/en/artificial-intelligence/documents/enhance-ai-workloads-built-in-accelerators-pdf for more information on optimizing AI workloads.

[18] See intel.com/3gen-xeon-config [77] for configuration and test information.

[19] ServeTheHome. "AWS EC2 m6 Instances: Why Acceleration Matters." Sponsored by Intel. November 2021. servethehome.com/aws-ec2-m6-instance-intel-ice-lake-and-graviton-2-acceleration-matters.

[20] ServeTheHome. "AWS EC2 m6 Instances: Why Acceleration Matters." Sponsored by Intel. November 2021. servethehome.com/aws-ec2-m6-instance-intel-ice-lake-and-graviton-2-acceleration-matters. AWS M6i.4xlarge instances with Intel acceleration technologies deliver up to 6.4x better price performance on ResNet-50 inference throughput than M6g.4xlarge instances with AWS Graviton2 processors. Pricing information from US-West-2 region as of January 5, 2023. https://aws.amazon.com/ec2/pricing/on-demand/.

[21] For more information about tuning your deep learning applications in AWS "6i" instances, see intel.com/content/www/us/en/developer/articles/guide/deep-learning-with-avx512-and-dl-boost.html.

[22] The use of FP32 in their test was explained by ServeTheHome as follows: "Note: Graviton2 does not have the same INT-8 support so we got a very poor result. We did not want to use it as a misleading baseline." See ServeTheHome. "AWS EC2 m6 Instances: Why Acceleration Matters." November 2021. servethehome.com/aws-ec2-m6-instance-intel-ice-lake-and-graviton-2-acceleration-matters.

[23] Intel. "AWS Activates 4th Gen Intel Xeon Scalable Processors for New EC2 R7i Instances." November 2022. intel.com/content/www/us/en/newsroom/news/aws-activates-4th-gen-xeon-scalable-processors.

[24] See intel.com/3gen-xeon-config [138] for configuration and test information. AWS C6i instances with Intel acceleration technologies deliver up to 2.28x better price performance and deliver up to 2.67x better performance on NGINX CPS than C7g instances with AWS Graviton3 processors. Pricing information from US-West-2 as of January 5, 2023. https://aws.amazon.com/ec2/pricing/on-demand/.

[25] See intel.com/3gen-xeon-config [138] for configuration and test information.

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.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Printed in USA 0823/KK/PRW/PDF Please Recycle 353347-001US