Get more from your Amazon EC2 investment by migrating your workloads to next-generation EC2 instances

Instances based on Intel Xeon Scalable Processors and the AWS Nitro System delivered greater performance at a lower price than comparable previous-generation instances

Many IT professionals spend their workdays attending to urgent matters such as the release of a mandatory upgrade, a security breach, or a hardware failure. When putting out fires is part of your job description, it's easy to adopt the attitude "If it isn't broken, don't fix it." But just because a solution is functional doesn't mean your company is getting the best value.

Previous-generation Amazon Elastic Compute Cloud (Amazon EC2) instances are powered by the Intel® Xeon® E5 family of processors, based on the Broadwell or Haswell microarchitecture. For companies running workloads on these instances, nothing is broken. However, they could enjoy increased performance while also lowering costs by moving these workloads to next-generation instances powered by Intel Xeon Scalable Processors, which are based on the Skylake microarchitecture and feature the Intel Advanced Vector Extensions 512 (AVX-512) architecture.

We conducted hands-on testing to measure performance improvements when moving from previous-generation Amazon EC2 instances to next-generation instances based on Intel Skylake processors. For three instance families—general purpose, compute-optimized, and memory-optimized—we ran a workload optimized for Intel AVX-512 and found performance improvements ranging from 1.7X to 1.8X compared to a workload optimized for Intel Advanced Vector Extensions 2.0 (Intel AVX2) processor architecture running on the instances based on the Intel Broadwell or Haswell processors. We combined these wins with the lower price of the new instances and found that they could deliver as much as twice the performance per dollar spent.

Improve performance by up to **1.79X***

Get up to **2X** the performance per dollar*

For legacy applications, increase performance per dollar by up to **21.9%**

* For a LINPACK workload optimized for AVX-512 running on EC2 instances based on Intel Xeon Scalable Processors compared to EC2 instances based on the Intel Xeon E5 processor family.

Increase performance per dollar with next-generation EC2 instances

If your company is running workloads in Amazon EC2 instances based on previous-generation Intel Broadwell or Haswell processors, moving them to next-generation instances powered by Intel Skylake processors might not be high on your to-do list. After all, the applications are running fine.

It's possible that when you set up these instances, you optimized your workloads for Intel AVX2. Maybe you're wondering whether it's worth the effort to optimize them for Intel AVX-512. We found the answer is yes, and by doing so, you could double the performance per dollar.

What if you didn't optimize your applications for Intel AVX-512 right away? Would there any point in moving these applications to instances based on Intel Xeon Scalable processors? Again, the answer is yes. In our testing, the new instances boosted performance even with legacy applications. When we combine these improvements with the lower price of these instances, we see performance-per-dollar increases of as much as 21.9 percent.

About Intel AVX-512

Intel Advanced Vector Extensions 512 is a set of instructions that can boost performance on a variety of compute-intensive workloads such as data analytics, scientific simulations, and audio/video processing.

Intel AVX-512 has 512-bit vector operations capabilities, which doubles the 256-bit vector operations capabilities available in Intel AVX2.

Learn more at https://www.intel.com/content/www/us/en/architecture-and-technology/avx-512-overview.html.

About Amazon Web Services (AWS)

Amazon Web Services provides IT infrastructure as a service, also known as cloud computing. According to AWS, "One of the key benefits of cloud computing is the opportunity to replace up-front capital infrastructure expenses with low variable costs that scale with your business. With the Cloud, businesses no longer need to plan for and procure servers and other IT infrastructure weeks or months in advance. Instead, they can instantly spin up hundreds or thousands of servers in minutes and deliver results faster."[1]

Learn more at https://aws.amazon.com/about-aws/.

Next-generation EC2 instances delivered greater performance

To demonstrate performance improvements when moving to next-generation Amazon EC2 instances based on Intel Skylake processors, we used the LINPACK Benchmark. We tested the following six instances:

- General-purpose instances—m4.large vs. m5.large
- Compute-optimized instances—c4.large vs. c5.large
- Memory-optimized instances—r4.large vs. r5.large

The M4/C4/R4 instances are based on previous-generation Intel Broadwell or Haswell processors, and the M5/C5/R5 instances are based on Intel Skylake processors.

In our testing, we used the Shared-memory (SMP) version of Intel Distribution for LINPACK Benchmark, which is AVX aware. It detects the AVX version for the processor on which the workload is running, and the workload takes advantage of the Intel AVX architecture capabilities of that processor. This means the workload used AVX2 instructions on the previous-generation EC2 instances and AVX-512 instructions on the next-generation instances. LINPACK reports performance using the GFlop, a unit that represents one billion floating point operations per second.

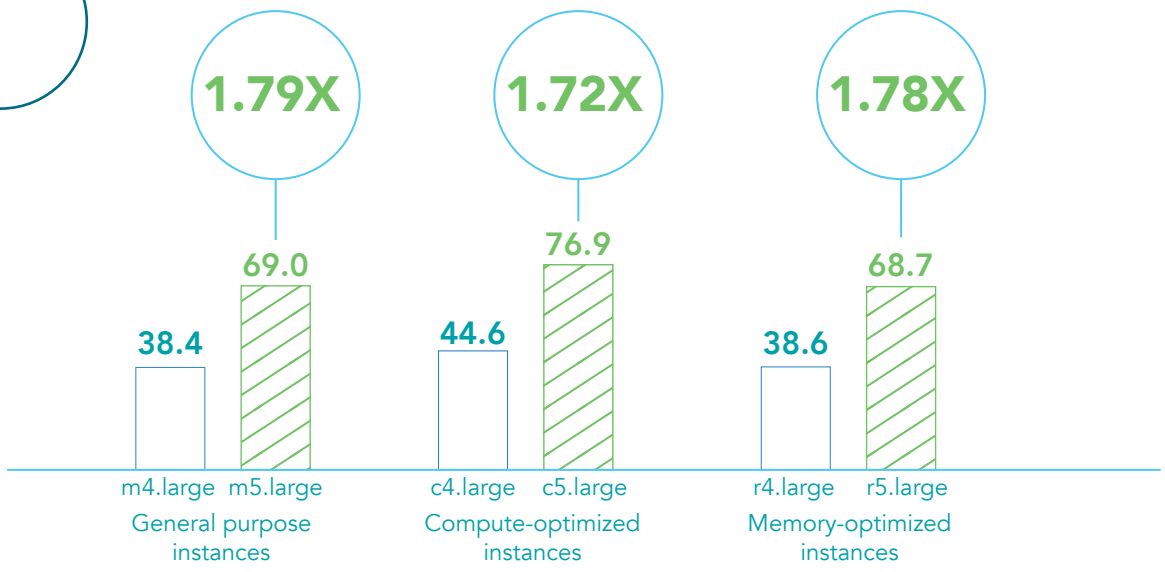
The chart below shows the peak measured GFlops LINPACK reported for the six instances we tested. When moving to the M5/C5/R5 instances, performance increased on all three instance types. The improvement ranged from 72 percent to 79 percent, which means the next-generation instances could perform more than 1.79 times as much work as the previous-generation instances. Being able to perform more work with the same number of instances, or the same amount of work with fewer instances, can translate to savings.

About the Intel Xeon Scalable Processor family

Intel Xeon Scalable processors, the latest generation of server processors from Intel, are based on the Intel Skylake microarchitecture. The family comprises four feature configurations designed for various workloads: Platinum, Gold, Silver, and Bronze. The general-purpose and memory-optimized Amazon EC2 instances we tested used the Intel Xeon Platinum 8175M processor, and the compute-optimized instances used the Intel Xeon Platinum 8124M processor.

Performance with LINPACK optimized for instance-specific Intel AVX architectures

Peak measured GFlops – Higher is better



In addition, we ran the LINPACK workload against the M5/C5/R5 instances using the Intel AVX2 flags, to determine the performance users could expect when migrating a legacy application without making any additional optimizations for AVX-512. As the table below shows, under these conditions, performance again improved on all three instance types when moving to the M5/C5/R5 instances.

Performance with LINPACK optimized for Intel AVX2 architecture
(Peak measured Gflops – higher is better)

	Previous-generation EC2 instance	Next-generation EC2 instance	Percentage improvement
General purpose instances	38.4	**39.7**	**3.4%**
Compute-optimized instances	44.6	**46.2**	**3.6%**
Memory-optimized instances	38.6	**40.1**	**3.9%**

Next-generation EC2 instances were less expensive

Below, we present the pricing for the Amazon EC2 instances we tested, On-Demand instances in the US East region, January 15, 2019.[2] The hourly cost for next-generation instances based on Intel Skylake processors ranged from 4 to 15 percent lower than the cost for previous-generation instances based on Intel Broadwell or Haswell processors.

Price per hour for Amazon EC2 On-Demand instances (US East region)
(lower is better)

	Previous-generation EC2 instance	Next-generation EC2 instance	Percentage difference
General purpose instances	0.100	**0.096**	**-4.0%**
Compute-optimized instances	0.100	**0.085**	**-15.0%**
Memory-optimized instances	0.133	**0.126**	**-5.3%**

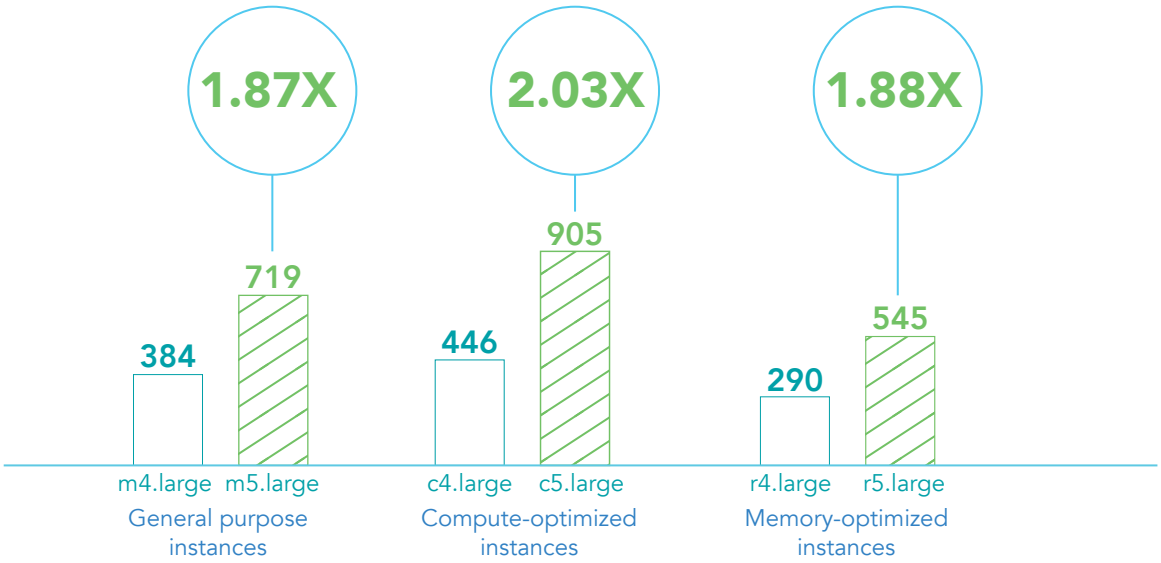
Putting performance and pricing together

As we've noted, in our testing with LINPACK, the next-generation AWS instances based on Intel Skylake processors delivered better performance and cost less than the older instances. In this section, we combine these data points to quantify the improved value a company might see by moving their applications to the new instances. (Note that every application is unique, and performance improvements will not always match those we found in our LINPACK simulation.)

If we start with the performance each instance delivered (in terms of peak measured Gflops reported by LINPACK) and then divide that number by the hourly cost of the instance, we get a performance/price metric. The chart below presents our findings for the first round of testing, with applications optimized for the Intel AVX-512 architecture. As it shows, the new instances delivered up to twice the LINPACK performance per dollar.

Performance/price: Peak measured Gflops/hourly cost with LINPACK optimized for instance-specific Intel AVX architectures

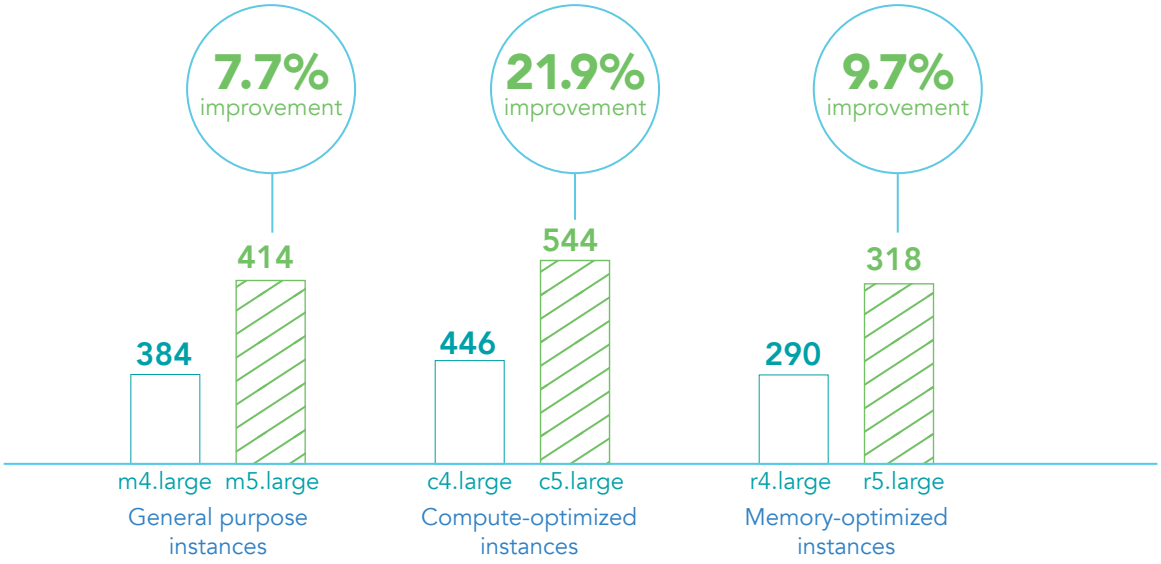
Higher is better



The chart below shows the results from the second round of testing, when LINPACK used the Intel AVX2 flags and thus did not take advantage of the new processor capability. Under these conditions, the price/performance of the AWS instances based on Intel Xeon Scalable Processors was up to 21.9 percent greater than that of the older instances.

Performance/price: Peak measured Gflops/hourly cost with LINPACK optimized for Intel AVX2 architecture

Higher is better



Other considerations: Variants and the AWS Nitro System

In our study, we investigated the performance and pricing advantages of migrating workloads to three of the many next-generation EC2 instances Amazon offers. Below, we address some additional benefits that we didn't test.

Increased memory

For some of the instance families, the memory footprint is greater on the next-generation instances (see the science behind the report for detailed configuration information).

Variants

Amazon offers several EC2 variants, including those with local-attached disk, enhanced networking, and bare metal options.

Local-attached disk

For applications that need access to high-speed, ultra-low latency local storage, instances such as EC2 M5d, C5d, and R5d can be an appropriate choice. These instances use NVMe SSD block-level storage that is physically connected to the host server. According to Amazon, "These instances are a great fit for applications that need access to high-speed, low latency local storage, to temporary storage of data such as batch and log processing, and to high-speed caches and scratch files."[3]

Enhanced networking

For applications that need access to high-speed networking, EC2 C5n instances, which provide up to 100 Gbps of networking throughput, can be an appropriate choice. Amazon states that "Customers can also take advantage of this improved network performance to accelerate data transfer to and from S3, reducing the data ingestion wait time for applications and speeding up delivery of results. A wide range of applications such as High Performance Computing (HPC), analytics, machine learning, Big Data and data lake applications can benefit from these instances."[4]

Bare metal

For applications that need direct access to the Intel Xeon Scalable processor and memory resources of the underlying server, Amazon EC2 bare metal instances can be an appropriate choice. According to Amazon, "These instances are ideal for workloads that require access to the hardware feature set (such as Intel VT-x), for applications that need to run in non-virtualized environments for licensing or support requirements, or for customers who wish to use their own hypervisor."[5]

AWS Nitro System

Next-generation EC2 instances are powered by the AWS Nitro System, which combines dedicated hardware with a lightweight hypervisor. Amazon describes the AWS Nitro System as "a rich collection of building blocks that offloads many of the traditional virtualization functions to dedicated hardware and software to deliver high performance, high availability, and high security while also reducing virtualization overhead."[6]

Technology changes continuously, and you often have no choice about updating or upgrading. That can make it tempting to delay moving cloud applications that are functioning fine where they are. However, as we learned in our study, by keeping your workloads in EC2 instances based on older architecture, you're leaving value on the table.

We found between the lower cost and improved performance of next-generation EC2 instances based on Intel Xeon Scalable Processors using AVX-512, you could double your performance per dollar by moving some applications and re-optimizing them. For some applications still using instructions for Intel AVX2, moving to the new instances could boost your performance per dollar by 21.9 percent.[7] These numbers make it clear that moving your workloads is a good idea.

---

1 Amazon Web Services, "About AWS," accessed January 15, 2019, https://aws.amazon.com/about-aws/.

2 Amazon Web Services, "Amazon EC2 pricing," accessed January 15, 2019, https://aws.amazon.com/ec2/pricing/on-demand/.

3 Amazon Web Services, "Amazon EC2 C5d, M5d, and R5d Instances are Now Available in Additional AWS Regions," accessed March 19, 2019, https://aws.amazon.com/about-aws/whats-new/2018/12/amazon-ec2-c5d-m5d-and-r5d-instances-are-now-available-in-additional-aws-regions/.

4 Amazon Web Services, "Introducing Amazon EC2 C5n Instances Featuring 100 Gbps of Network Bandwidth," accessed March 19, 2019, https://aws.amazon.com/about-aws/whats-new/2018/11/introducing-amazon-ec2-c5n-instances/.

5 Amazon Web Services, "Introducing Five New Amazon EC2 Bare Metal Instances," accessed March 19, 2019, https://aws.amazon.com/about-aws/whats-new/2019/02/introducing-five-new-amazon-ec2-bare-metal-instances/.

6 Amazon Web Services, "Amazon EC2 C5 Instances," accessed March 19, 2019, https://aws.amazon.com/ec2/instance-types/c5/.

7 These calculations reflect the LINPACK workload results from our testing. Performance improvements on other applications will vary.

Read the science behind this report at http://facts.pt/3uqnzk4 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Amazon Web Services.