Amazon EC2 is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Q: What can I do with Amazon EC2?

Just as Amazon Simple Storage Service (Amazon S3) enables storage in the cloud, Amazon EC2 enables "compute" in the cloud. The Amazon EC2 simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.

Q: How can I get started with Amazon EC2?

To sign up for Amazon EC2, select the "Sign up for This Web Service" button on the Amazon EC2 detail page. You must have an AWS account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon EC2 signup process. After signing up, please refer to the Amazon EC2 documentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for Amazon EC2?

Amazon EC2 registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a couple of minutes and involves receiving a phone call during the registration process and entering a PIN using the phone key pad.

Q: What can developers now do that they could not before?

Until now, small developers did not have the capital to acquire massive compute resources and ensure they had the capacity they needed to handle unexpected spikes in load. Amazon EC2 helps developers use Amazon's own benefits of massive scale with no upfront investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure they have the compute capacity they need to meet their business requirements.

The “Elastic” nature of the service allows developers to instantly scale to meet spikes in traffic or demand. When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Q: How do I run systems in the Amazon EC2 environment?

Once you have set up your account and select or create your AMIs, you are ready to boot your instance. You can start your AMI on any number of On-Demand instances by using the RunInstances API call. You simply need to indicate how many instances you wish to launch. If you wish to run more than your On-Demand quota, complete the Amazon EC2 instance request form.

If Amazon EC2 is able to fulfill your request, RunInstances will return success, and we will start launching your instances. You can check on the status of your instances using the DescribeInstances API call. You can also programmatically terminate any number of your instances using the TerminateInstances API call.

If you have a running instance using an Amazon EBS boot partition, you can also use the StopInstances API call to release the compute resources but preserve the data on the boot partition. You can use the StartInstances API when you are ready to restart the associated instance with the Amazon EBS boot partition.

In addition, you have the option to use Spot Instances to reduce your computing costs when you have flexibility in when your applications can run. Read more about Spot Instances for a more detailed explanation on how Spot Instances work.

If you prefer, you can also perform all these actions from the AWS Management Console or through the command line using our command line tools, which have been implemented with this web service API.

Q: What is the difference between using the local instance store and Amazon Elastic Block Store (Amazon EBS) for the root device?

When you launch your Amazon EC2 instances you have the ability to store your root device data on Amazon EBS or the local instance store. By using Amazon EBS, data on the root device will persist independently from the lifetime of the instance. This enables you to stop and restart the instance at a subsequent time, which is similar to shutting down your laptop and restarting it when you need it again.

Alternatively, the local instance store only persists during the life of the instance. This is an inexpensive way to launch instances where data is not stored to the root device. For example, some customers use this option to run large web sites where each instance is a clone to handle web traffic.

Q: How quickly will systems be running?

It typically takes less than 10 minutes from the issue of the RunInstances call to the point where all requested instances begin their boot sequences. This time depends on a number of factors including: the size of your AMI, the number of instances you are launching, and how recently you have launched that AMI. Images launched for the first time may take slightly longer to boot.

Q: How do I load and store my systems with Amazon EC2?

Amazon EC2 allows you to set up and configure everything about your instances from your operating system up to your applications. An Amazon Machine Image (AMI) is simply a packaged-up environment that includes all the necessary bits to set up and boot your instance. Your AMIs are your unit of deployment. You might have just one AMI or you might compose your system out of several building block AMIs (e.g., webservers, appservers, and databases). Amazon EC2 provides a number of tools to make creating an AMI easy. Once you create a custom AMI, you will need to bundle it. If you are bundling an image with a root device backed by Amazon EBS, you can simply use the bundle command in the AWS Management Console. If you are bundling an image with a boot partition on the instance store, then you will need to use the AMI Tools to upload it to Amazon S3. Amazon EC2 uses Amazon EBS and Amazon S3 to provide reliable, scalable storage of your AMIs so that we can boot them when you ask us to do so.

Or, if you want, you don’t have to set up your own AMI from scratch. You can choose from a number of globally available AMIs that provide useful instances. For example, if you just want a simple Linux server, you can choose one of the standard Linux distribution AMIs.

Q: How do I access my systems?

The RunInstances call that initiates execution of your application stack will return a set of DNS names, one for each system that is being booted. This name can be used to access the system exactly as you would if it were in your own data center. You own that machine while your operating system stack is executing on it.

Q: Is Amazon EC2 used in conjunction with Amazon S3?

Yes, Amazon EC2 is used jointly with Amazon S3 for instances with root devices backed by local instance storage. By using Amazon S3, developers have access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. In order to execute systems in the Amazon EC2 environment, developers use the tools provided to load their AMIs into Amazon S3 and to move them between Amazon S3 and Amazon EC2. See How do I load and store my systems with Amazon EC2? for more information about AMIs.

We expect developers to find the combination of Amazon EC2 and Amazon S3 to be very useful. Amazon EC2 provides cheap, scalable compute in the cloud while Amazon S3 allows users to store their data reliably.

Q: How many instances can I run in Amazon EC2?

You are limited to running On-Demand Instances per your vCPU-based On-Demand Instance limit, purchasing 20 Reserved Instances, and requesting Spot Instances per your dynamic Spot limit per region. New AWS accounts may start with limits that are lower than the limits described here.

If you need more instances, complete the Amazon EC2 limit increase request form with your use case, and your limit increase will be considered. Limit increases are tied to the region they were requested for.

Q: Are there any limitations in sending email from Amazon EC2 instances?

Yes. In order to maintain the quality of Amazon EC2 addresses for sending email, we enforce default limits on the amount of email that can be sent from EC2 accounts. If you wish to send larger amounts of email from EC2, you can apply to have these limits removed from your account by filling out this form.

Q: How quickly can I scale my capacity both up and down?

Amazon EC2 provides a truly elastic computing environment. Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can commission one, hundreds or even thousands of server instances simultaneously. When you need more instances, you simply call RunInstances, and Amazon EC2 will typically set up your new instances in a matter of minutes. Of course, because this is all controlled with web service APIs, your application can automatically scale itself up and down depending on its needs.

Q: What operating system environments are supported?

Amazon EC2 currently supports a variety of operating systems including: Amazon Linux, Ubuntu, Windows Server, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, openSUSE Leap, Fedora, Fedora CoreOS, Debian, CentOS, Gentoo Linux, Oracle Linux, and FreeBSD. We are looking for ways to expand it to other platforms.

Q: Does Amazon EC2 use ECC memory?

In our experience, ECC memory is necessary for server infrastructure, and all the hardware underlying Amazon EC2 uses ECC memory.

Q: How is this service different than a plain hosting service?

Traditional hosting services generally provide a pre-configured resource for a fixed amount of time and at a predetermined cost. Amazon EC2 differs fundamentally in the flexibility, control and significant cost savings it offers developers, allowing them to treat Amazon EC2 as their own personal data center with the benefit of Amazon.com's robust infrastructure.

When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Secondly, many hosting services don't provide full control over the compute resources being provided. Using Amazon EC2, developers can choose not only to initiate or shut down instances at any time, they can completely customize the configuration of their instances to suit their needs – and change it at any time. Most hosting services cater more towards groups of users with similar system requirements, and so offer limited ability to change these.

Finally, with Amazon EC2 developers enjoy the benefit of paying only for their actual resource consumption – and at very low rates. Most hosting services require users to pay a fixed, upfront fee irrespective of their actual computing power used, and so users risk overbuying resources to compensate for the inability to quickly scale up resources within a short time frame.

EC2 On-Demand Instance limits

Q: What is changing?

Amazon EC2 is transitioning On-Demand Instance limits from the current instance count-based limits to the new vCPU-based limits to simplify the limit management experience for AWS customers. Usage toward the vCPU-based limit is measured in terms of number of vCPUs (virtual central processing units) for the Amazon EC2 Instance Types to launch any combination of instance types that meet your application needs.

Q: What are vCPU-based limits?

You are limited to running one or more On-Demand Instances in an AWS account, and Amazon EC2 measures usage towards each limit based on the total number of vCPUs (virtual central processing unit) that are assigned to the running On-Demand instances in your AWS account. The following table shows the number of vCPUs for each instance size. The vCPU mapping for some instance types may differ; see Amazon EC2 Instance Types for details.

Instance Size	vCPUs

nano	1
micro	1
small	1
medium	1
large	2
xlarge	4
2xlarge	8
3xlarge	12
4xlarge	16
8xlarge	32
9xlarge	36
10xlarge	40
12xlarge	48
16xlarge	64
18xlarge	72
24xlarge	96
32xlarge	128

Q: How many On-Demand Instances can I run in Amazon EC2?

There are five vCPU-based instance limits; each defines the amount of capacity you can use of a given instance family. All usage of instances in a given family, regardless of generation, size, or configuration variant (e.g. disk, processor type), will accrue towards the family's total vCPU limit, listed in the table below. New AWS accounts may start with limits that are lower than the limits described here.

On-Demand Instance Limit Name	Default vCPU Limit
Running On-Demand Standard (A, C, D, H, I, M, R, T, Z) instances	1152 vCPUs
Running On-Demand F instances	128 vCPUs
Running On-Demand G instances	128 vCPUs
Running On-Demand Inf instances	128 vCPUs
Running On-Demand P instances	128 vCPUs
Running On-Demand X instances	128 vCPUs

Q: Are these On-Demand Instance vCPU-based limits regional?

Yes, the On-Demand Instance limits for an AWS account are set on a per-region basis.

Q: Will these limits change over time?

Yes, limits can change over time. Amazon EC2 is constantly monitoring your usage within each region and your limits are raised automatically based on your use of EC2.

Q: How can I request a limit increase?

Even though EC2 automatically increases your On-Demand Instance limits based on your usage, if needed you can request a limit increase from the Limits Page on Amazon EC2 console, the Amazon EC2 service page on the Service Quotas console, or the Service Quotas API/CLI.

Q: How can I calculate my new vCPU limit?

You can find the vCPU mapping for each of the Amazon EC2 Instance Types or use the simplified vCPU Calculator to compute the total vCPU limit requirements for your AWS account.

Q: Do vCPU limits apply when purchasing Reserved Instances or requesting Spot Instances?

No, the vCPU-based limits only apply to running On-Demand instances and Spot Instances.

Q: How can I view my current On-Demand Instance limits?

You can find your current On-Demand Instance limits on the EC2 Service Limits page in the Amazon EC2 console, or from the Service Quotas console and APIs.

Q: Will this affect running instances?

No, opting into vCPU-based limits will not affect any running instances.

Q: Can I still launch the same number of instances?

Yes, the vCPU-based instance limits allow you to launch at least the same number of instances as count-based instance limits.

Q: Will I be able to view instance usage against these limits?

With the Amazon CloudWatch metrics integration, you can view EC2 usage against limits in the Service Quotas console. Service Quotas also enables customers to use CloudWatch for configuring alarms to warn customers of approaching limits. In addition, you can continue to track and inspect your instance usage in Trusted Advisor and Limit Monitor.

Q: Will I still be able to use the DescribeAccountAttributes API?

With the vCPU limits, we no longer have total instance limits governing the usage. Hence the DescribeAccountAttributes API will no longer return the max-instances value. Instead you can now use the Service Quotas APIs to retrieve information about EC2 limits. You can find more information about the Service Quotas APIs in the AWS documentation.

Q: Will the vCPU limits have an impact on my monthly bill?

No. EC2 usage is still calculated either by the hour or the second, depending on which AMI you're running and the instance type and size you've launched.

Q: Will vCPU limits be available in all Regions?

vCPU-based instance limits are available in all commercial AWS Regions.

Changes to EC2 SMTP endpoint policy

Q: What is changing?

As of Januaury 7, 2020, Amazon EC2 began rolling out a change to restrict email traffic over port 25 by default to protect customers and other recipients from spam and email abuse. Port 25 is typically used as the default SMTP port to send emails. AWS accounts that have requested and had Port 25 throttles removed in the past will not be impacted by this change.

Q: I have a valid use case for sending emails to port 25 from EC2. How can I have these port 25 restrictions removed?

If you have a valid use case for sending emails to port 25 (SMTP) from EC2, please submit a Request to Remove Email Sending Limitations to have these restrictions lifted. You can alternately send emails using a different port, or leverage an existing authenticated email relay service such as Amazon Simple Email Service (Amazon SES).

Service level agreement (SLA)

Q: What does your Amazon EC2 Service Level Agreement guarantee?

Our SLA guarantees a Monthly Uptime Percentage of at least 99.99% for Amazon EC2 and Amazon EBS within a Region.

Q: How do I know if I qualify for an SLA Service Credit?

You are eligible for an SLA credit for either Amazon EC2 or Amazon EBS (whichever was Unavailable, or both if both were Unavailable) if the Region that you are operating in has an Monthly Uptime Percentage of less than 99.99% during any monthly billing cycle. For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, see the Amazon Compute Service Level Agreement.

Instance types

Accelerated Computing instances | Burstable instances | Compute Optimized instances | HPC Optimized instances | General Purpose instances | High Memory instances | Memory Optimized instances | Previous Generation instances | Storage Optimized instances

Accelerated Computing instances

Q: What are Accelerated Computing instances?

The Accelerated Computing instance category includes instance families that use hardware accelerators, or co-processors, to perform some functions, such as floating-point number calculation and graphics processing, more efficiently than is possible in software running on CPUs. Amazon EC2 provides three types of Accelerated Computing instances – GPU compute instances for general-purpose computing, GPU graphics instances for graphics intensive applications, and FPGA programmable hardware compute instances for advanced scientific workloads.

Q: When should I use GPU Graphics and Compute instances?

GPU instances work best for applications with massive parallelism such as workloads using thousands of threads. Graphics processing is an example with huge computational requirements, where each of the tasks is relatively small, the set of operations performed form a pipeline, and the throughput of this pipeline is more important than the latency of the individual operations. To be able to build applications that exploit this level of parallelism, one needs GPU device specific knowledge by understanding how to program against various graphics APIs (DirectX, OpenGL) or GPU compute programming models (CUDA, OpenCL).

Q: What applications can benefit from P4d?

Some of the applications that we expect customers to use P4d for are machine learning (ML) workloads like natural language understanding, perception model training for autonomous vehicles, image classification, object detection and recommendation engines. The increased GPU performance can significantly reduce the time to train and the additional GPU memory will help customers train larger, more complex models. HPC customers can use P4's increased processing performance and GPU memory for seismic analysis, drug discovery, DNA sequencing, and insurance risk modeling.

Q: How do P4d instances compare to P3 instances?

P4 instances feature NVIDIA's latest generation A100 Tensor Core GPUs to provide on average 2.5X increase in TFLOP performance over the previous generation V100 along with 2.5X the GPU memory. P4 instances feature Cascade Lake Intel CPU that has 24C per socket and an additional instruction set for vector neural network instructions. P4 instances will have 1.5X the total system memory and 4X the networking throughput of P3dn or 16x compared to P3.16xl. Another key difference is that the NVSwitch GPU interconnect throughput will double what was possible on P3 so each GPU can communicate with every other GPU at the same 600GB/s bidirectional throughput and with single-hop latency. This allows application development to consider multiple GPUs and memories as a single large GPU and a unified pool of memory. P4d instances are also deployed in tightly coupled hyperscale clusters, called EC2 UltraClusters, that enable you to run the most complex multi-node ML training and HPC applications.

Q: What are EC2 UltraClusters and how can I get access?

P4d instances are deployed in hyperscale clusters called EC2 UltraClusters. Each EC2 UltraCluster is comprised of more than 4,000 NVIDIA A100 Tensor Core GPUs, Petabit-scale networking, and scalable low latency storage with FSx for Lustre. Each EC2 UltraCluster is one of the

world's top supercomputers. Anyone can easily spin up P4d instances in EC2 SuperClusters. For additional help, contact us.

Q: Will AMIs that I used on P3 and P3dn work on P4?

The P4 AMIs will need new NVIDIA drivers for the A100 GPUs and a newer version of the ENA driver installed. P4 instances are powered by Nitro System and they require AMIs with NVMe and ENA driver installed. P4 also comes with new Intel Cascade Lake CPUs, which come with an updated instruction set, so we recommend using the latest distributions of ML frameworks, which take advantage of these new instruction sets for data pre-processing.

Q: How are P3 instances different from G3 instances?

P3 instances are the next-generation of EC2 general-purpose GPU computing instances, powered by up to 8 of the latest-generation NVIDIA Tesla V100 GPUs. These new instances significantly improve performance and scalability, and add many new features, including new Streaming Multiprocessor (SM) architecture for machine learning (ML)/deep learning (DL) performance optimization, second-generation NVIDIA NVLink high-speed GPU interconnect, and highly tuned HBM2 memory for higher-efficiency.

G3 instances use NVIDIA Tesla M60 GPUs and provide a high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA Tesla M60 GPUs support NVIDIA GRID Virtual Workstation features, and H.265 (HEVC) hardware encoding. Each M60 GPU in G3 instances supports 4 monitors with resolutions up to 4096x2160, and is licensed to use NVIDIA GRID Virtual Workstation for one Concurrent Connected User. Example applications of G3 instances include 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming, video encoding, and other server-side graphics workloads.

Q: What are the benefits of NVIDIA Volta GV100 GPUs?

The new NVIDIA Tesla V100 accelerator incorporates the powerful new Volta GV100 GPU. GV100 not only builds upon the advances of its predecessor, the Pascal GP100 GPU, it significantly improves performance and scalability, and adds many new features that improve programmability. These advances will supercharge HPC, data center, supercomputer, and deep learning systems and applications.

Q: Who will benefit from P3 instances?

P3 instances with their high computational performance will benefit users in artificial intelligence (AI), machine learning (ML), deep learning (DL) and high performance computing (HPC) applications. Users include data scientists, data architects, data analysts, scientific researchers, ML engineers, IT managers and software developers. Key industries include transportation, energy/oil & gas, financial services (banking, insurance), healthcare, pharmaceutical, sciences, IT, retail, manufacturing, high-tech, transportation, government, and academia, among many others.

Q: What are some key use cases of P3 instances?

P3 instances use GPUs to accelerate numerous deep learning systems and applications including autonomous vehicle platforms, speech, image, and text recognition systems, intelligent video analytics, molecular simulations, drug discovery, disease diagnosis, weather forecasting, big data analytics, financial modeling, robotics, factory automation, real-time language translation, online search optimizations, and personalized user recommendations, to name just a few.

Q: Why should customers use GPU-powered Amazon P3 instances for AI/ML and HPC?

GPU-based compute instances provide greater throughput and performance because they are designed for massively parallel processing using thousands of specialized cores per GPU, versus CPUs offering sequential processing with a few cores. In addition, developers have built hundreds of GPU-optimized scientific HPC applications such as quantum chemistry, molecular dynamics, and meteorology, among many others. Research indicates that over 70% of the most popular HPC applications provide built-in support for GPUs.

Q: How are G3 instances different from P2 instances?

G3 instances use NVIDIA Tesla M60 GPUs and provide a high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA Tesla M60 GPUs support NVIDIA GRID Virtual Workstation features, and H.265 (HEVC) hardware encoding. Each M60 GPU in G3 instances supports 4 monitors with resolutions up to 4096x2160, and is licensed to use NVIDIA GRID Virtual Workstation for one Concurrent Connected User. Example applications of G3 instances include 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming, video encoding, and other server-side graphics workloads.

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 25 Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory, making them ideal for deep learning, high performance databases, computational fluid dynamics, computational finance, seismic analysis, molecular modeling, genomics, rendering, and other server-side GPU compute workloads.

Q: How are P3 instances different from P2 instances?

P3 Instances are the next-generation of EC2 general-purpose GPU computing instances, powered by up to 8 of the latest-generation NVIDIA Volta GV100 GPUs. These new instances significantly improve performance and scalability and add many new features, including new Streaming Multiprocessor (SM) architecture, optimized for machine learning (ML)/deep learning (DL) performance, second-generation NVIDIA NVLink high-speed GPU interconnect, and highly tuned HBM2 memory for higher-efficiency.

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 25 Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory.

Q: What APIs and programming models are supported by GPU Graphics and Compute instances?

P3 instances support CUDA 9 and OpenCL, P2 instances support CUDA 8 and OpenCL 1.2 and G3 instances support DirectX 12, OpenGL 4.5, CUDA 8, and OpenCL 1.2.

Q: Where do I get NVIDIA drivers for P3 and G3 instances?

There are two methods by which NVIDIA drivers may be obtained. There are listings on the AWS Marketplace that offer Amazon Linux AMIs and Windows Server AMIs with the NVIDIA drivers pre-installed. You may also launch 64-bit, HVM AMIs and install the drivers yourself. You must visit the NVIDIA driver website and search for the NVIDIA Tesla V100 for P3, NVIDIA Tesla K80 for P2, and NVIDIA Tesla M60 for G3 instances.

Q: Which AMIs can I use with P3, P2 and G3 instances?

You can currently use Windows Server, SUSE Enterprise Linux, Ubuntu, and Amazon Linux AMIs on P2 and G3 instances. P3 instances only support HVM AMIs. If you want to launch AMIs with operating systems not listed here, contact AWS Customer Support with your request or reach out through EC2 Forums.

Q: Does the use of G3 instances require third-party licenses?

Aside from the NVIDIA drivers and GRID SDK, the use G3 instances does not necessarily require any third-party licenses. However, you are responsible for determining whether your content or technology used on G3 instances requires any additional licensing. For example, if you are streaming content you may need licenses for some or all of that content. If you are using third-party technology such as operating systems, audio and/or video encoders, and decoders from Microsoft, Thomson, Fraunhofer IIS, Sisvel S.p.A., MPEG-LA, and Coding Technologies, please consult these providers to determine if a license is required. For example, if you leverage the on-board h.264 video encoder on the NVIDIA GRID GPU you should reach out to MPEG-LA for guidance, and if you use mp3 technology you should contact Thomson for guidance.

Q: Why am I not getting NVIDIA GRID features on G3 instances using the driver downloaded from the NVIDIA website?

The NVIDIA Tesla M60 GPU used in G3 instances requires a special NVIDIA GRID driver to enable all advanced graphics features, and 4 monitors support with resolution up to 4096x2160. You need to use an AMI with NVIDIA GRID driver pre-installed, or download and install the NVIDIA GRID driver following the AWS documentation.

Q: Why am I unable to see the GPU when using Microsoft Remote Desktop?

When using Remote Desktop, GPUs using the WDDM driver model are replaced with a non-accelerated Remote Desktop display driver. In order to access your GPU hardware, you need to utilize a different remote access tool, such as VNC.

Q: What is Amazon EC2 F1?

Amazon EC2 F1 is a compute instance with programmable hardware you can use for application acceleration. The new F1 instance type provides a high performance, easy to access FPGA for developing and deploying custom hardware accelerations.

Q: What are FPGAs and why do I need them?

FPGAs are programmable integrated circuits that you can configure using software. By using FPGAs you can accelerate your applications up to 30x when compared with servers that use CPUs alone. And, FPGAs are reprogrammable, so you get the flexibility to update and optimize your hardware acceleration without having to redesign the hardware.

Q: How does F1 compare with traditional FPGA solutions?

F1 is an AWS instance with programmable hardware for application acceleration. With F1, you have access to FPGA hardware in a few simple clicks, reducing the time and cost of full-cycle FPGA development and scale deployment from months or years to days. While FPGA technology has been available for decades, adoption of application acceleration has struggled to be successful in both the development of accelerators and the business model of selling custom hardware for traditional enterprises, due to time and cost in development infrastructure, hardware design, and at-scale deployment. With this offering, customers avoid the undifferentiated heavy lifting associated with developing FPGAs in on-premises data centers.

Q: What is an Amazon FPGA Image (AFI)?

The design that you create to program your FPGA is called an Amazon FPGA Image (AFI). AWS provides a service to register, manage, copy, query, and delete AFIs. After an AFI is created, it can be loaded on a running F1 instance. You can load multiple AFIs to the same F1 instance, and can switch between AFIs in runtime without reboot. This lets you quickly test and run multiple hardware accelerations in rapid sequence. You can also offer to other customers on the AWS Marketplace a combination of your FPGA acceleration and an AMI with custom software or AFI drivers.

Q: How do I list my hardware acceleration on the AWS Marketplace?

You would develop your AFI and the software drivers/tools to use this AFI. You would then package these software tools/drivers into an Amazon Machine Image (AMI) in an encrypted format. AWS manages all AFIs in the encrypted format you provide to maintain the security of your code. To sell a product in the AWS Marketplace, you or your company must sign up to be an AWS Marketplace reseller, you would then submit your AMI ID and the AFI ID(s) intended to be packaged in a single product. AWS Marketplace will take care of cloning the AMI and

AFI(s) to create a product, and associate a product code to these artifacts, such that any end-user subscribing to this product code would have access to this AMI and the AFI(s).

Q: What is available with F1 instances?

For developers, AWS is providing a Hardware Development Kit (HDK) to help accelerate development cycles, a FPGA Developer AMI for development in the cloud, an SDK for AMIs running the F1 instance, and a set of APIs to register, manage, copy, query, and delete AFIs. Both developers and customers have access to the AWS Marketplace where AFIs can be listed and purchased for use in application accelerations.

Q: Do I need to be an FPGA expert to use an F1 instance?

AWS customers subscribing to an F1-optimized AMI from AWS Marketplace do not need to know anything about FPGAs to take advantage of the accelerations provided by the F1 instance and the AWS Marketplace. Simply subscribe to an F1-optimized AMI from the AWS Marketplace with an acceleration that matches the workload. The AMI contains all the software necessary for using the FPGA acceleration. Customers need only write software to the specific API for that accelerator and start using the accelerator.

Q: I'm an FPGA developer; how do I get started with F1 instances?

Developers can get started on the F1 instance by creating an AWS account and downloading the AWS Hardware Development Kit (HDK). The HDK includes documentation on F1, internal FPGA interfaces, and compiler scripts for generating AFI. Developers can start writing their FPGA code to the documented interfaces included in the HDK to create their acceleration function. Developers can launch AWS instances with the FPGA Developer AMI. This AMI includes the development tools needed to compile and simulate the FPGA code. The Developer AMI is best run on the latest C5, M5, or R4 instances. Developers should have experience in the programming languages used for creating FPGA code (i.e. Verilog or VHDL) and an understanding of the operation they wish to accelerate.

Q: I'm not an FPGA developer; how do I get started with F1 instances?

Customers can get started with F1 instances by selecting an accelerator from the AWS Marketplace, provided by AWS Marketplace sellers, and launching an F1 instance with that AMI. The AMI includes all of the software and APIs for that accelerator. AWS manages programming the

FPGA with the AFI for that accelerator. Customers do not need any FPGA experience or knowledge to use these accelerators. They can work completely at the software API level for that accelerator.

Q: Does AWS provide a developer kit?

Yes. The Hardware Development Kit (HDK) includes simulation tools and simulation models for developers to simulate, debug, build, and register their acceleration code. The HDK includes code samples, compile scripts, debug interfaces, and many other tools you will need to develop the FPGA code for your F1 instances. You can use the HDK either in an AWS provided AMI, or in your on-premises development environment. These models and scripts are available publicly with an AWS account.

Q: Can I use the HDK in my on-premises development environment?

Yes. You can use the Hardware Development Kit HDK either in an AWS-provided AMI, or in your on-premises development environment.

Q: Can I add an FPGA to any EC2 instance type?

No. F1 instances comes in two instance sizes: f1.2xlarge, f1.4xlarge, and f1.16 xlarge.

Q: How do I use the Inferentia chip in Inf1 instances?

You can start your workflow by building and training your model in one of the popular ML frameworks such as TensorFlow, PyTorch, or MXNet using GPU instances such as P4, P3, or P3dn. Once the model is trained to your required accuracy, you can use the ML framework's API to invoke Neuron, a software development kit for Inferentia, to compile the model for execution on Inferentia chips, load it in to Inferentia's memory, and then execute inference calls. In order to get started quickly, you can use AWS Deep Learning AMIs that come pre-installed with ML frameworks and the Neuron SDK. For a fully managed experience you will be able to use Amazon SageMaker, which will enable you to seamlessly deploy your trained models on Inf1 instances.

Q: When would I use Inf1 vs. C6i or C5 vs. G4 instances for inference?

Customers running machine learning models that are sensitive to inference latency and throughput can use Inf1 instances for high-performance cost-effective inference. For those ML models that are less sensitive to inference latency and throughput, customers can use EC2 C6i or C5 instances and utilize the AVX-512/VNNI instruction set. For ML models that require access to NVIDIA's CUDA, CuDNN or TensorRT libraries, we recommend using G4 instances.

Model Characteristics and Libraries Used	EC2 Inf1	EC2 C6i or C5	EC2 G4
Models that benefit from low latency and high throughput at low cost	X		
Models not sensitive to latency and throughput		X	
Models requiring NVIDIA's developer libraries			X

Q: When should I choose Elastic Inference (EI) for inference vs Amazon EC2 Inf1 instances?

There are two cases where developers would choose EI over Inf1 instances: (1) if you need different CPU and memory sizes than what Inf1 offers, then you can use EI to attach acceleration to the EC2 instance with the right mix of CPU and memory for your application (2) if your performance requirements are significantly lower than what the smallest Inf1 instance provides, then using EI could be a more cost effective choice. For example, if you only need 5 TOPS, enough for processing up to 6 concurrent video streams, then using the smallest slice of EI with a C5.large instance could be up to 50% cheaper than using the smallest size of an Inf1 instance.

Q: What ML models types and operators are supported by EC2 Inf1 instances using the Inferentia chip?

Inferentia chips support the commonly used machine learning models such as single shot detector (SSD) and ResNet for image recognition/classification and Transformer and BERT for natural language processing and translation and many others. A list of supported operators can be found on GitHub.

Q: How do I take advantage of AWS Inferentia's NeuronCore Pipeline capability to lower latency?

Inf1 instances with multiple Inferentia chips, such as Inf1.6xlarge or Inf1.24xlarge, support a fast chip-to-chip interconnect. Using the Neuron Processing Pipeline capability, you can split your model and load it to local cache memory across multiple chips. The Neuron compiler uses ahead-of-time (AOT) compilation technique to analyze the input model and compile it to fit across the on-chip memory of single or multiple Inferentia chips. Doing so enables the Neuron Cores to have high-speed access to models and not require access to off-chip memory, keeping latency bounded while increasing the overall inference throughput.

Q: What is the difference between AWS Neuron and Amazon SageMaker Neo?

AWS Neuron is a specialized SDK for AWS Inferentia chips that optimizes the machine learning inference performance of Inferentia chips. It consists of a compiler, run-time, and profiling tools for AWS Inferentia and is required to run inference workloads on EC2 Inf1 instances. On the other hand, Amazon SageMaker Neo is a hardware agnostic service that consists of a compiler and run-time that enables developers to train machine learning models once, and run them on many different hardware platforms.

Q: How do I use the Trainium chips in Trn1 instances?

The Trainium software stack, AWS Neuron SDK, integrates with leading ML frameworks, such as PyTorch and TensorFlow, so you can get started with minimal code changes. To get started quickly, you can use AWS Deep Learning AMIs and AWS Deep Learning Containers, which come preconfigured with AWS Neuron. If you are using containerized applications, you can deploy AWS Neuron by using Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), or your preferred native container engine. AWS Neuron also supports Amazon SageMaker, which you can use to build, train, and deploy machine learning models.

Q: Where can I deploy deep learning models trained on Trn1?

You can deploy deep learning models trained on Trn1 instances on any other Amazon EC2 instance that supports deep learning use cases, including instances based on CPUs, GPUs, or other accelerators. You can also deploy models trained on Trn1 instances outside of AWS, such as on-premises data centers or in embedded devices at the edge. For example, you can train your models on Trn1 instances and deploy them on Inf1 instances, G5 instances, G4 instances, or compute devices at the edge.

Q: When would I use Trn1 instances over GPU-based instances for training ML models?

Trn1 instances are a good fit for your natural language processing (NLP), large language model (LLM), and computer vision (CV) model training use cases. Trn1 instances focus on accelerating model training to deliver high performance while also lowering your model training costs. If you have ML models that need third-party proprietary libraries or languages, for example NVIDIA CUDA, CUDA Deep Neural Network (cuDNN), or TensorRT libraries, we recommend using the NVIDIA GPU-based instances (P4, P3).

Burstable instances

Q: How are Burstable Performance Instances different?

Amazon EC2 allows you to choose between Fixed Performance Instances (e.g. C, M and R instance families) and Burstable Performance Instances (e.g. T2). Burstable Performance Instances provide a baseline level of CPU performance with the ability to burst above the baseline.

T2 instances' baseline performance and ability to burst are governed by CPU Credits. Each T2 instance receives CPU Credits continuously, the rate of which depends on the instance size. T2 instances accrue CPU Credits when they are idle, and consume CPU credits when they are active. A CPU Credit provides the performance of a full CPU core for one minute.

Model	vCPUs	CPU Credits / hour	Maximum CPU Credit Balance	Baseline CPU Performance
t2.nano	1	3	72	5% of a core
t2.micro	1	6	144	10% of a core
t2.small	1	12	288	20% of a core

t2.medium	2	24	576	40% of a core*
t2.large	2	36	864	60% of a core**
t2.xlarge	4	54	1,296	90% of a core***
t2.2xlarge	8	81	1,944	135% of a core****

** For the t2.medium, single threaded applications can use 40% of 1 core, or if needed, multithreaded applications can use 20% each of 2 cores.*

***For the t2.large, single threaded applications can use 60% of 1 core, or if needed, multithreaded applications can use 30% each of 2 cores.*

**** For the t2.xlarge, single threaded applications can use 90% of 1 core, or if needed, multithreaded applications can use 45% each of 2 cores or 22.5% of all 4 cores.*

***** For the t2.2xlarge, single threaded applications can use all of 1 core, or if needed, multithreaded applications can use 67.5% each of 2 cores or 16.875% of all 8 cores.*

Q: How do I choose the right Amazon Machine Image (AMI) for my T2 instances?

You will want to verify that the minimum memory requirements of your operating system and applications are within the memory allocated for each T2 instance size (for example, 512 MiB for t2.nano). Operating systems with Graphical User Interfaces (GUI) that consume significant memory and CPU, for example Microsoft Windows, might need a t2.micro or larger instance size for many use cases. You can find AMIs suitable for the t2.nano instance types on AWS Marketplace. Windows customers who do not need the GUI can use the Microsoft Windows Server 2012 R2 Core AMI.

Q: When should I choose a Burstable Performance Instance, such as T2?

T2 instances provide a cost-effective platform for a broad range of general purpose production workloads. T2 Unlimited instances can sustain high CPU performance for as long as required. If your workloads consistently require CPU usage much higher than the baseline, consider a dedicated CPU instances such as the M or C.

Q: How can I see the CPU Credit balance for each T2 instance?

You can see the CPU Credit balance for each T2 instance in EC2 per-Instance metrics in Amazon CloudWatch. T2 instances have four metrics, CPUCreditUsage, CPUCreditBalance, CPUSurplusCreditBalance and CPUSurplusCreditsCharged. CPUCreditUsage indicates the amount of CPU Credits used. CPUCreditBalance indicates the balance of CPU Credits. CPUSurplusCredit Balance indicates credits used for bursting in the absence of earned credits. CPUSurplusCreditsCharged indicates credits that are charged when average usage exceeds the baseline.

Q: What happens to CPU performance if my T2 instance is running low on credits (CPU Credit balance is near zero)?

If your T2 instance has a zero CPU Credit balance, performance will remain at baseline CPU performance. For example, the t2.micro provides baseline CPU performance of 10% of a physical CPU core. If your instance's CPU Credit balance is approaching zero, CPU performance will be lowered to baseline performance over a 15-minute interval.

Q: Does my T2 instance credit balance persist at stop / start?

No, a stopped instance does not retain its previously earned credit balance.

Q: Can T2 instances be purchased as Reserved Instances or Spot Instances?

T2 instances can be purchased as On-Demand Instances, Reserved Instances or Spot Instances.

Q: What are Amazon EC2 T4g instances?

Amazon EC2 T4g instances are the next-generation of general purpose burstable instances powered by Arm-based AWS Graviton2 processors. T4g instances deliver up to 40% better price performance over T3 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for T4g instances?

T4g instances deliver up to 40% better price performance over T3 instances for a wide variety of burstable general purpose workloads such as micro-services, low-latency interactive applications, small and medium databases, virtual desktops, development environments, code repositories, and business-critical applications. Customers deploying applications built on open source software across T instances will find the T4g instances an appealing option to realize the best price performance. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: How can customers get access to the T4g free trial?

Until December 31, 2024, all AWS customers will be enrolled automatically in the T4g free trial as detailed in the AWS Free Tier. During the free-trial period, customers who run a t4g.small instance will automatically get 750 free hours per month deducted from their bill during each month. The 750 hours are calculated in aggregate across all Regions in which the t4g.small instances are used. Customers must pay for surplus CPU credits when they exceed the instances allocated credits during the 750 free hours of the T4g free trial program. For more information about how CPU credits work, see Key concepts and definitions for burstable performance instances in the Amazon EC2 User Guide for Linux Instances.

Q: Who is eligible for the T4g free trial?

All existing and new customers with an AWS account can take advantage of the T4g free trial. The T4g free trial is available for a limited time until December 31, 2024. The start and end time of the free trial are based on the Coordinated Universal Time (UTC). The T4g free trial will be available in addition to the existing AWS Free Tier on t2.micro/t3.micro. Customers who have exhausted their t2.micro (or t3.micro, depending on the Region) Free Tier usage can still benefit from the T4g free trial.

Q: What is the regional availability of T4g free trial?

The T4g free trial is currently available across these AWS Regions: US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), South America (Sao Paulo), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), and Europe (Stockholm). It is currently not available in the China (Beijing) and China (Ningxia) Regions.

As part of the free trial, customers can run t4g.small instances across one or multiple Regions from a single cumulative bucket of 750 free hours per month until December 31, 2024. For example, a customer can run t4g.small in Oregon for 300 hours for a month and run another t4g.small in Tokyo for 450 hours during the same month. This would add up to 750 hours per month of the free-trial limit.

Q: Is there an additional charge for running specific AMIs under the T4g free trial?

Under the t4g.small free trial, there will be no Amazon Machine Image (AMI) charge for Amazon Linux 2, RHEL and SUSE Linux AMIs that are available through the EC2 console Quick Start for the first 750 free hours per month. After 750 free hours per month, regular On-Demand prices, including AMI charge (if any), will apply. The applicable software fees for AWS Marketplace offers with AMI fulfillment options is not included in the free trial. Only the t4g.small infrastructure cost is included and covered under the free trial.

Q: How will the t4g.small free trial be reflected on my AWS bill?

The T4g free trial has a monthly billing cycle that starts on the first of every month and ends on the last day of that month. Under the T4g free-trial billing plan, customers using t4g.small will see a $0 line item on their bill under the On-Demand pricing plan for the first 750 aggregate hours of usage for every month during the free-trial period. Customers can start any time during the free-trial period and get 750 free hours for the remainder of that month. Any unused hours from the previous month will not be carried over. Customers can launch multiple t4g.small instances under the free trial. Customers will be notified automatically through email using AWS Budgets when their aggregate monthly usage reaches 85% of 750 free hours. When the aggregate instance usage exceeds 750 hours for the monthly billing cycle, customers will be charged based on regular On-Demand pricing for the exceeded hours for that month. For customers with a Compute Savings Plan or T4g Instance Savings Plan, Savings Plan (SV) discount will be applied to On-Demand pricing for hours exceeding the 750 free

trial hours. If customers have purchased the T4g Reserved Instance (RI) plan, the RI plan applies first to any usage on an hourly basis. For any remaining usage after the RI plan has been applied, the free trial billing plan is in effect.

Q: If customers sign up for consolidated billing (or a single payer account), can they get the T4g free trial for each account that is tied to the payer account?

No, customers who use consolidated billing to consolidate payment across multiple accounts will have access to one free trial per Organization. Each payer account gets a total aggregate of 750 free hours a month. For more details about consolidated billing, see Consolidated billing for AWS Organizations in the AWS Billing and Cost Management User Guide.

Q: Will customers get charged for surplus CPU credits as a part of T4g free trial?

Customers must pay for surplus CPU credits when they exceed the instances allocated credits during the 750 free hours of the T4g free trial program. For details about how CPU credits work, see Key concepts and definitions for burstable performance instances in the Amazon EC2 User Guide for Linux Instances.

Q: At the end of the free trial, how will customers be billed for t4g.small instances?

Starting January 1, 2025, customers running on t4g.small instances will be automatically switched from the free trial plan to the On-Demand pricing plan (or Reserved Instance (RI)/Savings Plan (SV) plan, if purchased). Accumulated credits will be set to zero. Customers will receive an email notification seven days before the end of the free trial period stating that the free trial period will be ending in seven days. Starting January 1, 2025, if the RI plan is purchased, the RI plans will apply. Otherwise, customers will be charged regular On-Demand pricing for t4g.small instances. For customers who have the T4g Instance Savings Plan or a Compute Savings Plan, t4g.small instance billing will apply the Savings Plan discount on their On-Demand pricing.

Compute Optimized instances

Q: When should I use Compute Optimized instances?

Compute Optimized instances are designed for applications that benefit from high compute power. These applications include compute-intensive applications like high-performance web servers, high-performance computing (HPC), scientific modelling, distributed analytics and machine learning inference.

Q: What are Amazon EC2 C7g instances?

Amazon EC2 C7g instances, powered by the latest generation AWS Graviton3 processors, provide the best price performance in Amazon EC2 for compute-intensive workloads. C7g instances are ideal for high performance computing (HPC), batch processing, electronic design automation (EDA), gaming, video encoding, scientific modeling, distributed analytics, CPU-based machine learning (ML) inference, and ad-serving. They offer up to 25% better performance over the sixth generation AWS Graviton2-based C6g instances.

Q: What are Amazon EC2 C6g instances?

Amazon EC2 C6g instances are the next-generation of compute-optimized instances powered by Arm-based AWS Graviton2 Processors. C6g instances deliver up to 40% better price performance over C5 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for C6g instances?

C6g instances deliver significant price performance benefits for compute-intensive workloads such as high performance computing (HPC), batch processing, ad serving, video encoding, gaming, scientific modelling, distributed analytics, and CPU-based machine learning inference. Customers deploying applications built on open source software across C instances family will find the C6g instances an appealing option to realize the best price performance. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on C6g instances?

C6g instances are EBS-optimized by default and offer up to 19,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. C6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage are also available through the C6gd instance types.

Q: Which network interface is supported on C6g instances?

C6g instances support ENA based Enhanced Networking. With ENA, C6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the C6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS, Amazon ECR) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. Applications that are based on interpreted languages (such as Java, Node, Python) not reliant on native CPU instruction sets should run with minimal to no changes. Applications developed using compiled languages (C, C++, GoLang) will need to be re-compiled to generate Arm binaries. The Arm architecture is well supported in these popular programming languages and modern code usually requires a simple 'Make' command. Refer to the Getting Started guide on GitHub for more details.

Q: Will there be more compute choices offered with the C6 instance families?

Yes, we plan to offer Intel and AMD CPU powered instances in the future as part of the C6 instance families.

Q: Can I launch C4 instances as Amazon EBS-optimized instances?

Each C4 instance type is EBS-optimized by default. C4 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on C4 instances, launching a C4 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q: How can I use the processor state control feature available on the c4.8xlarge instance?

The c4.8xlarge instance type provides the ability for an operating system to control processor C-states and P-states. This feature is currently available only on Linux instances. You may want to change C-state or P-state settings to increase processor performance consistency, reduce latency, or tune your instance for a specific workload. By default, Amazon Linux provides the highest-performance configuration that is optimal for most customer workloads; however, if your application would benefit from lower latency at the cost of higher single- or dual-core frequencies, or from lower-frequency sustained performance as opposed to bursty Turbo Boost frequencies, then you should consider experimenting with the C-state or P-state configuration options that are available to these instances. For additional information on this feature, see the Amazon EC2 User Guide section on Processor State Control.

Q: Which instances are available within Compute Optimized instances category?

C6g instances: Amazon EC2 C6g instances are powered by Arm-based AWS Graviton2 processors. They deliver up to 40% better price performance over C5 instances and are ideal for running advanced compute-intensive workloads. This includes workloads such as high performance computing (HPC), batch processing, ad serving, video encoding, gaming, scientific modelling, distributed analytics, and CPU-based machine learning inference.

C6a instances: C6a instances are powered by 3rd generation AMD EPYC processors with an all-core turbo frequency of 3.6 GHz, offer up to 15% better price performance over C5a instances for a wide variety of workloads, and support always-on memory encryption using AMD Transparent Single Key Memory Encryption (TSME). C6a instances provide new instance sizes with up to 192 vCPUs and 384 GiB of memory, double that of the largest C5a instance. C6a also gives customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, more than twice that of C5a instances.

C6i instances: C6i instances are powered by 3rd generation Intel Xeon Scalable processors with an all-core turbo frequency of 3.5 GHz, offer up to 15% better price performance over C5 instances for a wide variety of workloads, and always-on memory encryption using Intel Total Memory encryption (TME). C6i instances provide a new instance size (c6i.32xlarge) with 128 vCPUs and 256 GiB of memory, 33% more than the largest C5 instance. They also provide up to 9% higher memory bandwidth per vCPU compared to C5 instances. C6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of C5 instances. C6i are also

available with local NVMe-based SSD block-level storage (C6id instances) for applications that need high-speed, low-latency local storage. Compared to previous generation C5d instances, C6id instances offer up to 138% higher TB storage per vCPU and 56% lower cost per TB.

C5 instances: C5 instances are based on Intel Xeon Platinum processors, part of the Intel Xeon Scalable (codenamed Skylake-SP or Cascade Lake) processor family, are available in 9 sizes, and offer up to 96 vCPUs and 192 GiB memory. C5 instances deliver 25% improvement in price/performance compared to C4 instances. The C5d instances have local NVMe storage for workloads that require very low latency and storage access with high random read and write IOPS ability.

C5a instances: C5a instances deliver leading x86 price-performance for a broad set of compute-intensive workloads including batch processing, distributed analytics, data transformations, log analysis, and web applications. C5a instances feature 2nd Gen 3.3GHz AMD EPYC processors with up to 96 vCPUs and up to 192 GiB of memory. The C5ad instances have local NVMe storage for workloads that require very low latency and storage access with high random read and write IOPS ability.

C5n instances: C5n instances are ideal for applications requiring high network bandwidth and packet rate. The C5n instances are ideal for applications like HPC, data lakes, network appliances as well as applications that require inter-node communication and the Message Passing Interface (MPI). C5n offer a choice of Intel Xeon Platinum 3.0 GHz processors with up to 72 vCPUs and 192GiB of Memory.

C4 instances: C4 instances are based on Intel Xeon E5-2666 v3 (codenamed Haswell) processors. C4 instances are available in 5 sizes and offer up to 36 vCPUs and 60 GiB memory.

Q: Why should customers choose C6i instances over C5 instances?

C6i instances offer up to 15% better price performance over C5 instances, and always-on memory encryption using Intel Total Memory encryption (TME). C6i instances provide a new instance size (c6i.32xlarge) with 128 vCPUs and 256 GiB of memory, 33% more than the largest C5 instance. They also provide up to 9% higher memory bandwidth per vCPU compared to C5 instances. C6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of C5 instances.

Q: Why should customers choose C5 instances over C4 instances?

The generational improvement in CPU performance and lower price of C5 instances, which combined result in a 25% price/performance improvement relative to C4 instances, benefit a broad spectrum of workloads that currently run on C3 or C4 instances. For floating point intensive applications, Intel AVX-512 enables significant improvements in delivered TFLOPS by effectively extracting data level parallelism. Customers looking for absolute performance for graphics rendering and HPC workloads that can be accelerated with GPUs or FPGAs should also evaluate other instance families in the Amazon EC2 portfolio that include those resources to find the ideal instance for their workload.

Q: Which storage interface is supported on C5 instances?

C5 instances will support only NVMe EBS device model. EBS volumes attached to C5 instances will appear as NVMe devices. NVMe is a modern storage interface that provides latency reduction and results in increased disk I/O and throughput.

Q: Why does the total memory reported by the operating system not exactly match the advertised memory on instance types?

Portions of the EC2 instance memory are reserved and used by the virtual BIOS for video RAM, DMI, and ACPI. In addition, for instances that are powered by the AWS Nitro Hypervisor, a small percentage of the instance memory is reserved by the Amazon EC2 Nitro Hypervisor to manage virtualization.

High Performance Computing Optimized instances

Q: Which instances are available within the high performance computing (HPC) instances category?

Hpc7g instances: Hpc7g instances enable the best price performance for HPC workloads on AWS. They deliver up to 70% better performance and almost 3x better price performance compared to previous-generation AWS Graviton-based instances for compute-intensive HPC workloads. Hpc7g instances are powered by AWS Graviton 3E processors and provide up to 35% higher vector instruction performance compared to existing AWS Graviton3 instances. These instances provide up to 2x better floating-point performance compared to instances powered by Graviton2 processors. Hpc7g instances are built on the AWS Nitro System and provide 200 Gbps network bandwidth for low-latency internode communication for tightly coupled workloads that require highly parallelized, clustered compute resources.

Hpc7a instances: Amazon Elastic Compute Cloud (Amazon EC2) Hpc7a instances, powered by 4th Gen AMD EPYC processors, deliver up to 2.5x better performance compared to Amazon EC2 Hpc6a instances. Hpc7a instances feature 2x higher core density (up to 192 cores), 2.1x higher memory bandwidth throughput (up to 768 GB of memory), and 3x higher network bandwidth compared to Hpc6a instances. These instances offer 300 Gbps of Elastic Fabric Adapter (EFA) network bandwidth, powered by the AWS Nitro System, for fast and low latency inter-node communications.

Hpc6id instances: Hpc6id instances are powered by 64 cores of Intel 3rd Gen Xeon Scalable processors that run at frequencies up to 3.5 GHz for increased efficiency. These instances are designed to improve performance for memory-bound workloads by offering 5 GB/s memory bandwidth per vCPU. Hpc6id instances offer 200 Gbps EFA networking for high-throughput internode communications to help you run your HPC workloads at scale.

Hpc6a instances: Hpc6a instances are powered by 96 cores of 3rd Gen AMD EPYC processors with an all-core turbo frequency of 3.6 GHz and 384 GiB RAM. Hpc6a instances offer 100 Gbps EFA networking enabled for high throughput internode communications to help you run your HPC workloads at scale.

Q: How are Hpc7g instances different from other EC2 instances?

Hpc7g instances are optimized to deliver capabilities suited for compute-intensive HPC workloads. Hpc7g instances are based on Arm-based Graviton3E processors that provide up to 35% higher vector instruction performance compared to existing instances based on Graviton3 processors. These instances deliver 64 physical cores, 128 GiB memory, and 200 Gbps network bandwidth optimized for traffic between instances in the same VPC and support EFA for increased network performance. Hpc7g instances are available in single Availability Zone deployments, enabling workloads to achieve the low-latency network performance necessary for tightly coupled node-to-node communication for HPC applications.

Q: Which pricing models do Hpc7g instances support?

Hpc7g instances are available for purchase through the 1- and 3-year Amazon EC2 Instance Savings Plans, Compute Savings Plans, EC2 On-Demand Instances, and EC2 Reserved Instances.

Q: Which AMIs are supported on Hpc7g instances?

Hpc7g instances support Amazon EBS backed AMIs only.

Q: How are Hpc7a instances different from other EC2 instances?

HPC-optimized EC2 Hpc7a instances are ideal for applications that benefit from high-performance processors such as large, complex simulations including computational fluid dynamics (CFD), numerical weather prediction, and multiphysics simulations. Hpc7a instances are designed to help you run tightly coupled, x86-based HPC workloads with better performance. Hpc7a instances feature 4th Gen AMD EPYC processors with 2x higher core density (up to 192 cores), 2.1x higher memory bandwidth throughput (768 GB of memory), and 3x higher network bandwidth compared to Hpc6a instances. These instances offer 300 Gbpsof EFA network bandwidth, powered by the AWS Nitro System, for for fast and low latency internode communications.

Q: Which pricing models do Hpc7a instances support?

Hpc7a instances are available for purchase through the 1- and 3-year Amazon EC2 Instance Savings Plans, Compute Savings Plans, EC2 On-Demand Instances, and EC2 Reserved Instances.

Q: Which AMIs are supported on Hpc7a instances?

Hpc7a instances support Amazon Linux 2, Amazon Linux, Ubuntu 18.04 or later, Red Hat Enterprise Linux 7.6 or later, SUSE Linux Enterprise Server 12 SP3 or later, CentOS 7 or later, and FreeBSD 11.1 or later.

Q: Which pricing models do Hpc6id instances support?

Hpc6id instances are available for purchase through the 1-year and 3-year Amazon EC2 Instance Savings Plans, Compute Savings Plans, EC2 On-Demand Instances, and EC2 Reserved Instances.

Q: How are Hpc6id instances different from other EC2 instances?

Hpc6id instances are optimized to deliver capabilities suited for memory-bound, data-intensive HPC workloads. Hyperthreading is disabled to increase per-vCPU CPU throughput and up to 5 GB/s memory bandwidth per vCPU. These instances deliver 200 Gbps network bandwidth optimized for traffic between instances in the same virtual private cloud (VPC), and support EFA for increased network performance. To optimize Hpc6id instances networking for tightly coupled workloads, you can access EC2 Hpc6id instances in a single Availability Zone in each Region.

Q: Which AMIs are supported on Hpc6id instances?

Hpc6id supports Amazon Linux 2, Amazon Linux, Ubuntu 18.04 or later, Red Hat Enterprise Linux 7.4 or later, SUSE Linux Enterprise Server 12 SP2 or later, CentOS 7 or later, Windows Server 2008 R2 or earlier, and FreeBSD 11.1 or later.

Q: Which AMIs are supported on Hpc6a instances?

Hpc6a instances support Amazon Linux 2, Amazon Linux, Ubuntu 18.04 or later, Red Hat Enterprise Linux 7.4 or later, SUSE Linux Enterprise Server 12 SP2 or later, CentOS 7 or later, and FreeBSD 11.1 or later. These instances also support Windows Server 2012, 2012 R2, 2016, and 2019.

Q: Which pricing models do Hpc6a instances support?

Hpc6a instances are available for purchase through 1-year and 3-year Standard Reserved Instances, Convertible Reserved Instances, Savings Plans, and On-Demand Instances.

General Purpose instances

Q: How do M7i instances compare to M7i-flex? When should I use M7i-flex instead of M7i instances?

M7i-flex instances are a lower-priced variant of M7i instances that offer 19% better price performance over M6i instances. M7i-flex instances can be used to run a majority of general-purpose workloads that benefit from the latest generation performance but do not fully utilize compute resources. M7i-flex instances are designed to deliver a baseline CPU performance with the ability to scale up to the full CPU performance 95% of the time. M7i-flex instances are ideal for workloads that fit on instance sizes up to 8xlarge (32 vCPUs and 128 GB), including web and application servers, virtual desktops, microservices, databases, and enterprise applications. You can use M7i instances for workloads that need the largest instance sizes or high sustained CPU, network, or EBS performance, such as large application servers, large databases, gaming servers, CPU-based machine learning, and video streaming.

Q: What performance do M7i-flex instances provide?

M7i-flex instances provide reliable CPU resources to deliver a baseline CPU performance of 40%, designed to meet the compute requirements of the majority of general-purpose workloads. For times when workloads need more performance, M7i-flex instances provide the ability to scale up to 100% CPU for 95% of the time over a 24-hour window.

Q: What are some other use cases for M7i-flex instances?

The M7i-flex instances provide a compelling upgrade path for workloads running on T3 larger-sized instances (large to 2xlarge) by offering better price performance, a fixed hourly price that includes baseline CPU and additional CPU usage beyond baseline, and larger instance sizes up to 8xlarge (32vCPUs and 128 GB). M7i-flex instances offer a simplified way to optimize your EC2 usage without CPU credits.

Q: What are Amazon EC2 M6g instances?

Amazon EC2 M6g instances are the next-generation of general-purpose instances powered by Arm-based AWS Graviton2 Processors. M6g instances deliver up to 40% better price/performance over M5 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are the specifications of the new AWS Graviton2 Processors?

The AWS Graviton2 processors deliver up to 7x performance, 4x the number of compute cores, 2x larger caches, 5x faster memory, and 50% faster per core encryption performance than first generation AWS Graviton processors. Each core of the AWS Graviton2 processor is a single-threaded vCPU. These processors also offer always-on fully encrypted DRAM memory, hardware acceleration for compression workloads, dedicated engines per vCPU that double the floating-point performance for workloads such as video encoding, and instructions for int8/fp16 CPU-based machine learning inference acceleration. The CPUs are built utilizing 64-bit Arm Neoverse cores and custom silicon designed by AWS on the advanced 7 nm manufacturing technology.

Q: Is memory encryption supported by AWS Graviton2 processors?

AWS Graviton2 processors support always-on 256-bit memory encryption to further enhance security. Encryption keys are securely generated within the host system, do not leave the host system, and are irrecoverably destroyed when the host is rebooted or powered down. Memory encryption does not support integration with AWS Key Management Service (AWS KMS) and customers cannot bring their own keys.

Q: What are some of the ideal use cases for M6g instances?

M6g instances deliver significant performance and price performance benefits for a broad spectrum of general-purpose workloads such as application servers, gaming servers, microservices, mid-size databases, and caching fleets. Customers deploying applications built on open source software across the M instances will find the M6g instances an appealing option to realize the best price performance. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on M6g instances?

M6g instances are EBS-optimized by default and offer up to 19,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. M6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage are also available through the M6gd instance types.

Q: Which network interface is supported on M6g instances?

M6g instances support ENA based Enhanced Networking. With ENA, M6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the M6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS, Amazon ECR) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. Applications that are based on interpreted languages (such as Java, Node, Python) not reliant on native CPU instruction sets should run with minimal to no changes. Applications developed using compiled languages (C, C++, GoLang) will need to be re-compiled to generate Arm binaries. The Arm architecture is well supported in these popular programming languages and modern code usually requires a simple 'Make' command. Refer to the Getting Started guide on GitHub for more details.

Q: What are Amazon EC2 A1 instances?

Amazon EC2 A1 instances are general purpose instances powered by the first-generation AWS Graviton Processors that are custom designed by AWS.

Q: What are the specifications of the first-generation AWS Graviton Processors?

AWS Graviton processors are custom designed by AWS utilizing Amazon's extensive expertise in building platform solutions for cloud applications running at scale. These processors are based on the 64-bit Arm instruction set and feature Arm Neoverse cores as well as custom silicon designed by AWS. The cores operate at a frequency of 2.3 GHz.

Q: When should I use A1 instances?

A1 instances deliver significant cost savings for scale-out workloads that can fit within the available memory footprint. A1 instances are ideal for scale-out applications such as web servers, containerized microservices, and data/log processing. These instances will also appeal to

developers, enthusiasts, and educators across the Arm developer community.

Q: Will customers have to modify applications and workloads to be able to run on the A1 instances?

The changes required are dependent on the application. Applications based on interpreted or run-time compiled languages (e.g. Python, Java, PHP, Node.js) should run without modifications. Other applications may need to be recompiled and those that don't rely on x86 instructions will generally build with minimal to no changes.

Q: Which operating systems/AMIs are supported on A1 Instances?

The following AMIs are supported on A1 instances: Amazon Linux 2, Ubuntu 16.04.4 or newer, Red Hat Enterprise Linux (RHEL) 7.6 or newer, SUSE Linux Enterprise Server 15 or newer. Additional AMI support for Fedora, Debian, NGINX Plus are also available through community AMIs and the AWS Marketplace. EBS backed HVM AMIs launched on A1 instances require NVMe and ENA drivers installed at instance launch.

Q: Are there specific AMI requirements to run on M6g and A1 instances?

You will need to use the “arm64” AMIs with the M6g and A1 instances. x86 AMIs are not compatible with M6g and A1 instances.

Q: When should customers use A1 instances versus the new M6g instances?

A1 instances continue to offer significant cost benefits for scale-out workloads that can run on multiple smaller cores and fit within the available memory footprint. The new M6g instances are a good fit for a broad spectrum of applications that require more compute, memory, networking resources and/or can benefit from scaling up across platform capabilities. M6g instances will deliver the best price-performance within the instance family for these applications. M6g supports up to 16xlarge instance size (A1 supports up to 4xlarge), 4GB of memory per vCPU (A1 supports 2GB memory per vCPU), and up to 25 Gbps of networking bandwidth (A1 supports up to 10 Gbps).

Q: What are the various storage options available to A1 customers?

A1 instances are EBS-optimized by default and offer up to 3,500 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. A1 instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. A1 instances will not

support the blkfront interface.

Q: Which network interface is supported on A1 instances?

A1 instances support ENA based Enhanced Networking. With ENA, A1 instances can deliver up to 10 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Do A1 instances support the AWS Nitro System?

Yes, A1 instances are powered by the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: Why should customers choose EC2 M5 Instances over EC2 M4 Instances?

Compared with EC2 M4 Instances, the new EC2 M5 Instances deliver customers greater compute and storage performance, larger instance sizes for less cost, consistency and security. The biggest benefit of EC2 M5 Instances is based on its usage of the latest generation of Intel Xeon Scalable processors (Skylake-SP or Cascade Lake), which deliver up to 20% improvement in price/performance compared to M4. With AVX-512 support in M5 vs. the older AVX2 in M4, customers will gain 2x higher performance in workloads requiring floating point operations. M5 instances offer up to 25 Gbps of network bandwidth and up to 10 Gbps of dedicated bandwidth to Amazon EBS. M5 instances also feature significantly higher networking and Amazon EBS performance on smaller instance sizes with EBS burst capability.

Q: Why should customers choose M6i instances over M5 instances?

Amazon M6i instances are powered by 3rd generation Intel Xeon Scalable processors (code named Ice Lake) with an all-core turbo frequency of 3.5 GHz, offer up to 15% better compute price performance over M5 instances, and always-on memory encryption using Intel Total Memory Encryption (TME). Amazon EC2 M6i instances are the first to use a lower-case “i” to indicate they are Intel-powered instances. M6i instances provide a new instance size (m6i.32xlarge) with 128 vCPUs and 512 GiB of memory, 33% more than the largest M5 instance. They also provide up to 20% higher memory bandwidth per vCPU compared to M5 instances, allowing customers to efficiently perform real-time analysis for data-intensive AI/ML, gaming, and High Performance Computing (HPC) applications. M6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of M5 instances. M6i also allows customers to use

Elastic Fabric Adapter on the 32xlarge size, enabling low latency and high scale inter-node communication. For optimal networking performance on these new instances, Elastic Network Adapter (ENA) driver update may be required. For more information on optimal ENA driver for M6i, see this article.

Q: How does support for Intel AVX-512 benefit customers who use the EC2 M5 family or the M6i family?

Intel Advanced Vector Extensions 512 (AVX-512) is a set of new CPU instructions available on the latest Intel Xeon Scalable processors, that can accelerate performance for workloads and usages such as scientific simulations, financial analytics, artificial intelligence, machine learning/deep learning, 3D modeling and analysis, image and video processing, cryptography and data compression, among others. Intel AVX-512 offers exceptional processing of encryption algorithms, helping to reduce the performance overhead for cryptography, which means customers who use the EC2 M5 family or M6i family can deploy more secure data and services into distributed environments without compromising performance.

Q: What are M5zn instances?

M5zn instances are a variant of the M5 general purpose instances that are powered by the fastest Intel Xeon Scalable processor in the cloud, with an all-core turbo frequency of up to 4.5 GHz, along with 100 Gbps networking and support for Amazon EFA. M5zn instances are an ideal fit for workloads such as gaming, financial applications, simulation modeling applications such as those used in the automotive, aerospace, energy, and telecommunication industries, and other High Performance Computing applications.

Q: How are M5zn instances different than z1d instances?

z1d instances are a memory-optimized instance, and feature a high frequency version of the Intel Xeon Scalable processors (up to 4.0 GHz), along with local NVMe storage. M5zn instances are a general purpose instance, and feature a high frequency version of the 2nd Generation Intel Xeon Scalable processors up to 4.5 GHz), along with up to 100 Gbps networking performance, and support for EFA. M5zn instances offer improved price performance compared to z1d.

High Memory instances

Q: What are EC2 High Memory instances?

Amazon EC2 High Memory instances offer 3, 6, 9, 12, 18, or 24 TiB of memory in a single instance. These instances are designed to run large in-memory databases, including production installations of SAP HANA, in the cloud.

EC2 High Memory instances with 3, 6, 9, and 12 TiB of memory are powered by an 8-socket platform with Intel® Xeon® Platinum 8176M (Skylake) processors. EC2 High Memory instances with 18 and 24 TiB of memory are the first Amazon EC2 instances powered by an 8-socket platform with 2nd Generation Intel® Xeon® Scalable (Cascade Lake) processors that are optimized for mission-critical enterprise workloads. EC2 High Memory instances deliver high networking throughput and low-latency with up to 100 Gbps of aggregate network bandwidth using Amazon Elastic Network Adapter (ENA)-based Enhanced Networking. EC2 High Memory instances are EBS-Optimized by default, and support encrypted and unencrypted EBS volumes.

Q: Are High Memory instances certified by SAP to run SAP HANA workloads?

High Memory instances are certified by SAP for running Business Suite on HANA, the next-generation Business Suite S/4HANA, Data Mart Solutions on HANA, Business Warehouse on HANA, and SAP BW/4HANA in production environments. For details, see SAP's Certified and Supported SAP HANA Hardware Directory.

Q: What instance types are available for High Memory instances?

High Memory instances are available as both bare metal and virtualized instances, giving customers the choice to have direct access to the underlying hardware resources, or to take advantage of the additional flexibility that virtualized instances offer including On-Demand and 1-year and 3-year Savings Plan purchase options. Please check out available options for High Memory instances in the Memory optimized section of EC2 Instance types page.

Q: What are some of the benefits of using High Memory Virtualized instances over High Memory Bare Metal instances?

Benefits of High Memory virtual instances over High Memory Metal instances include – significantly better launch/reboot times, flexible purchase options (On-Demand, Savings Plan, Reserved Instances, Dedicated Hosts), choice of tenancy type, self-service options and support

for a higher number of EBS volumes (27 vs 19).

Q: When should a High Memory 'Metal' instance be used vs using High Memory 'Virtualized' instance?

Though High Memory 'Virtualized' instances are in-general recommended to be used, there are specific situations where only High Memory Metal instances can work. These situation include – when using OS versions that are not supported on High Memory Virtual instances OR when using applications that need to run in non-virtualized mode to meet licensing / support requirements OR when using applications that require access to hardware feature set (such as Intel VT-x) OR when using custom hypervisor (e.g, ESXi).

Q: How do I migrate from High Memory metal instances to High Memory virtualized instances?

You can migrate your High Memory metal instance to a virtualized instance in just few steps. 1/Stop your instance, 2/ Change the instance and tenancy type through EC2 API and 3/ Start your instance back up. If you are using Red Hat Enterprise Linux for SAP or SUSE Linux Enterprise Server for SAP, you need to ensure that your operating system and kernel versions are compatible with virtualized High Memory instances. For further details, see Migrating SAP HANA on AWS to an EC2 High Memory Instance documentation.

Q: What are the storage options available with High Memory instances?

High Memory instances support Amazon EBS volumes for storage. High Memory instances are EBS-optimized by default, and offer up to 38 Gbps of storage bandwidth.:

Q: Which storage interface is supported on High Memory instances?

High Memory instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces. EBS volumes attached to High Memory instances appear as NVMe devices. NVMe is an efficient and scalable storage interface, which is commonly used for flash based SSDs and provides latency reduction and results in increased disk I/O and throughput. The EBS volumes are attached and detached by PCI hotplug.

Q: What network performance is supported on High Memory instances?

High Memory instances use the Elastic Network Adapter (ENA) for networking and enable Enhanced Networking by default. With ENA, High Memory instances can utilize up to 100 Gbps of network bandwidth.

Q: Can I run High Memory instances in my existing Amazon Virtual Private Cloud (Amazon VPC)?

You can run High Memory instances in your existing and new Amazon VPCs.

Q: What is the underlying hypervisor on High Memory instances?

High Memory instances use the lightweight Nitro Hypervisor that is based on core KVM technology.

Q: Do High Memory instances enable CPU power management state control?

Yes. You can configure C-states and P-states on High Memory instances. You can use C-states to enable higher turbo frequencies (as much as 4.0 GHz). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

Q: What purchase options are available for High Memory instances?

EC2 High Memory virtualized instances (e.g. u-6tb1.112xlarge) are available for purchase via On-Demand, 1-Yr and 3-Yr Savings Plan, and 1-Yr and 3-Yr Reserved Instance. EC2 High Memory metal instances (e.g. u-6tb1.metal) are only available for purchase as EC2 Dedicated Hosts on a 1-Yr and 3-Yr Reservation.

Q: What is the lifecycle of a Dedicated Host?

Once a Dedicated Host is allocated within your account, it will be standing by for your use. You can then launch an instance with a tenancy of "host" using the RunInstances API, and can also stop/start/terminate the instance through the API. You can use the AWS Management Console to manage the Dedicated Host and the instance.

Q: Can I launch, stop/start, and terminate High Memory instances using AWS CLI/SDK?

You can launch, stop/start, and terminate instances using AWS CLI/SDK.

Q: Which AMIs are supported with High memory instances?

EBS-backed HVM AMIs with support for ENA networking can be used with High Memory instances. The latest Amazon Linux, Red Hat Enterprise Linux, SUSE Enterprise Linux Server, and Windows Server AMIs are supported. Operating system support for SAP HANA workloads on High Memory instances include: SUSE Linux Enterprise Server 12 SP3 for SAP, Red Hat Enterprise Linux 7.4 for SAP, Red Hat Enterprise Linux 7.5 for SAP, SUSE Linux Enterprise Server 12 SP4 for SAP, SUSE Linux Enterprise Server 15 for SAP, Red Had Enterprise Linux 7.6 for SAP. Refer to SAP's Certified and Supported SAP HANA Hardware Directory for latest detail on supported operating systems.

Q: Are there standard SAP HANA reference deployment frameworks available for the High Memory instance and the AWS Cloud?

You can use the AWS Quick Start reference SAP HANA deployments to rapidly deploy all the necessary SAP HANA building blocks on High Memory instances following SAP's recommendations for high performance and reliability. AWS Quick Starts are modular and customizable, so you can layer additional functionality on top or modify them for your own implementations.

Memory Optimized instances

Q: When should I use memory-optimized instances?

Memory-optimized instances offer large memory size for memory intensive applications including in-memory applications, in-memory databases, in-memory analytics solutions, HPC, scientific computing, and other memory-intensive applications.

Q: What are Amazon EC2 R6g instances?

Amazon EC2 R6g instances are the next-generation of memory-optimized instances powered by Arm-based AWS Graviton2 Processors. R6g instances deliver up to 40% better price performance over R5 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for R6g instances?

R6g instances deliver significant price performance benefits for memory-intensive workloads such as instances and are ideal for running memory-intensive workloads such as open-source databases, in-memory caches, and real time big data analytics. Customers deploying applications built on open source software across R instances will find the R6g instances an appealing option to realize the best price performance within the instance family. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on R6g instances?

R6g instances are EBS-optimized by default and offer up to 19,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. R6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage are also available through the R6gd instance types.

Q: Which network interface is supported on R6g instances?

R6g instances support ENA based Enhanced Networking. With ENA, R6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the R6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS, Amazon ECR) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. Applications that are based on interpreted languages (such as Java, Node, Python) not reliant on native CPU instruction sets should run with minimal to no changes. Applications developed using compiled languages (C, C++, GoLang) will need to be re-compiled to generate Arm binaries. The Arm architecture is well supported in these popular programming languages and modern code usually requires a simple ‘Make’ command. Refer to the Getting Started guide on GitHub for more details.

Q: Why should you choose R6i instances over R5 instances?

Amazon R6i instances are powered by 3rd Generation Intel Xeon Scalable processors (Ice Lake) with an all-core turbo frequency of 3.5 GHz, offer up to 15% better compute price performance over R5 instances, and always-on memory encryption using Intel Total Memory Encryption (TME). Amazon EC2 R6i instances use a lower-case “i” to indicate they are Intel-powered instances. R6i instances provide a new instance size (r6i.32xlarge) with 128 vCPUs and 1,024 GiB of memory, 33% more than the largest R5 instance. They also provide up to 20% higher memory bandwidth per vCPU compared to R5 instances, allowing you to efficiently perform real-time analysis for data-intensive AI/ML, gaming, and high performance computing (HPC) applications. R6i instances also give you up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of R5 instances. With R6i instances, you can use Elastic Fabric Adapter allows customers to use Elastic Fabric Adapter (EFA) on the 32xlarge and metal sizes, enabling low-latency and high-scale inter-node communication. For optimal networking performance on these new instances, Elastic Network Adapter (ENA) driver update may be required. For more information about an optimal ENA driver for R6i, see "What do I need to do before migrating my EC2 instance to a sixth-generation instance?" on Knowledge Center.

Q: What are Amazon EC2 R5b instances?

R5b instances are EBS-optimized variants of memory-optimized R5 instances that deliver up to 3x better EBS performance compared to same sized R5 instances. R5b instances deliver up to 60 Gbps bandwidth and 260K IOPS of EBS performance, the fastest block storage performance on EC2. They are built on the AWS Nitro System, which is a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for R5b instances?

R5b instances are ideal for large relational database workloads, including Microsoft SQL Server, SAP HANA, IBM DB2, and Oracle that run performance intensive applications such as commerce platforms, ERP systems, and health record systems. Customers looking to migrate large on-premises workloads with large storage performance requirements to AWS will find R5b instances to be a good fit.

Q: What are the various storage options available on R5b instances?

R5b instances are EBS-optimized by default and offer up to 60,000 Mbps of dedicated EBS bandwidth and 260K IOPS for both encrypted and unencrypted EBS volumes. R5b instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. R5b is supported by all volume types, with the exception of io2 volumes.

Q: When should I use R5b instances?

Customers running workloads such as large relational databases and data analytics that want to take advantage of the increased EBS storage network performance can use R5b instances to deliver higher performance and bandwidth. Customers can also lower costs by migrating their workloads to smaller size R5b instances or by consolidating workloads on fewer R5b instances.

Q: What are the storage options available with High Memory instances?

High Memory instances support Amazon EBS volumes for storage. High Memory instances are EBS-optimized by default, and offer up to 38Gbps of storage bandwidth to both encrypted and unencrypted EBS volumes.

Q: What are Amazon EC2 X2gd instances?

Amazon EC2 X2gd instances are the next generation of memory-optimized instances powered by AWS-designed Arm-based AWS Graviton2 processors. X2gd instances deliver up to 55% better price performance compared to x86-based X1 instances and offer the lowest cost per GiB of memory in Amazon EC2. They are the first of the X instances to be built on the AWS Nitro System, which is a combination of dedicated hardware and Nitro hypervisor.

Q: What workloads are suited for X2gd instances?

X2gd is ideal for customers with Arm-compatible memory bound scale-out workloads such as Redis and Memcached in-memory databases, that need low latency memory access and benefit from more memory per vCPU. X2gd is also well suited for relational databases such as PostgreSQL, MariaDB, MySQL, and RDS Aurora. Customers who run memory intensive workloads such as Apache Hadoop, real-time analytics, and real-time caching servers will benefit from 1:16 vCPU to memory ratio of X2gd. Single threaded workloads such as EDA backend

verification jobs will benefit from physical core and more memory of X2gd instances, allowing them to consolidate more workloads on to a single instance. X2gd instance also feature local NVMe SSD block storage to improve response times by acting as a caching layer.

Q: When should I use X2gd instances compared to the X1, X2i, or R instances?

X2gd instances are suitable for Arm-compatible memory bound scale-out workloads such as in-memory databases, memory analytics applications, open-source relational database workloads, EDA workloads, and large caching servers. X2gd instances offer customers the lowest cost per gigabyte of memory within EC2, with sizes up to 1 TiB. X2iezn, X2idn, X2iedn, X1, and X1e instances use x86 processors and are suitable for memory-intensive enterprise-class, scale-up workloads such as Windows workloads, in-memory databases (e.g. SAP HANA), and relational databases (e.g. OracleDB). Customers can leverage the x86-based X instances for larger memory sizes up to 4 TiB. R6g and R6gd instances are suitable for workloads such as web applications, databases, and search indexing queries that need more vCPUs during times of heavy data processing. Customers running memory bound workloads that need less than 1 TiB memory and have dependency on x86 instruction set such as Windows applications, and applications like Oracle or SAP can leverage R5 instances and R6 instances.

Q: When should I use X2idn and X2iedn instances?

X2idn and X2iedn instances are powered by 3rd generation Intel Xeon Scalable processors with an all-core turbo frequency up to 3.5 GHz and deliver up to 50% higher compute price performance than comparable X1 instances. X2idn and X2iedn instances both include up to 3.8 TB of local NVMe SSD storage and up to 100 Gbps of networking bandwidth, while X2idn offers up to 2 TiB of memory and X2iedn offers up to 4 TiB of memory. X2idn and X2iedn instances are SAP-Certified and are a great fit for workloads such as small-to large-scale traditional and in-memory databases, and analytics.

Q: When should I use X2iezn instances?

X2iezn instances feature the fastest Intel Xeon Scalable processors in the cloud and are a great fit for workloads that need high single-threaded performance combined with a high memory-to-vCPU ratio and high speed networking. X2iezn instances have an all-core turbo frequency up to 4.5 GHz, feature a 32:1 ratio of memory to vCPU, and deliver up to 55% higher compute price performance compared to X1e instances. X2iezn instances are a great fit for electronic design automation (EDA) workloads like physical verification, static timing analysis, power signoff, and full chip gate-level simulation.

Q: Which operating systems/AMIs are supported on X2gd instances?

The following AMIs are supported: Amazon Linux 2, Ubuntu 18.04 or newer, Red Hat Enterprise Linux 8.2 or newer, and SUSE Enterprise Server 15 or newer. Customers will find additional AMIs such as Fedora, Debian, NetBSD, and CentOS available through community AMIs and the AWS Marketplace. For containerized applications, Amazon ECS and EKS optimized AMIs are available as well.

Q: When should I use X1 instances?

X1 instances are ideal for running in-memory databases like SAP HANA, big data processing engines like Apache Spark or Presto, and high performance computing (HPC) applications. X1 instances are certified by SAP to run production environments of the next-generation Business Suite S/4HANA, Business Suite on HANA (SoH), Business Warehouse on HANA (BW), and Data Mart Solutions on HANA on the AWS cloud.

Q: Do X1 and X1e instances enable CPU power management state control?

Yes. You can configure C-states and P-states on x1e.32xlarge, x1e.16xlarge, x1e.8xlarge, x1.32xlarge and x1.16xlarge instances. You can use C-states to enable higher turbo frequencies (as much as 3.1 GHz with one or two core turbo). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

x1e.32xlarge will also support Windows Server 2012 R2 and 2012 RTM. x1e.xlarge, x1e.2xlarge, x1e.4xlarge, x1e.8xlarge, x1e.16xlarge and x1.32xlarge will also support Windows Server 2012 R2, 2012 RTM and 2008 R2 64bit (Windows Server 2008 SP2 and older versions will not be supported) and x1.16xlarge will support Windows Server 2012 R2, 2012 RTM, 2008 R2 64bit, 2008 SP2 64bit, and 2003 R2 64bit (Windows Server 32bit versions will not be supported).

Q: Are there standard SAP HANA reference deployment frameworks available for the High Memory instance and the AWS?

You can use AWS Launch Wizard for SAP or AWS Quick Start reference SAP HANA deployments to rapidly deploy all the necessary SAP HANA building blocks on High Memory instances following recommendations from AWS and SAP for high performance and reliability.

Previous Generation instances

Q: Why don't I see M1, C1, CC2 and HS1 instances on the pricing pages any more?

These have been moved to the Previous Generation Instance page.

Q: Are these Previous Generation instances still being supported?

Yes. Previous Generation instances are still fully supported.

Q: Can I still use/add more Previous Generation instances?

Yes. Previous Generation instances are still available as On-Demand, Reserved Instances, and Spot Instance, from our APIs, CLI and EC2 Management Console interface.

Q: Are my Previous Generation instances going to be deleted?

No. Until instances reach end of life and are fully deprecated, previous generation instances will be fully functional and will not be deleted because of this change. If AWS decides to deprecate previous generation instances due to end of life considerations, you will be notified of that change.

Q: Are Previous Generation instances being discontinued soon?

With any rapidly evolving technology the latest generation will typically provide the best performance for the price and we encourage our customers to take advantage of technological advancements. If AWS decides to deprecate previous generation instances due to end of life considerations, you will be notified of that change.

Q: Will my Previous Generation instances I purchased as a Reserved Instance be affected or changed?

No. Your Reserved Instances will not change, and the Previous Generation instances are not going away.

Storage Optimized instances

Q: What is a Dense-storage Instance?

Dense-storage instances are designed for workloads that require high sequential read and write access to very large data sets, such as Hadoop distributed computing, massively parallel processing data warehousing, and log processing applications. The Dense-storage instances offer the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q: How do dense-storage instances compare to High I/O instances?

High I/O instances (Im4gn, Is4gen, I4i, I3, I3en) are targeted at workloads that demand low latency and high random I/O in addition to moderate storage density and provide the best price/IOPS across other EC2 instance types. Dense-storage instances (D3, D3en, D2) and HDD-storage instances (H1) are optimized for applications that require high sequential read/write access and low cost storage for very large data sets and provide the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q: How much disk throughput can Dense-storage and HDD-storage instances deliver?

The largest current generation of Dense HDD-storage instances, d3en.12xlarge, can deliver up to 6.2 GiB/s read and 6.2 GiB/s write disk throughput with a 128k block size. Please see the product detail page for additional performance information. To ensure the best disk throughput performance from your D2, D3 and D3en instances on Linux, we recommend that you use the most recent version of the Amazon Linux AMI, or another Linux AMI with a kernel version of 3.8 or later that supports persistent grants—an extension to the Xen block ring protocol that significantly improves disk throughput and scalability.

Q: Do Dense-storage and HDD-storage instances provide any failover mechanisms or redundancy?

D2 and H1 instances provide notifications for hardware failures. Like all instance storage, Dense HDD-storage volumes persist only for the life of the instance. Hence, we recommend that you build a degree of redundancy (e.g. RAID 1/5/6) or use file systems (e.g. HDFS and MapR-FS) that support redundancy and fault tolerance. You can also back up data periodically to more data storage solutions such as Amazon EBS or Amazon S3.

Q: How do dense HDD-storage instances differ from Amazon EBS?

Amazon EBS offers simple, elastic, reliable (replicated), and persistent block level storage for Amazon EC2 while abstracting the details of the underlying storage media in use. Amazon EC2 instance instances with local HDD or NVMe storage provide directly attached, high performance storage building blocks that can be used for a variety of storage applications. Dense-storage instances are specifically targeted at customers who want high sequential read/write access to large data sets on local storage, e.g. for Hadoop distributed computing and massively parallel processing data warehousing.

Q: Can I launch dense HDD-storage instances as Amazon EBS optimized instances?

Each HDD-storage instance type (H1, D2, D3, and D3en) is EBS optimized by default. Since this feature is always enabled, launching one of these instances explicitly as EBS optimized will not affect the instance's behavior. For more information, see Amazon EBS–optimized instances.

Q: Can I launch D2 instances as Amazon EBS optimized instances?

Each D2 instance type is EBS optimized by default. D2 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on D2 instances, launching a D2 instance explicitly as EBS optimized will not affect the instance's behavior.

Q: What is a High I/O instance?

High I/O instances use NVMe based local instance storage to deliver very high, low latency, I/O capacity to applications, and are optimized for applications that require millions of IOPS. Like Cluster instances, High I/O instances can be clustered via cluster placement groups for low latency networking.

Q: Are all features of Amazon EC2 available for High I/O instances?

High I/O instances support all Amazon EC2 features. Im4gn, Is4gen, I4i, I3 and I3en instances offer NVMe only storage, while previous generation I2 instances allow legacy blkfront storage access.

Q: AWS has other database and Big Data offerings. When or why should I use High I/O instances?

High I/O instances are ideal for applications that require access to millions of low latency IOPS, and can leverage data stores and architectures that manage data redundancy and availability. Example applications are:

- NoSQL databases like Cassandra and MongoDB
- In-memory databases like Aerospike
- Elasticsearch and analytics workloads
- OLTP systems

Q: Do High I/O instances provide any failover mechanisms or redundancy?

Like other Amazon EC2 instance types, instance storage on Im4gn, Is4gen, I4i, I3 and I3en instances persists during the life of the instance. Customers are expected to build resilience into their applications. We recommend using databases and file systems that support redundancy and fault tolerance. Customers should back up data periodically to Amazon S3 for improved data durability.

Q: Do High I/O instances support TRIM?

The TRIM command allows the operating system to inform SSDs which blocks of data are no longer considered in use and can be wiped internally. In the absence of TRIM, future write operations to the involved blocks can slow down significantly. Im4gn, Is4gen, I4i, I3 and I3en instances support TRIM.

Q: How do D3 and D3en instances compare to D2 instances?

D3 and D3en instances offer improved specifications over D2 on the following compute, storage and network attributes:

- D3 and D3en instances offer up to 30% higher compute performances than equivalent D2 instances. Exact performance benefit will depend on the specific workload.
- D3 and D3en instances provide up to 45% and 100% higher disk throughput than D2 instances, respectively.
- D3 instances are available at a price that is 5% lower than D2 instances. D3en instances lower cost per TB of storage by up to 80% compared to D2 instances.
- D3 and D3en instances offer Intel Advanced Vector Extensions (AVX 512), which offer up to 2X the FLOPS per cycle compared to AVX 2 on D2.
- D3en instances offer a new instance size (12xl) with 48 vCPUs and 7 TB of storage per vCPU for 336 TB of total storage, but have half the memory per vCPU compared to D2 and 48 TB of total storage.
- D3 and D3en instances offer up to 25 Gbps and 75 Gbps of network bandwidth respectively on their largest sizes to meet customer needs for network performance for running big data workloads and file system clusters.

Q: Do D3 and D3en instances encrypt storage volumes and network traffic?

Yes; data written onto the storage volumes will be encrypted at rest using AES-256-XTS. Network traffic between D3 and D3en instances in the same VPC or a peered VPC are encrypted by default using a 256-bit key.

Storage

Amazon Elastic Block Store (EBS) | Amazon Elastic File System (EFS) | NVMe Instance storage

Amazon Elastic Block Store (Amazon EBS)

Q: What happens to my data when a system terminates?

The data stored on a local instance store will persist only as long as that instance is alive. However, data that is stored on an Amazon EBS volume will persist independently of the life of the instance. Therefore, we recommend that you use the local instance store for temporary data and, for data requiring a higher level of durability, we recommend using Amazon EBS volumes or backing up the data to Amazon S3. If you are using an Amazon EBS volume as a root partition, you will need to set the Delete On Terminate flag to "N" if you want your Amazon EBS volume to persist outside the life of the instance.

Q: What kind of performance can I expect from Amazon EBS volumes?

Amazon EBS provides four current generation volume types that are divided into two major categories: SSD-backed storage for transactional workloads and HDD-backed storage for throughput intensive workloads. These volume types differ in performance characteristics and price, allowing you to tailor your storage performance and cost to the needs of your applications. For more information, see the Amazon EBS overview. For additional information on performance, see the Amazon EC2 User Guide's EBS Performance section.

Q: What are Throughput Optimized HDD (st1) and Cold HDD (sc1) volume types?

ST1 volumes are backed by hard disk drives (HDDs) and are ideal for frequently accessed, throughput intensive workloads with large datasets and large I/O sizes, such as MapReduce, Kafka, log processing, data warehouse, and ETL workloads. These volumes deliver performance in terms of throughput, measured in MB/s, and include the ability to burst up to 250 MB/s per TB, with a baseline throughput of 40 MB/s per TB and a maximum throughput of 500 MB/s per volume. ST1 is designed to deliver the expected throughput performance 99% of the time and has enough I/O credits to support a full-volume scan at the burst rate.

SC1 volumes are backed by HDDs and provide the lowest cost per GB of all EBS volume types. It is ideal for less frequently accessed workloads with large, cold datasets. Similar to st1, sc1 provides a burst model: these volumes can burst up to 80 MB/s per TB, with a baseline throughput of 12 MB/s per TB and a maximum throughput of 250 MB/s per volume. For infrequently accessed data, sc1 provides extremely inexpensive storage. SC1 is designed to deliver the expected throughput performance 99% of the time and has enough I/O credits to support a full-volume scan at the burst rate.

To maximize the performance of st1 and sc1, we recommend using EBS-optimized EC2 instances.

Q: Which volume type should I choose?

Amazon EBS includes two major categories of storage: SSD-backed storage for transactional workloads (performance depends primarily on IOPS) and HDD-backed storage for throughput workloads (performance depends primarily on throughput, measured in MB/s). SSD-backed volumes are designed for transactional, IOPS-intensive database workloads, boot volumes, and workloads that require high IOPS. SSD-backed volumes include Provisioned IOPS SSD (io1 and io2) and General Purpose SSD (gp2 and gp3). HDD-backed volumes are designed for throughput-intensive and big-data workloads, large I/O sizes, and sequential I/O patterns. HDD-backed volumes include Throughput Optimized HDD (st1) and Cold HDD (sc1). For more information, see the Amazon EBS overview.

Q: Do you support multiple instances accessing a single volume?

Yes, you can enable Multi-Attach on an EBS Provisioned IOPS io1 volume to allow a volume to be concurrently attached to up to sixteen Nitro-based EC2 instances within the same Availability Zone. For more information on Amazon EBS Multi-Attach, see the EBS product page.

Q: Will I be able to access my EBS snapshots using the regular Amazon S3 APIs?

No, EBS snapshots are only available through the Amazon EC2 APIs.

Q: Do volumes need to be un-mounted in order to take a snapshot? Does the snapshot need to complete before the volume can be used again?

No, snapshots can be done in real time while the volume is attached and in use. However, snapshots only capture data that has been written to your Amazon EBS volume, which might exclude any data that has been locally cached by your application or OS. In order to ensure consistent snapshots on volumes attached to an instance, we recommend cleanly detaching the volume, issuing the snapshot command, and then reattaching the volume. For Amazon EBS volumes that serve as root devices, we recommend shutting down the machine to take a clean snapshot.

Q: Are snapshots versioned? Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier, and customers can create volumes based on any of their existing snapshots.

Q: What charges apply when using Amazon EBS shared snapshots?

If you share a snapshot, you won't be charged when other users make a copy of your snapshot. If you make a copy of another user's shared volume, you will be charged normal EBS rates.

Q: Can users of my Amazon EBS shared snapshots change any of my data?

Users who have permission to create volumes based on your shared snapshots will first make a copy of the snapshot into their account. Users can modify their own copies of the data, but the data on your original snapshot and any other volumes created by other users from your original snapshot will remain unmodified.

Q: How can I discover Amazon EBS snapshots that have been shared with me?

You can find snapshots that have been shared with you by selecting "Private Snapshots" from the viewing dropdown in the Snapshots section of the AWS Management Console. This section will list both snapshots you own and snapshots that have been shared with you.

Q: How can I find what Amazon EBS snapshots are shared globally?

You can find snapshots that have been shared globally by selecting "Public Snapshots" from the viewing dropdown in the Snapshots section of the AWS Management Console.

Q: Do you offer encryption on Amazon EBS volumes and snapshots?

Yes. EBS offers seamless encryption of data volumes and snapshots. EBS encryption better enables you to meet security and encryption compliance requirements.

Q: How can I find a list of Amazon Public Data Sets?

All information on Public Data Sets is available in our Public Data Sets Resource Center. You can also obtain a listing of Public Data Sets within the AWS Management Console by choosing “Amazon Snapshots” from the viewing dropdown in the Snapshots section.

Q: Where can I learn more about EBS?

See Amazon EBS FAQ.

Amazon Elastic File System (Amazon EFS)

Q: How do I access a file system from an Amazon EC2 instance?

To access your file system, you mount the file system on an Amazon EC2 Linux-based instance using the standard Linux mount command and the file system’s DNS name. Once you’ve mounted, you can work with the files and directories in your file system just like you would with a local file system.

Amazon EFS uses the NFSv4.1 protocol. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the Amazon EFS Getting Started guide.

Q: What Amazon EC2 instance types and AMIs work with Amazon EFS?

Amazon EFS is compatible with all Amazon EC2 instance types and is accessible from Linux-based AMIs. You can mix and match the instance types connected to a single file system. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the Amazon EFS Getting Started guide.

Q: How do I load data into a file system?

You can load data into an Amazon EFS file system from your Amazon EC2 instances or from your on-premises datacenter servers.

Amazon EFS file systems can be mounted on an Amazon EC2 instance, so any data that is accessible to an Amazon EC2 instance can also be read and written to Amazon EFS. To load data that is not currently stored on the Amazon cloud, you can use the same methods you use to transfer files to Amazon EC2 today, such as Secure Copy (SCP).

Amazon EFS file systems can also be mounted on an on-premises server, so any data that is accessible to an on-premises server can be read and written to Amazon EFS using standard Linux tools. For more information about accessing a file system from an on-premises server, please see the On-premises Access section of the Amazon EFS FAQ.

For more information about moving data to the Amazon cloud, please see the Cloud Data Migration page.

Q: How do I access my file system from outside my VPC?

Amazon EC2 instances within your VPC can access your file system directly. On-premises servers can mount your file systems via an AWS Direct Connect connection to your VPC.

Q: How many Amazon EC2 instances can connect to a file system?

Amazon EFS supports one to thousands of Amazon EC2 instances connecting to a file system concurrently.

Q: Where can I learn more about EFS?

You can visit the Amazon EFS FAQ page.

NVMe Instance storage

Q: Is data stored on Amazon EC2 NVMe instance storage encrypted?

Yes, all data is encrypted in an AWS Nitro hardware module prior to being written on the locally attached SSDs offered via NVMe instance storage.

Q: What encryption algorithm is used to encrypt Amazon EC2 NVMe instance storage?

Amazon EC2 NVMe instance storage is encrypted using an XTS-AES-256 block cipher.

Q: Are encryption keys unique to an instance or a particular device for NVMe instance storage?

Encryption keys are securely generated within the Nitro hardware module, and are unique to each NVMe instance storage device that is provided with an EC2 instance.

Q: What is the lifetime of encryption keys on NVMe instance storage?

All keys are irrecoverably destroyed on any de-allocation of the storage, including instance stop and instance terminate actions.

Q: Can I disable NVMe instance storage encryption?

No, NVMe instance storage encryption is always on, and cannot be disabled.

Q: Do the published IOPS performance numbers on I3 and I3en include data encryption?

Yes, the documented IOPS numbers for Im4gn, Is4gen, I4i, I3 and I3en NVMe instance storage include encryption.

Q: Does Amazon EC2 NVMe instance storage support AWS Key Management Service (KMS)?

No, disk encryption on NVMe instance storage does not support integration with AWS KMS system. Customers cannot bring in their own keys to use with NVMe instance storage.

Networking and security

Elastic Network Adapter (ENA) Express | Elastic Fabric Adapter (EFA) | Enhanced networking | Elastic Load Balancing | Elastic IP | Security

Elastic Network Adapter (ENA) Express

Q: What is ENA Express?

ENA Express is an enhancement on the Elastic Network Adapter that brings the Scalable Reliable Datagram (SRD) protocol to traditional TCP and UDP networking. Transparent to the application, ENA Express improves single flow bandwidths and reduces tail latencies in throughput intensive workloads.

Q: How does ENA Express work?

When configured, ENA Express works between any two supported instances in an Availability Zone (AZ). ENA Express detects compatibility between your EC2 instances and establishes an SRD connection when both communicating instances have ENA Express enabled. Once a connection is established, your traffic can take advantage of SRD and its performance benefits.

Q: When should I use ENA Express?

ENA Express works best for applications requiring high, single-flow throughput, like distributed storage systems and live media encoding. These workloads require high single flow bandwidth and low tail latency.

Q: How do I enable ENA Express?

ENA Express can be enabled on a per-ENI basis. While attaching a network card to an instance or while running a modify command, ENA Express can be enabled. ENA Express must be enabled on both communicating ENIs to establish point-to-point communication with it. Additionally, if you are using Jumbo Frames, you must adjust your maximum MTU to 8900 to use ENA Express.

Q: What protocols are supported by ENA Express?

ENA Express supports TCP by default. UDP can optionally be enabled through an API argument or within the management console.

Q: What instances are supported?

ENA Express is supported on Graviton-, Intel-, and AMD-based EC2 instances. It is supported on compute-optimized, memory-optimized, general purpose and storage-optimized based instances. For a complete list of supported instances, please see the ENA Express user guide.

Q: What is the difference between Elastic Fabric Adapter (EFA) and ENA Express?

EFA is a network interface built for HPC and ML applications, and it also leverages the SRD protocol. EFA requires a different network programming model, which uses the LibFabric interface to pass communication to the ENI. Unlike EFA, ENA Express helps you run your application transparently on TCP and UDP. Additionally, ENA Express allows for intra-Availability Zone (AZ) communication, while EFA is currently limited to communication within the same subnet.

Q: What happens if I'm running ENA Express on one instance, and it is communicating with another instance that doesn't support ENA Express or hasn't enabled it on the ENI?

ENA Express will detect if ENA Express has been enabled on another instance. If that instance doesn't support or hasn't enabled ENA Express, your instance will fallback to normal ENA operation. You will not be able to achieve any of the SRD performance benefits in this case, but there are no adverse effects either.

Q: What operating systems are supported?

The SRD functionality will be supported on all operating systems, but please note that ENA Express monitoring metrics will be available on only the EthTool in the latest Amazon Linux AMI or by installing the ENA driver version 2.8.0 or later from GitHub, with all operating systems supporting the metrics in the future.

Q: What monitoring tools are available to track this?

ENA Express offers EthTool counters to track packets that are eligible for SRD transmission in addition to those actually sent and received with SRD. Additionally, EthTool will support an SRD resource utilization metric on a percent basis, providing insight into when you should consider scaling out your architecture. Finally, a Boolean will provide on and off status for ENA Express and the UDP protocol.

Q: Where is ENA Express available?

ENA Express is available in all commercial Regions. It can be used to establish communication between any two enabled instances within the same AZ.

Q: Are there any additional costs to running ENA Express?

No, ENA Express is free to use.

Elastic Fabric Adapter (EFA)

Q: Why should I use EFA?

EFA brings the scalability, flexibility, and elasticity of cloud to tightly coupled HPC applications. With EFA, tightly coupled HPC applications have access to lower and more consistent latency and higher throughput than traditional TCP channels, enabling them to scale better. EFA support can be enabled dynamically, on-demand on any supported EC2 instance without pre-reservation, giving you the flexibility to respond to changing business/workload priorities.

Q: What types of applications can benefit from using EFA?

HPC applications distribute computational workloads across a cluster of instances for parallel processing. Examples of HPC applications include computational fluid dynamics (CFD), crash simulations, and weather simulations. HPC applications are generally written using the

Message Passing Interface (MPI) and impose stringent requirements for inter-instance communication in terms of both latency and bandwidth. Applications using MPI and other HPC middleware that supports the libfabric communication stack can benefit from EFA.

Q: How does EFA communication work?

EFA devices provide all ENA devices' functionalities plus a new OS bypass hardware interface that allows user-space applications to communicate directly with the hardware-provided reliable transport functionality. Most applications will use existing middleware, such as the MPI, to interface with EFA. AWS has worked with a number of middleware providers to ensure support for the OS bypass functionality of EFA. Please note that communication using the OS bypass functionality is limited to instances within a single subnet of a virtual private cloud (VPC).

Q: Which instance types support EFA?

For a full list of supported EC2 instances, refer to this page in our documentation.

Q: What are the differences between an EFA ENI and an ENA ENI?

An ENA ENI provides traditional IP networking features necessary to support VPC networking. An EFA ENI provides all the functionality of an ENA ENI, plus hardware support for applications to communicate directly with the EFA ENI without involving the instance kernel (OS-bypass communication) using an extended programming interface. Due to the advanced capabilities of the EFA ENI, EFA ENIs can only be attached at launch or to stopped instances.

Q: What are the pre-requisites to enabling EFA on an instance?

EFA support can be enabled either at the launch of the instance or added to a stopped instance. EFA devices cannot be attached to a running instance.

Enhanced networking

Q: What networking capabilities are included in this feature?

We currently support enhanced networking capabilities using SR-IOV (Single Root I/O Virtualization). SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization compared to traditional implementations. For supported Amazon EC2 instances, this feature provides higher packet per second (PPS) performance, lower inter-instance latencies, and very low network jitter.

Q: Why should I use Enhanced Networking?

If your applications benefit from high packet-per-second performance and/or low latency networking, Enhanced Networking will provide significantly improved performance, consistence of performance and scalability.

Q: How can I enable Enhanced Networking on supported instances?

In order to enable this feature, you must launch an HVM AMI with the appropriate drivers. The instances listed as current generation use ENA for enhanced networking. Amazon Linux AMI includes both of these drivers by default. For AMIs that do not contain these drivers, you will need to download and install the appropriate drivers based on the instance types you plan to use. You can use Linux or Windows instructions to enable Enhanced Networking in AMIs that do not include the SR-IOV driver by default. Enhanced Networking is only supported in Amazon VPC.

Q: Do I need to pay an additional fee to use Enhanced Networking?

No, there is no additional fee for Enhanced Networking. To take advantage of Enhanced Networking you need to launch the appropriate AMI on a supported instance type in a VPC.

Q: Which instance types support Enhanced Networking?

Depending on your instance type, you can enable enhanced networking by using one of the following mechanisms:

Intel 82599 Virtual Function (VF) interface - The Intel 82599 Virtual Function interface supports network speeds of up to 10 Gbps for supported instance types. C3, C4, D2, I2, M4 (excluding m4.16xlarge), and R3 instances use the Intel 82599 VF interface for enhanced networking.

Elastic Network Adapter (ENA) - The Elastic Network Adapter (ENA) supports network speeds of up to 200 Gbps for supported instance types. The instances listed as current generation use ENA for enhanced networking, with the exception of C4, D2, and M4 instances smaller than m4.16xlarge.

Q: What does it mean to have multiple network cards for an EC2 instance? Why are they needed?

Newer generation EC2 instances use Nitro network cards for VPC data-plane offloading. To provide higher network bandwidth and improved packet-rate performance, you can configure specific EC2 instances to use multiple network cards for packet processing, ultimately increasing overall system performance.

Q: Which instance types support multiple network cards?

Multiple network cards are supported on accelerated instances such as p4d.24xlarge, and network optimized instances, such as c6in.32xlarge. For a full list of instances supporting multiple network cards, see the Elastic network interfaces.

Q: What is the default number of network interfaces a multiple card instance can launch with?

This depends on the instance type. Accelerated instances, such as p4 scale, up to 15 network interfaces per network card. High network instances such as the recently launched c6in instances support an aggregate 14 network interfaces split evenly (7 and 7) across the two network cards. For information about the network interface scale per network cards, see Network cards.

Elastic Load Balancing

Q: What load balancing options does the Elastic Load Balancing service offer?

Elastic Load Balancing offers two types of load balancers that both feature high availability, automatic scaling, and robust security. These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request.

Q: When should I use the Classic Load Balancer and when should I use the Application Load Balancer?

The Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while the Application Load Balancer is ideal for applications needing advanced routing capabilities, microservices, and container-based architectures. Please visit Elastic Load Balancing for more information.

Elastic IP

Q: Why am I limited to 5 Elastic IP addresses per region?

Public (IPV4) internet addresses are a scarce resource. There is only a limited amount of public IP space available, and Amazon EC2 is committed to helping use that space efficiently.

By default, all accounts are limited to 5 Elastic IP addresses per region. If you need more than 5 Elastic IP addresses, we ask that you apply for your limit to be raised. We will ask you to think through your use case and help us understand your need for additional addresses. You can apply for more Elastic IP addresses here. Any increases will be specific to the region they have been requested for.

Q: Why am I charged when my Elastic IP address is not associated with a running instance?

In order to help ensure our customers are efficiently using the Elastic IP addresses, we impose a small hourly charge for each address when it is not associated with a running instance.

Q: Do I need one Elastic IP address for every instance that I have running?

No. You do not need an Elastic IP address for all your instances. By default, every instance comes with a private IP address and an internet routable public IP address. The private IP address remains associated with the network interface when the instance is stopped and restarted, and is released when the instance is terminated. The public address is associated exclusively with the instance until it is stopped, terminated or replaced with an Elastic IP address. These IP addresses should be adequate for many applications where you do not need a long lived internet routable end point. Compute clusters, web crawling, and backend services are all examples of applications that typically do not require Elastic IP addresses.

Q: How long does it take to remap an Elastic IP address?

The remap process currently takes several minutes from when you instruct us to remap the Elastic IP until it fully propagates through our system.

Q: Can I configure the reverse DNS record for my Elastic IP address?

All Elastic IP addresses come with reverse DNS, in a standard template of the form ec2-1-2-3-4.region.compute.amazonaws.com. For customers requiring custom reverse DNS settings for internet-facing applications that use IP-based mutual authentication (such as sending email from EC2 instances), you can configure the reverse DNS record of your Elastic IP address by filling out this form. Alternatively, please contact AWS Customer Support if you want AWS to delegate the management of the reverse DNS for your Elastic IPs to your authoritative DNS name servers (such as Amazon Route 53), so that you can manage your own reverse DNS PTR records to support these use-cases. Note that a corresponding forward DNS record pointing to that Elastic IP address must exist before we can create the reverse DNS record.

Security

Q: How do I prevent other people from viewing my systems?

You have complete control over the visibility of your systems. The Amazon EC2 security systems allow you to place your running instances into arbitrary groups of your choice. Using the web services interface, you can then specify which groups may communicate with which other groups, and also which IP subnets on the Internet may talk to which groups. This allows you to control access to your instances in our highly dynamic environment. Of course, you should also secure your instance as you would any other server.

Q: Can I get a history of all EC2 API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of all EC2 API calls (including VPC and EBS) made on your account, you simply turn on CloudTrail in the AWS Management Console. For more information, visit the CloudTrail home page.

Q: Where can I find more information about security on AWS?

For more information on security on AWS please refer to our Amazon Web Services: Overview of Security Processes white paper and to our Amazon EC2 running Windows Security Guide.

Management

Amazon CloudWatch | Amazon EC2 Auto Scaling | Hibernate | VM Import/Export

Amazon CloudWatch

Q: What is the minimum time interval granularity for the data that Amazon CloudWatch receives and aggregates?

Metrics are received and aggregated at 1 minute intervals.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

You can retrieve metrics data for any Amazon EC2 instance up to 2 weeks from the time you started to monitor it. After 2 weeks, metrics data for an Amazon EC2 instance will not be available if monitoring was disabled for that Amazon EC2 instance. If you want to archive metrics beyond 2 weeks you can do so by calling mon-get-stats command from the command line and storing the results in Amazon S3 or Amazon SimpleDB.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?

Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 2 weeks.

Q: Does the Amazon CloudWatch monitoring charge change depending on which type of Amazon EC2 instance I monitor?

No, the Amazon CloudWatch monitoring charge does not vary by Amazon EC2 instance type.

Q: Why does the graphing of the same time window look different when I view in 5 minute and 1 minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period, the single data point is placed at the 1 minute mark. We recommend using a 1 minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Amazon EC2 Auto Scaling

Q: Can I automatically scale Amazon EC2 Auto Scaling Groups?

Yes. Amazon EC2 Auto Scaling is a fully managed service designed to launch or terminate Amazon EC2 instances automatically to help ensure you have the correct number of Amazon EC2 instances available to handle the load for your application. EC2 Auto Scaling helps you maintain application availability through fleet management for EC2 instances, which detects and replaces unhealthy instances, and by scaling your Amazon EC2 capacity up or down automatically according to conditions you define. You can use EC2 Auto Scaling to automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs.

Allocation strategies in EC2 Auto Scaling determine how Spot Instances in your fleet are fulfilled from Spot Instance pools. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing. The lowest-price allocation strategy launches Spot Instances strictly based on diversification across 'N' lowest priced pools.

For more information see the Amazon EC2 Auto Scaling FAQ.

Hibernate

Q: Why should I hibernate an instance?

You can hibernate an instance to get your instance and applications up and running quickly, if they take a long time to bootstrap (e.g. load memory caches). You can start instances, bring them to a desired state and hibernate them. These "pre-warmed" instances can then be resumed to reduce the time it takes for an instance to return to service. Hibernation retains memory state across Stop/Start cycles.

Q: What happens when I hibernate my instance?

When you hibernate an instance, data from your EBS root volume and any attached EBS data volumes is persisted. Additionally, contents from the instance's memory (RAM) are persisted to EBS root volume. When the instance is restarted, it returns to its previous state and

reloads the RAM contents.

Q: What is the difference between hibernate and stop?

In the case of hibernate, your instance gets hibernated and the RAM data persisted. In the case of Stop, your instance gets shut down and RAM is cleared.

In both the cases, data from your EBS root volume and any attached EBS data volumes is persisted. Your private IP address remains the same, as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow. Stop and hibernate are available for Amazon EBS backed instances only. Local instance storage is not persisted.

Q: How much does it cost to hibernate an instance?

Hibernating instances are charged at standard EBS rates for storage. As with a stopped instance, you do not incur instance usage fees while an instance is hibernating.

Q: How can I hibernate an instance?

Hibernation needs to be enabled when you launch the instance. Once enabled, you can use the StopInstances API with an additional 'Hibernate' parameter to trigger hibernation. You can also do this through the console by selecting your instance, then clicking Actions> Instance State > Stop - Hibernate. For more information on using hibernation, refer to the user guide.

Q: How can I resume a hibernating instance?

You can resume by calling the StartInstances API as you would for a regular stopped instance. You can also do this through the console by selecting your instance, then clicking Actions > Instance State > Start.

Q: Can I enable hibernation on an existing instance?

No, you cannot enable hibernation on an existing instance (running or stopped). This needs to be enabled during instance launch.

Q: How can I tell that an instance is hibernated?

You can tell that an instance is hibernated by looking at the state reason. It should be 'Client.UserInitiatedHibernate'. This is visible on the console under "Instances - Details" view or in the DescribeInstances API response as the "reason" field.

Q: What is the state of an instance when it is hibernating?

Hibernated instances are in 'Stopped' state.

Q: What data is saved when I hibernate an instance?

EBS volume storage (boot volume and attached data volumes) and memory (RAM) are saved. Your private IP address remains the same (for VPC), as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow.

Q: Where is my data stored when I hibernate an instance?

As with the Stop feature, root device and attached device data are stored on the corresponding EBS volumes. Memory (RAM) contents are stored on the EBS root volume.

Q: Is my memory (RAM) data encrypted when it is moved to EBS?

Yes, RAM data is always encrypted when it is moved to the EBS root volume. Encryption on the EBS root volume is enforced at instance launch time. This is to ensure protection for any sensitive content that is in memory at the time of hibernation.

Q: How long can I keep my instance hibernated?

We do not support keeping an instance hibernated for more than 60 days. You need to resume the instance and go through Stop and Start (without hibernation) if you wish to keep the instance around for a longer duration. We are constantly working to keep our platform up-to-date with upgrades and security patches, some of which can conflict with the old hibernated instances. We will notify you for critical updates that require you to resume the hibernated instance to perform a shutdown or a reboot.

Q: What are the prerequisites to hibernate an instance?

To use hibernation, the root volume must be an encrypted EBS volume. The instance needs to be configured to receive the ACPID signal for hibernation (or use the Amazon published AMIs that are configured for hibernation). Additionally, your instance should have sufficient space available on your EBS root volume to write data from memory.

Q: Which instances and operating systems support hibernation?

For instances running Amazon Linux, Amazon Linux 2, Ubuntu, and Windows, Hibernation is supported across C3, C4, C5, C5d, I3, M3, M4, M5, M5a, M5ad, M5d, R3, R4, R5, R5a, R5ad, R5d, T2, T3, and T3a instances.

For instances running CentOS, Fedora, and Red Hat Enterprise Linux, Hibernation is supported across C5, C5d, M5, M5a, M5ad, M5d, R5, R5a, R5ad, R5d, T3, and T3a instances.

For Windows, Hibernation is supported for instances up to 16 GB of RAM. For other operating systems, Hibernation is supported for instances with less than 150 GB of RAM. To review the list of supported OS versions and instance types, refer to the user guide.

Q: Should I use specific Amazon Machine Image (AMIs) if I want to hibernate my instance?

You can use any AMI that is configured to support hibernation. You can use AWS published AMIs that support hibernation by default. Alternatively, you can create a custom image from an instance after following the hibernation pre-requisite checklist and configuring your instance appropriately.

Q: What if my EBS root volume is not large enough to store memory state (RAM) for hibernation?

To enable hibernation, space is allocated on the root volume to store the instance memory (RAM). Make sure that the root volume is large enough to store the RAM contents and accommodate your expected usage, e.g. OS, applications. If the EBS root volume does not have enough space, hibernation will fail and the instance will get shut down instead.

VM Import/Export

Q: What is VM Import/Export?

VM Import/Export enables customers to import Virtual Machine (VM) images in order to create Amazon EC2 instances. Customers can also export previously imported EC2 instances to create VMs. Customers can use VM Import/Export to leverage their previous investments in building VMs by migrating their VMs to Amazon EC2.

Q: What operating systems are supported?

VM Import/Export currently supports Windows and Linux VMs, including multiple editions of Windows Server, Red Hat Enterprise Linux (RHEL), CentOS, Ubuntu, Debian and others. For more details on VM Import, including supported file formats, architectures, and operating system configurations, please see the VM Import/Export section of the VM Import/Export.

Q: What VM file formats are supported?

You can import VMware ESX VMDK images, Citrix Xen VHD images, Microsoft Hyper-V VHD images and RAW images as Amazon EC2 instances. You can export EC2 instances to VMware ESX VMDK, VMware ESX OVA, Microsoft Hyper-V VHD or Citrix Xen VHD images. For a full list of supported operating systems, please see What operating systems are supported?

Q: What is VMDK?

VMDK is a file format that specifies a virtual machine hard disk encapsulated within a single file. It is typically used by virtual IT infrastructures such as those sold by VMware, Inc.

Q: How do I prepare a VMDK file for import using the VMware vSphere client?

The VMDK file can be prepared by calling File-Export-Export to OVF template in VMware vSphere Client. The resulting VMDK file is compressed to reduce the image size and is compatible with VM Import/Export. No special preparation is required if you are using the

Amazon EC2 VM Import Connector vApp for VMware vCenter.

Q: What is VHD?

VHD (Virtual Hard Disk) is a file format that specifies a virtual machine hard disk encapsulated within a single file. The VHD image format is used by virtualization platforms such as Microsoft Hyper-V and Citrix Xen.

Q: How do I prepare a VHD file for import from Citrix Xen?

Open Citrix XenCenter and select the virtual machine you want to export. Under the Tools menu, choose "Virtual Appliance Tools" and select "Export Appliance" to initiate the export task. When the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q: How do I prepare a VHD file for import from Microsoft Hyper-V?

Open the Hyper-V Manager and select the virtual machine you want to export. In the Actions pane for the virtual machine, select "Export" to initiate the export task. Once the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q: Are there any other requirements when importing a VM into Amazon EC2?

The virtual machine must be in a stopped state before generating the VMDK or VHD image. The VM cannot be in a paused or suspended state. We suggest that you export the virtual machine with only the boot volume attached. You can import additional disks using the ImportVolume command and attach them to the virtual machine using AttachVolume. Additionally, encrypted disks (e.g. Bit Locker) and encrypted image files are not supported. You are also responsible for ensuring that you have all necessary rights and licenses to import into AWS and run any software included in your VM image.

Q: Does the virtual machine need to be configured in any particular manner to enable import to Amazon EC2?

Ensure Remote Desktop (RDP) or Secure Shell (SSH) is enabled for remote access and verify that your host firewall (Windows firewall, iptables, or similar), if configured, allows access to RDP or SSH. Otherwise, you will not be able to access your instance after the import is complete. Please also ensure that Windows VMs are configured to use strong passwords for all users including the administrator and that Linux VMs are configured with a public key for SSH access.

Q: How do I import a virtual machine to an Amazon EC2 instance?

You can import your VM images using the Amazon EC2 API tools:

- Import the VMDK, VHD or RAW file via the ec2-import-instance API. The import instance task captures the parameters necessary to properly configure the Amazon EC2 instance properties (instance size, Availability Zone, and security groups) and uploads the disk image into Amazon S3.
- If ec2-import-instance is interrupted or terminates without completing the upload, use ec2-resume-import to resume the upload. The import task will resume where it left off.
- Use the ec2-describe-conversion-tasks command to monitor the import progress and obtain the resulting Amazon EC2 instance ID.
- Once your import task is completed, you can boot the Amazon EC2 instance by specifying its instance ID to the ec2-run-instances API.
- Finally, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3 as it is no longer needed.

Alternatively, if you use the VMware vSphere virtualization platform, you can import your virtual machine to Amazon EC2 using a graphical user interface provided through AWS Management Portal for vCenter. Please refer to the Getting Started Guide in AWS Management Portal for vCenter. AWS Management Portal for vCenter includes integrated support for VM Import. Once the portal is installed within vCenter, you can right-click on a VM and select “Migrate to EC2” to create an EC2 instance from the VM. The portal will handle exporting the VM from vCenter, uploading it to S3, and converting it into an EC2 instance for you, with no additional work required. You can also track the progress of your VM migrations within the portal.

Q: How do I export an Amazon EC2 instance back to my on-premise virtualization environment?

You can export your Amazon EC2 instance using the Amazon EC2 CLI tools:

- Export the instance using the ec2-create-instance-export-task command. The export command captures the parameters necessary (instance ID, S3 bucket to hold the exported image, name of the exported image, VMDK, OVA or VHD format) to properly export the instance to your chosen format. The exported file is saved in an S3 bucket that you previously created.
- Use ec2-describe-export-tasks to monitor the export progress.
- Use ec2-cancel-export-task to cancel an export task prior to completion.

Q: Are there any other requirements when exporting an EC2 instance using VM Import/Export?

You can export running or stopped EC2 instances that you previously imported using VM Import/Export. If the instance is running, it will be momentarily stopped to snapshot the boot volume. EBS data volumes cannot be exported. EC2 instances with more than one network interface cannot be exported.

Q: Can I export Amazon EC2 instances that have one or more EBS data volumes attached?

Yes, but VM Import/Export will only export the boot volume of the EC2 instance.

Q: What does it cost to import a virtual machine?

You will be charged standard Amazon S3 data transfer and storage fees for uploading and storing your VM image file. Once your VM is imported, standard Amazon EC2 instance hour and EBS service fees apply. If you no longer wish to store your VM image file in S3 after the import process completes, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3.

Q: What does it cost to export a VM?

You will be charged standard Amazon S3 storage fees for storing your exported VM image file. You will also be charged standard S3 data transfer charges when you download the exported VM file to your on-premise virtualization environment. Finally, you will be charged standard EBS charges for storing a temporary snapshot of your EC2 instance. To minimize storage charges, delete the VM image file in S3 after downloading it to your virtualization environment.

Q: When I import a VM of Windows Server 2003 or 2008, who is responsible for supplying the operating system license?

When you launch an imported VM using Microsoft Windows Server 2003 or 2008, you will be charged standard instance hour rates for Amazon EC2 running the appropriate Windows Server version, which includes the right to utilize that operating system within Amazon EC2. You are responsible for ensuring that all other installed software is properly licensed.

So then, what happens to my on-premise Microsoft Windows license key when I import a VM of Windows Server 2003 or 2008? Since your on-premise Microsoft Windows license key that was associated with that VM is not used when running your imported VM as an EC2 instance, you can reuse it for another VM within your on-premise environment.

Q: Can I continue to use the AWS provided Microsoft Windows license key after exporting an EC2 instance back to my on-premises virtualization environment?

No. After an EC2 instance has been exported, the license key utilized in the EC2 instance is no longer available. You will need to reactivate and specify a new license key for the exported VM after it is launched in your on-premises virtualization platform.

Q: When I import a VM with Red Hat Enterprise Linux (RHEL), who is responsible for supplying the operating system license?

When you import Red Hat Enterprise Linux (RHEL) VM images, you can use license portability for your RHEL instances. With license portability, you are responsible for maintaining the RHEL licenses for imported instances, which you can do using Cloud Access subscriptions for Red Hat Enterprise Linux. Please contact Red Hat to learn more about Cloud Access and to verify your eligibility.

Q: How long does it take to import a virtual machine?

The length of time to import a virtual machine depends on the size of the disk image and your network connection speed. As an example, a 10 GB Windows Server 2008 SP2 VMDK image takes approximately 2 hours to import when it’s transferred over a 10 Mbps network connection. If you have a slower network connection or a large disk to upload, your import may take significantly longer.

Q: In which Amazon EC2 Regions can I use VM Import/Export?

Visit the Region Table page to see product service availability by Region.

Q: How many simultaneous import or export tasks can I have?

Each account can have up to five active import tasks and five export tasks per region.

Q: Can I run imported virtual machines in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, you can launch imported virtual machines within Amazon VPC.

Q: Can I use the AWS Management Console with VM Import/Export?

No. VM Import/Export commands are available via EC2 CLI and API. You can also use the AWS Management Portal for vCenter to import VMs into Amazon EC2. Once imported, the resulting instances are available for use via the AWS Management Console.

Billing and purchase options

Billing | Data transfer fees when moving all data off AWS | Convertible Reserved Instances | EC2 Fleet | Amazon EC2 Capacity Blocks for ML | On-Demand Capacity Reservation | Reserved Instances | Reserved Instance Marketplace | Savings Plans | Spot Instances

Billing

Q: How will I be charged and billed for my use of Amazon EC2?

You pay only for what you use. Displayed pricing is an hourly rate but depending on which instances you choose, you pay by the hour or second (minimum of 60 seconds) for each instance type. Partial instance-hours consumed are billed based on instance usage. Data transferred between AWS services in different regions is charged at standard inter-region data transfer rates. Usage for other Amazon Web Services is billed separately from Amazon EC2.

For EC2 pricing information, please visit the pricing section on the EC2 detail page.

Q: When does billing of my Amazon EC2 systems begin and end?

Billing commences when Amazon EC2 initiates the boot sequence of an AMI instance. Billing ends when the instance terminates, which could occur through a web services command, by running "shutdown -h", or through instance failure. When you stop an instance, we shut it down but don't charge hourly usage for a stopped instance, or data transfer fees, but we do charge for the storage for any Amazon EBS volumes. To learn more, visit the AWS Documentation.

Q: What defines billable EC2 instance usage?

Instance usages are billed for any time your instances are in a "running" state. If you no longer wish to be charged for your instance, you must "stop" or "terminate" the instance to avoid being billed for additional instance usage. Billing starts when an instance transitions into the running state.

Q: If I have two instances in different availability zones, how will I be charged for regional data transfer?

Each instance is charged for its data in and data out at corresponding Data Transfer rates. Therefore, if data is transferred between these two instances, it is charged at "Data Transfer Out from EC2 to Another AWS Region" for the first instance and at "Data Transfer In from Another AWS Region" for the second instance. Please refer to this page for detailed data transfer pricing.

Q: If I have two instances in different Regions, how will I be charged for data transfer?

Each instance is charged for its data in and data out at Inter-Region Data Transfer rates. Therefore, if data is transferred between these two instances, it is charged at Inter-Region Data Transfer Out for the first instance and at Inter-Region Data Transfer In for the second instance.

Q: How will my monthly bill show per-second versus per-hour?

Although EC2 charges in your monthly bill will now be calculated based on a per second basis, for consistency, the monthly EC2 bill will show cumulative usage for each instance that ran in a given month in decimal hours. For example, an instance running for 1 hour 10 minutes and 4

seconds would look like 1.1677. Read this blog for an example of the detailed billing report.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Data transfer fees when moving all data off AWS

Q: Will I incur any data transfer out to the internet charges when I move my data out of AWS?

AWS offers eligible customers free data transfer out to the internet when they move all of their data off of AWS, in accordance with the process below.

Q: I want to move my data out of AWS. How do I request free data transfer out to the internet?

Complete the following steps:

1) If you have a dedicated AWS account team, contact them first and inform them of your plans. In some cases, if you have a negotiated commitment with AWS, you'll want to discuss your options with your AWS account team.

2) Review the criteria and process described on this page.

3) Contact AWS Customer Support and indicate that your request is for “free data transfer to move off AWS.” AWS Customer Support will ask that you provide information, so they can review your moving plans, evaluate whether you qualify for free data transfer out, and calculate a proper credit amount.

4) If AWS Customer Support approves your move, you will receive a temporary credit for the cost of data transfer out based on the volume of all data you have stored across AWS services at the time of AWS' calculation. AWS Customer Support will notify you if you are approved, and you will then have 60 days to complete your move off of AWS. The credit will count against data transfer out usage only, and it will not be applied to other service usage. After your move away from AWS services, within the 60-day period, you must delete all remaining data and workloads from your AWS account, or you can close your AWS account.

Free data transfers for moving IT providers are also subject to the following criteria:

a) Only customers with an active AWS account in good standing are eligible for free data transfer out.

b) If you have less than 100 GB of data stored in your AWS account you may move this data off of AWS for free under AWS's existing 100 GB monthly free tier for data transfer out. Customers with less than 100 GB of data stored in their AWS account are not eligible for additional credits.

c) AWS will provide you with free data transfer out to the internet when you move all of your data off of AWS. If you only want to move your total usage of a single service, but not everything, contact AWS Customer Support.

d) If your plans change, or you cannot complete your move off of AWS within 60 days, you must notify AWS Customer Support.

e) Standard services charges for use of AWS services are not included. Only data transfer out charges in support of your move off of AWS are eligible for credits. However, data transfer out from specialized data transfer services, such as Amazon CloudFront, AWS Direct Connect, AWS Snow Family, and AWS Global Accelerator, are not included.

f) AWS may review your service usage to verify compliance with these requirements. If we determine your use of data transfer out was for a purpose other than moving off of AWS, we may charge you for the data transfer out that had been credited.

g) AWS may make changes with respect to free data transfers out to the internet at any time.

Q: Why do I have to request AWS' pre-approval for free data transfer out to the internet before moving my data out of AWS?

AWS customers make hundreds of millions of data transfers each day, and we generally don't know the reason for any given data transfer. For example, customers may be transferring data to an end user of their application, to a visitor of their website, or to another cloud or on-premises environment for backup purposes. Accordingly, the only way we know that your data transfer is to support your move off of AWS is if you tell us beforehand.

Convertible Reserved Instances

Q: What is a Convertible RI?

A Convertible RI is a type of Reserved Instance with attributes that can be changed during the term.

Q: When should I purchase a Convertible RI instead of a Standard RI?

The Convertible RI is useful for customers who can commit to using EC2 instances for a three-year term in exchange for a significant discount on their EC2 usage, are uncertain about their instance needs in the future, or want to benefit from changes in price.

Q: What term length options are available on Convertible RIs?

Like Standard RIs, Convertible RIs are available for purchase for a one-year or three-year term.

Q: Can I exchange my Convertible RI to benefit from a Convertible RI matching a different instance type, operating system, tenancy, or payment option?

Yes, you can select a new instance type, operating system, tenancy, or payment option when you exchange your Convertible RIs. You also have the flexibility to exchange a portion of your Convertible RI or merge the value of multiple Convertible RIs in a single exchange.

Q: Can I transfer a Convertible or Standard RI from one region to another?

No, an RI is associated with a specific region, which is fixed for the duration of the reservation's term.

Q: How do I change the configuration of a Convertible RI?

You can change the configuration of your Convertible RI using the EC2 Management Console or the GetReservedInstancesExchangeQuote API. You also have the flexibility to exchange a portion of your Convertible RI or merge the value of multiple Convertible RIs in a single exchange. Click here to learn more about exchanging Convertible RIs.

Q: Do I need to pay a fee when I exchange my Convertible RIs?

No, you do not pay a fee when you exchange your RIs. However you may need to pay a one-time true-up charge that accounts for differences in pricing between the Convertible RIs that you have and the Convertible RIs that you want.

Q: How do Convertible RI exchanges work?

When you exchange one Convertible RI for another, EC2 ensures that the total value of the Convertible RIs is maintained through a conversion. So, if you are converting your RI with a total value of $1000 for another RI, you will receive a quantity of Convertible RIs with a value that's equal to or greater than $1000. You cannot convert your Convertible RI for Convertible RI(s) of a lesser total value.

Q: Can you define total value?

The total value is the sum of all expected payments that you'd make during the term for the RI.

Q: Can you walk me through how the true-up cost is calculated for a conversion between two All Upfront Convertible RIs?

Sure, let's say you purchased an All Upfront Convertible RI for $1000 upfront, and halfway through the term you decide to change the attributes of the RI. Since you're halfway through the RI term, you have $500 left of prorated value remaining on the RI. The All Upfront Convertible RI that you want to convert into costs $1,200 upfront today. Since you only have half of the term left on your existing Convertible RI, there is $600 of value remaining on the desired new Convertible RI. The true-up charge that you'll pay will be the difference in upfront value between original and desired Convertible RIs, or $100 ($600 - $500).

Q: Can you walk me through a conversion between No Upfront Convertible RIs?

Unlike conversions between Convertible RIs with an upfront value, since you're converting between RIs without an upfront cost, there will not be a true-up charge. However, the amount you pay on an hourly basis before the exchange will need to be greater than or equal to the amount you pay on a total hourly basis after the exchange.

For example, let's say you purchased one No Upfront Convertible RI (A) with a $0.10/hr rate, and you decide to exchange Convertible RI A for another RI (B) that costs $0.06/hr. When you convert, you will receive two RIs of B because the amount that you pay on an hourly basis must be greater than or equal to the amount you're paying for A on an hourly basis.

Q: Can I customize the number of instances that I receive as a result of a Convertible RI exchange?

No, EC2 uses the value of the Convertible RIs you're trading in to calculate the minimal number of Convertible RIs you'll receive while ensuring the result of the exchange gives you Convertible RIs of equal or greater value.

Q: Are there exchange limits for Convertible RIs?

No, there are no exchange limits for Convertible RIs.

Q: Do I have the freedom to choose any instance type when I exchange my Convertible RIs?

No, you can only exchange into Convertible RIs that are currently offered by AWS.

Q: Can I upgrade the payment option associated with my Convertible RI?

Yes, you can upgrade the payment option associated with your RI. For example, you can exchange your No Upfront RIs for Partial or All Upfront RIs to benefit from better pricing. You cannot change the payment option from All Upfront to No Upfront, and cannot change from Partial Upfront to No Upfront.

Q: Do Convertible RIs allow me to benefit from price reductions when they happen?

Yes, you can exchange your RIs to benefit from lower pricing. For example, if the price of new Convertible RIs reduces by 10%, you can exchange your Convertible RIs and benefit from the 10% reduction in price.

EC2 Fleet

Q: What is Amazon EC2 Fleet?

With a single API call, EC2 Fleet lets you provision compute capacity across different instance types, Availability Zones and across On-Demand, Reserved Instances (RI) and Spot Instances purchase models to help optimize scale, performance and cost.

Q: If I currently use Amazon EC2 Spot Fleet should I migrate to Amazon EC2 Fleet?

If you are leveraging Amazon EC2 Spot Instances with Spot Fleet, you can continue to use that. Spot Fleet and EC2 Fleet offer the same functionality. There is no requirement to migrate.

Q: Can I use Reserved Instance (RI) discounts with Amazon EC2 Fleet?

Yes. Similar to other EC2 APIs or other AWS services that launch EC2 instances, if the On-Demand instance launched by EC2 Fleet matches an existing RI, that instance will receive the RI discount. For example, if you own Regional RIs for M4 instances and you have specified only M4 instances in your EC2 Fleet, RI discounts will be automatically applied to this usage of M4.

Q: Will Amazon EC2 Fleet failover to On-Demand if EC2 Spot capacity is not fully fulfilled?

No, EC2 Fleet will continue to attempt to meet your desired Spot capacity based on the number of Spot instances you requested in your Fleet launch specification.

Q: What is the pricing for Amazon EC2 Fleet?

EC2 Fleet comes at no additional charge; you only pay for the underlying resources that EC2 Fleet launches.

Q: Can you provide a real world example of how I can use Amazon EC2 Fleet?

There are a number of ways to take advantage of Amazon EC2 Fleet, such as in big data workloads, containerized application, grid processing workloads, etc. In this example of a genomic sequencing workload, you can launch a grid of worker nodes with a single API call: select your favorite instances, assign weights for these instances, specify target capacity for On-Demand and Spot Instances, and build a fleet within seconds to crunch through genomic data quickly.

Q: How can I allocate resources in an Amazon EC2 Fleet?

By default, EC2 Fleet will launch the On-Demand option that is the lowest price. For Spot Instances, EC2 Fleet provides three allocation strategies: capacity-optimized, lowest price and diversified. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing.

The lowest price strategy allows you to provision Spot Instances in pools that provide the lowest price per unit of capacity at the time of the request. The diversified strategy allows you to provision Spot Instances across multiple Spot pools and you can maintain your fleet's target capacity to increase application.

Q: Can I submit a multi-region Amazon EC2 Fleet request?

No, we do not support multi-region EC2 Fleet requests.

Q: Can I tag an Amazon EC2 Fleet?

Yes. You can tag an EC2 Fleet request to create business-relevant tag groupings to organize resources along technical, business, and security dimensions.

Q: Can I modify my Amazon EC2 Fleet?

Yes, you can modify the total target capacity of your EC2 Fleet when in maintain mode. You may need to cancel the request and submit a new one to change other request configuration parameters.

Q: Can I specify a different AMI for each instance type that I want to use?

Yes, simply specify the AMI you'd like to use in each launch specification you provide in your EC2 Fleet.

Amazon EC2 Capacity Blocks for ML

Q: What are Amazon EC2 Capacity Blocks for ML?

Amazon EC2 Capacity Blocks for ML allow you to reserve GPU instances in an Amazon EC2 UltraClusters to run your machine learning (ML) workloads. With Amazon EC2 Capacity Blocks, you can reserve GPU capacity starting on a future date for duration up to 14 days and in cluster sizes of one to 64 instances. When your EC2 Capacity Block reservation date and time arrives, you will be able to launch your instances and use them until your reservation time ends.

Q: Why should I use EC2 Capacity Blocks?

EC2 Capacity Blocks make it easy to access the highest-performing GPU instances in Amazon EC2 for ML, even in the face of industry-wide GPU shortages. Use EC2 Capacity Blocks to ensure capacity availability for GPU instances to plan your ML development with confidence. EC2 Capacity Blocks are delivered in EC2 UltraClusters so you can leverage the best network latency and throughput performance available in EC2.

Q: When should I use Amazon EC2 Capacity Blocks instead of On-Demand Capacity Reservations?

You should use EC2 Capacity Blocks when you need short-term capacity assurance to train or fine-tune ML models, run experiments, build prototypes, or handle surges in demand for ML applications. With EC2 Capacity Blocks, you can have peace of mind knowing you'll have access to GPU resources on a specific date to run your ML workloads. You should use On-Demand Capacity Reservations for all other workload types that need assurance, such as business-critical applications, regulatory requirements, or disaster recovery.

Q: How do I get started with EC2 Capacity Blocks?

You can search for available EC2 Capacity Blocks based on your capacity needs in the AWS Management Console, AWS Command Line Interface (AWS CLI), and AWS SDKs. Once you purchase an EC2 Capacity Block, a reservation is created in your account. When the EC2 Capacity Block start time arrives, EC2 will emit an event through Amazon EventBridge to indicate that the reservation is now active and available for use. To use an active EC2 Capacity Block, select the "Capacity Block" purchase option and target the capacity reservation ID for your EC2 Capacity Block while launching EC2 instances. As your EC2 Capacity Block end time approaches, EC2 will emit event through EventBridge letting you know your reservation is ending soon so you can checkpoint your workload. Around 30 minutes before your EC2 Capacity Block expires, AWS will begin terminating any running instances. The amount you are charged for your EC2 Capacity Block does not include the last 30 minutes of the reservation.

Q: Which instance types do EC2 Capacity Blocks support, and which AWS Regions are they available in?

EC2 Capacity Blocks support EC2 p5.48xlarge instances in the AWS US East (N. Virginia) and US East (Ohio) regions and EC2 p4d.24xlarge instances in US East (Ohio) and US West (Oregon) regions.

Q: What size options are available with EC2 Capacity Blocks?

EC2 Capacity Blocks are available in cluster sizes of 1, 2, 4, 8, 16, 32, and 64 instances, and they can be reserved for up to 14 days in one-day multiples.

Q: How far in advance can I reserve an EC2 Capacity Block?

You can purchase an EC2 Capacity Block as far out as eight weeks into the future. All EC2 Capacity Blocks reservations start at 11:30 AM Coordinated Universal Time (UTC).

Q: What happens if there are no EC2 Capacity Blocks available that meet my specifications?

If there are no EC2 Capacity Blocks that match your requirements, you can retry your request with different input parameters. We recommend that you use the widest date range possible in your search requests for the best chance at finding an EC2 Capacity Block.

Q: Can I modify or cancel my EC2 Capacity Block?

No, an EC2 Capacity Block cannot be modified or canceled once it is reserved.

Q: How much do EC2 Capacity Blocks cost?

When you search for an EC2 Capacity Block across dates, AWS returns the lowest-priced offering available that meets your specifications in the date range you provide. The price for an EC2 Capacity Block depends on total available supply and demand at the time you purchase the reservation. You can view the price of an EC2 Capacity Block offering before you reserve it, and the price of an EC2 Capacity Block is charged up front at the time the reservation is made. The price of an EC2 Capacity Block does not change after it is reserved. When you launch instances in an active EC2 Capacity Block, you will only be charged for the usage of any premium operating system on a pay-as-you-go basis.

Q: Do Savings Plans and Reserved Instances (RI) discounts apply to EC2 Capacity Blocks?

No, EC2 Capacity Blocks are not covered by Savings Plans or RI discounts.

Q: Can I use EC2 Capacity Blocks with Amazon SageMaker?

At this time, EC2 Capacity Blocks only support EC2 instances.

On-Demand Capacity Reservation

An On-Demand Capacity Reservation is an EC2 offering that you can use to create and manage reserved capacity on EC2. You can create an On-Demand Capacity Reservation by choosing an Availability Zone (AZ) and quantity (number of instances) along with other instance

specifications such as instance type and tenancy. Once created, the EC2 capacity is held for you regardless of whether you run the instances or not.

Q: How much do On-Demand Capacity Reservations cost?

When the On-Demand Capacity Reservation is active, you will pay equivalent instance charges whether you run the instances or not. If you do not use the reservation, the charge will show up as an unused reservation on your EC2 bill. When you run an instance that matches the attributes of a reservation, you just pay for the instance and nothing for the reservation. There are no upfront or additional charges.

For example, if you create an On-Demand Capacity Reservation for 20 c5.2xlarge instances and you run 15 c5.2xlarge instances, you will be charged for 15 instances and five unused instances in the reservation (effectively charged for 20 instances).

Q: Can I get a discount for On-Demand Capacity Reservation usage?

Yes. Savings Plans or Regional RI (RI scoped to a Region) discounts apply to On-Demand Capacity Reservations. When you are running an instance within your reservation, you are not charged for the reservation. Savings Plans or Regional RIs will apply to this usage as if it were On-Demand usage. When the reservation is not used, AWS Billing will automatically apply your discount when the attributes of the unused On-Demand Capacity Reservation match the attributes of an active Savings Plan or Regional RI.

For example, if you have a Regional RI for 10 c5.2xlarge instances and an unused On-Demand Capacity Reservation for 10 c5.2xlarge instances in the same Region, the RI discount will apply to all 10 instances on the reservation. Note that we apply Regional RI discounts preferentially to running instance usage before covering unused On-Demand Capacity Reservations. Meaning, if you have any other C5 instances running in the Region, we will apply the Regional RI first to those instances, and then we will apply the remaining discount to the unused On-Demand Capacity Reservation.

Note: A Regional RI is an EC2 RI scoped to an AWS Region. Zonal RIs (RIs scoped to an AZ within a Region) discounts do not apply to On-Demand Capacity Reservations, as Zonal RIs already come with a capacity reservation.

Q: When should I use Savings Plans, EC2 RIs, and On-Demand Capacity Reservations?

Use Savings Plans or Regional RIs to reduce your bill while committing to a one- or three-year term. Savings Plans offer significant savings over On-Demand, just like EC2 RIs, but automatically reduce customers' bills on compute usage across any AWS Region, even as usage changes. Use On-Demand Capacity Reservations if you need the additional confidence in your ability to launch instances. On-Demand Capacity Reservations can be created for any duration and can be managed independently of your Savings Plans or RIs. If you have Savings Plans or Regional RIs, they will automatically apply to matching On-Demand Capacity Reservations. This gives you the flexibility to selectively add On-Demand Capacity Reservations to a portion of your instance footprint and still reduce your bill for that usage.

Q: I have a Zonal RI (RI scoped to an AZ) that also provides a capacity reservation. How does this compare with an On-Demand Capacity Reservation?

A Zonal RI provides both a discount and a capacity reservation in a specific AZ in return for a one- to three-year commitment. An On-Demand Capacity Reservation allows you to create and manage reserved capacity independently of your RI commitment and term length.

You can use On-Demand Capacity Reservations with a Savings Plan or a Regional RI to get, at the minimum, all the benefits of a Zonal RI for no additional cost. You also get the enhanced flexibility of a Savings Plan (or Regional RI) and the features of an On-Demand Capacity Reservation: the ability to add or subtract from the reservation at any time, view reservation utilization in real time, and the ability to target an On-Demand Capacity Reservation for specific workloads.

Rescoping your Zonal RIs to a Region immediately gives you the AZ and instance-size flexibility in how RI discounts are applied. You can convert your Standard Zonal RIs to a Regional RI by modifying the scope of the RI from a specific AZ to a Region using the EC2 console or the ModifyReservedInstances API.

Q: I created an On-Demand Capacity Reservation. How can I use it?

An On-Demand Capacity Reservation is tied to a specific AZ and is, by default, automatically used by running instances in that AZ. When you launch new instances that match the reservation attributes, they will automatically match to the reservation.

You can also target a reservation for specific workloads/instances if you prefer. Refer to Linux or Windows technical documentation to learn more about the targeting option.

Q: How many instances am I allowed to reserve?

The number of instances that you are allowed to reserve is based on your account's On-Demand instance limit. You can reserve as many instances as that limit allows, minus the number of instances that are already running.

If you need a higher limit, contact your AWS sales representative or complete the Amazon EC2 instance request form with your use case and your instance increase will be considered. Limit increases are tied to the region they are requested for.

Q: Can I modify an On-Demand Capacity Reservation after it has started?

Yes. You can reduce the number of instances that you reserved at any time. You can also increase the number of instances (subject to availability). You can also modify the end time of your reservation. You cannot modify an On-Demand Capacity Reservation that has ended or has been deleted.

Q: Can I end an On-Demand Capacity Reservation after it has started?

Yes. You can end an On-Demand Capacity Reservation by canceling it using the console or API/SDK, or by modifying your reservation to specify an end time that makes it expire automatically. Running instances are unaffected by changes to your On-Demand Capacity Reservation, including deletion or expiration of a reservation.

Q: Where can I find more information about using On-Demand Capacity Reservations?

Refer to Linux or Windows technical documentation to learn about creating and using an On-Demand Capacity Reservation.

Q: Can I share an On-Demand Capacity Reservation with another AWS account?

Yes, you can share On-Demand Capacity Reservations with other AWS accounts or within your AWS Organization through AWS Resource Access Manager (AWS RAM). You can share EC2 On-Demand Capacity Reservations in three easy steps: Create a Resource Share using AWS RAM, add resources (On-Demand Capacity Reservations) to the Resource Share, and specify the target accounts that you wish to share the resources with.

Note that sharing of an On-Demand Capacity Reservation is not available to new AWS accounts or AWS accounts that have a limited billing history. New accounts that are linked to a qualified primary (payer) account or through an AWS Organization are exempt from this restriction.

Q: What happens when I share an On-Demand Capacity Reservation with another AWS account?

When an On-Demand Capacity Reservation is shared with other accounts, those accounts can consume the reserved capacity to run their EC2 instances. The exact behavior depends on the preferences set on the On-Demand Capacity Reservation. By default, On-Demand Capacity Reservations automatically match existing and new instances from other accounts that have shared access to the reservation. You can also target an On-Demand Capacity Reservation for specific workloads/instances. Individual accounts can control which of their instances consume On-Demand Capacity Reservations. Refer to Linux or Windows technical documentation to learn more about the instance matching options.

Q: Is there an additional charge for sharing a reservation?

No, there is no additional charge for sharing a reservation.

Q: Who gets charged when an On-Demand Capacity Reservation is shared across multiple accounts?

If multiple accounts are consuming an On-Demand Capacity Reservation, each account gets charged for its own instance usage. Unused reserved capacity, if any, gets charged to the account that owns the On-Demand Capacity Reservation. If there is a consolidated billing arrangement among the accounts that share an On-Demand Capacity Reservation, the primary account gets billed for instance usage across all the linked accounts.

Q: Can I prioritize access to an On-Demand Capacity Reservation among the AWS accounts that have shared access?

No. Instance spots in an On-Demand Capacity Reservation are available on a first-come, first-served basis to any account that has shared access.

Q: How can I communicate the AZ of an On-Demand Capacity Reservation with another account, given AZ name mappings could be different across AWS accounts?

You can now use an Availability Zone ID (AZ ID) instead of an AZ name. An AZ ID is a static reference and provides a consistent way of identifying the location of a resource across all your accounts. This makes it easier for you to provision resources centrally in a single account and share them across multiple accounts.

Q: Can I stop sharing my On-Demand Capacity Reservation once I have shared it?

Yes, you can stop sharing a reservation after you have shared it. When you stop sharing an On-Demand Capacity Reservation with specific accounts or stop sharing entirely, other accounts lose the ability to launch new instances into the On-Demand Capacity Reservation. Any capacity occupied by instances running from other accounts will be restored to the On-Demand Capacity Reservation for your use (subject to availability).

Q: Where can I find more information about sharing On-Demand Capacity Reservations?

Refer to Linux or Windows technical documentation to learn about sharing On-Demand Capacity Reservations.

Q: Can I get a discount for On-Demand Capacity Reservation usage?

Yes. Savings Plans or Regional RI discounts apply to On-Demand Capacity Reservations. AWS Billing automatically applies the discount when the attributes of an On-Demand Capacity Reservation match the attributes of a Savings Plan or Regional RI. When an On-Demand Capacity Reservation is used by an instance, you are only charged for the instance (with Savings Plan or RI discounts applied). Discounts are preferentially applied to instance usage before covering unused On-Demand Capacity Reservations.

Note: A Regional RI is an EC2 RI scoped to an AWS Region. Zonal RI (RIs scoped to an AZ within a Region) discounts do not apply to On-Demand Capacity Reservations, as Zonal RIs already come with a capacity reservation.

Reserved Instances

Q: What is a Reserved Instance?

A Reserved Instance (RI) is an EC2 offering that provides you with a significant discount on EC2 usage when you commit to a one-year or three-year term.

Q: What are the differences between Standard RIs and Convertible RIs?

Standard RIs offer a significant discount on EC2 instance usage when you commit to a particular instance family. Convertible RIs offer you the option to change your instance configuration during the term, and still receive a discount on your EC2 usage. For more information on Convertible RIs, please click here.

Q: Do RIs provide a capacity reservation?

Yes, when a Standard or Convertible RI is scoped to a specific Availability Zone (AZ), instance capacity matching the exact RI configuration is reserved for your use (these are referred to as “zonal RIs”). Zonal RIs give you additional confidence in your ability to launch instances when you need them.

You can also choose to forgo the capacity reservation and purchase Standard or Convertible RIs that are scoped to a region (referred to as “regional RIs”). Regional RIs automatically apply the discount to usage across AZs and instance sizes in a region, making it easier for you to take advantage of the RI’s discounted rate.

Q: When should I purchase a zonal RI?

If you want to take advantage of the capacity reservation, then you should buy an RI in a specific AZ.

Q: When should I purchase a regional RI?

If you do not require the capacity reservation, then you should buy a regional RI. Regional RIs provide AZ and instance size flexibility, which offer broader applicability of the RI's discounted rate.

Q: What are AZ and instance size flexibility?

AZ and instance size flexibility make it easier for you to take advantage of your regional RI's discounted rate. AZ flexibility applies your RI's discounted rate to usage in any AZ in a Region, while instance size flexibility applies your RI's discounted rate to usage of any size within an instance family. Let's say you own an m5.2xlarge Linux/Unix regional RI with default tenancy in US East (N. Virginia). Then this RI's discounted rate can automatically apply to two m5.xlarge instances in us-east-1a or four m5.large instances in us-east-1b.

Q: What types of RIs provide instance size flexibility?

Linux/Unix regional RIs with the default tenancy provide instance size flexibility. Instance size flexibility is not available on RIs of other platforms such as Windows, Windows with SQL Standard, Windows with SQL Server Enterprise, Windows with SQL Server Web, RHEL, and SLES or G4 instances.

Q: Do I need to take any action to take advantage of AZ and instance size flexibility?

Regional RIs do not require any action to take advantage of AZ and instance size flexibility.

Q: I own zonal RIs. How do I assign them to a region?

You can assign your Standard zonal RIs to a region by modifying the scope of the RI from a specific AZ to a Region from the EC2 console or by using the ModifyReservedInstances API.

Q: How do I purchase an RI?

To get started, you can purchase an RI from the EC2 console or by using the AWS CLI. Simply specify the instance type, platform, tenancy, term, payment option, and region or AZ.

Q: Can I purchase an RI for a running instance?

Yes, AWS will automatically apply an RI's discounted rate to any applicable instance usage from the time of purchase. Visit the Getting Started page to learn more.

Q: Can I control which instances are billed at the discounted rate?

No. AWS automatically optimizes which instances are charged at the discounted rate to ensure you always pay the lowest amount. For information about billing, and how it applies to RIs, see Billing Benefits and Payment Options.

Q: How does instance size flexibility work?

EC2 uses the scale shown below to compare different sizes within an instance family. In the case of instance size flexibility on RIs, this scale is used to apply the discounted rate of RIs to the normalized usage of the instance family. For example, if you have an m5.2xlarge RI that is scoped to a region, then your discounted rate could apply towards the usage of 1 m5.2xlarge or 2 m5.xlarge instances.

Click here to learn more about how instance size flexibility of RIs applies to your EC2 usage. And click here to learn about how instance size flexibility of RIs is presented in the Cost and Usage Report.

Instance Size	Normalization Factor
nano	0.25
micro	0.5
small	1
medium	2

large	4
xlarge	8
2xlarge	16
4xlarge	32
8xlarge	64
9xlarge	72
10xlarge	80
12xlarge	96
16xlarge	128
18xlarge	144
24xlarge	192
32xlarge	256

Q: Can I change my RI during its term?

Yes, you can modify the AZ of the RI, change the scope of the RI from AZ to Region (and the other way around), or modify instance sizes within the same instance family (on the Linux/Unix platform).

Q: Can I change the instance type of my RI during its term?

Yes. Convertible RIs offer you the option to change the instance type, operating system, tenancy or payment option of your RI during its term. Please refer to the Convertible RI section of the FAQ for additional information.

Q: What are the different payment options for RIs?

You can choose from three payment options when you purchase an RI. With the All Upfront option, you pay for the entire RI term with one upfront payment. With the Partial Upfront option, you make a low upfront payment and are then charged a discounted hourly rate for the instance for the duration of the RI term. The No Upfront option does not require any upfront payment and provides a discounted hourly rate for the duration of the term.

Q: When are RIs activated?

The billing discount and capacity reservation (if applicable) is activated once your payment has successfully been authorized. You can view the status (pending | active | retired) of your RIs on the "Reserved Instances" page of the Amazon EC2 console.

Q: Do RIs apply to Spot instances or instances running on a Dedicated Host?

No, RIs do not apply to Spot instances or instances running on Dedicated Hosts. To lower the cost of using Dedicated Hosts, purchase Dedicated Host Reservations.

Q: How do RIs work with Consolidated Billing?

Our system automatically optimizes which instances are charged at the discounted rate to ensure that the consolidated accounts always pay the lowest amount. If you own RIs that apply to an AZ, then only the account which owns the RI will receive the capacity reservation. However, the discount will automatically apply to usage in any account across your consolidated billing family.

Q: Can I get a discount on RI purchases?

Yes, EC2 provides tiered discounts on RI purchases. These discounts are determined based on the total list value (non-discounted price) for the active RIs you have per Region. Your total list value is the sum of all expected payments for an RI within the term, including both the upfront and recurring hourly payments. The tier ranges and corresponding discounts are shown below.

Tier Range of List Value	Discount on Upfront	Discount on Hourly
Less than $500k	0%	0%
$500k-$4M	5%	5%
$4M-$10M	10%	10%
More than $10M	Call Us	

Q: Can you help me understand how volume discounts are applied to my RI purchases?

Sure. Let's assume that you currently have $400,000 worth of active RIs in the US-east-1 region. Now, if you purchase RIs worth $150,000 in the same region, then the first $100,000 of this purchase would not receive a discount. However, the remaining $50,000 of this purchase would be discounted by 5 percent, so you would only be charged $47,500 for this portion of the purchase over the term based on your payment option.

To learn more, please visit the Understanding Reserved Instance Discount Pricing Tier portion of the Amazon EC2 User Guide.

Q: How do I calculate the list value of an RI?

Here is a sample list value calculation for three-year Partial Upfront Reserved Instances:

3yr Partial Upfront Volume Discount Value in US-East

	Upfront $	Recurring Hourly $	Recurring Hourly Value	List Value

m3.xlarge	$ 1,345	$ 0.060	$ 1,577	$ 2,922
c3.xlarge	$ 1,016	$ 0.045	$ 1,183	$ 2,199

Q: How are volume discounts calculated if I use Consolidated Billing?

If you leverage Consolidated Billing, AWS will use the aggregate total list price of active RIs across all of your consolidated accounts to determine which volume discount tier to apply. Volume discount tiers are determined at the time of purchase, so you should activate Consolidated Billing prior to purchasing RIs to ensure that you benefit from the largest possible volume discount that your consolidated accounts are eligible to receive.

Q: Do Convertible RIs qualify for Volume Discounts?

No, but the value of each Convertible RI that you purchase contributes to your volume discount tier standing.

Q: How do I determine which volume discount tier applies to me?

To determine your current volume discount tier, please consult the Understanding Reserved Instance Discount Pricing Tiers portion of the Amazon EC2 User Guide.

Q: Will the cost of my RIs change, if my future volume qualifies me for other discount tiers?

No. Volume discounts are determined at the time of purchase, therefore the cost of your RIs will continue to remain the same as you qualify for other discount tiers. Any new purchase will be discounted according to your eligible volume discount tier at the time of purchase.

Q: Do I need to take any action at the time of purchase to receive volume discounts?

No, you will automatically receive volume discounts when you use the existing PurchaseReservedInstance API or EC2 Management Console interface to purchase RIs. If you purchase more than $10M worth of RIs contact us about receiving discounts beyond those that are

automatically provided.

Reserved Instance Marketplace

Q: What is the Reserved Instance (RI) Marketplace?

The RI Marketplace is an online marketplace that provides AWS customers the flexibility to sell their Amazon EC2 RIs to other businesses and organizations. Customers can also browse the RI Marketplace to find an even wider selection of RI term lengths and pricing options sold by other AWS customers.

Q: When can I list an RI on the RI Marketplace?

You can list an RI when:

- You've registered as a seller in the RI Marketplace.
- You've paid for your RI.
- You've owned the RI for longer than 30 days.

Q: Can RIs be transferred?

EC2 Reserved Instances are only transferrable in accordance with the requirements of the RI Marketplace provided in AWS Service Terms and cannot otherwise be transferred.

Q: Can I sell any RI on the EC2 RI Marketplace?

No, AWS prohibits the resale of RIs purchased as part of a discount program per AWS Service Terms. Any All Upfront, Partial Upfront, or No Unfront RIs that were purchased directly from AWS or from EC2 RI Marketplace that received a discount from AWS (for example, RI Volume Discount or other discount programs) are not eligible for sale on the EC2 RI Marketplace.

Q: How will I register as a seller for the RI Marketplace?

To register for the RI Marketplace, you can enter the registration workflow by selling an RI from the EC2 Management Console or setting up your profile from the "Account Settings" page on the AWS portal. No matter the route, you will need to complete the following steps:

1. Start by reviewing the overview of the registration process.
2. Log in to your AWS Account.
3. Enter in the bank account into which you want us to disburse funds. Once you select "Continue," we will set that bank account as the default disbursement option.
4. In the confirmation screen, choose "Continue to Console to Start Listing."

If you exceed $20,000 in sales of RI, or plan to sell 50 or more RIs, you will need to provide tax information before you can list your RIs. Choose "Continue with Tax Interview." During the tax interview pipeline, you will be prompted to enter your company name, contact name, address, and Tax Identification Number using the TIMS workflow.

Additionally, if you plan to sell RIs worth more than $50,000 per year you will also need to file a limit increase.

Q: How will I know when I can start selling on the RI Marketplace?

You can start selling on the RI Marketplace after you have added a bank account through the registration pipeline. Once activation is complete, you will receive a confirmation email. However, it is important to note that you will not be able to receive disbursements until we are able to receive verification from your bank, which may take up to two weeks, depending on the bank you use.

Q: How do I list an RI for sale?

To list an RI, simply complete these steps in the Amazon EC2 console:

1. Select the RIs that you wish to sell, and choose "Sell Reserved Instances." If you have not completed the registration process, you will be prompted to register using the registration pipeline.

2. For each RI type, set the number of instances you'd like to sell, and the price for the one-time fee you would like to set. Note that you can set the one-time price to different amounts depending on the amount of time remaining so that you don't have to keep adjusting your one-time price if your RI doesn't sell quickly. By default you just need to set the current price and we will automatically decrease the one-time price by the same increment each month.
3. Once you have configured your listing, a final confirmation screen will appear. Choose "Sell Reserved Instance."

Q: Which RIs can I list for sale?

You can list any RIs that have been active for at least 30 days, and for which we have received payment. Typically, this means that you can list your reservations once they are in the **active** state. It is important to note that if you are an invoice customer, your RI can be in the **active** state prior to AWS receiving payment. In this case, your RI will not be listed until we have received your payment.

Q: How are listed RIs displayed to buyers?

RIs (both third-party and those offered by AWS) that have been listed on the RI Marketplace can be viewed in the "Reserved Instances" section of the Amazon EC2 console. You can also use the DescribeReservedInstancesListings API call.

The listed RIs are grouped based on the type, term remaining, upfront price, and hourly price. This makes it easier for buyers to find the right RIs to purchase.

Q: How much of my RI term can I list?

You can sell an RI for the term remaining, rounded down to the nearest month. For example, if you had 9 months and 13 days remaining, you will list it for sale as a 9-month-term RI.

Q: Can I remove my RI after I've listed it for sale?

Yes, you can remove your RI listings at any point until a sale is pending (meaning a buyer has bought your RI and confirmation of payment is pending).

Q: Which pricing dimensions can I set for the RIs that I want to list?

Using the RI Marketplace, you can set an upfront price you'd be willing to accept. You cannot set the hourly price (which will remain the same as was set on the original RI), and you will not receive any funds collected from payments associated with the hourly prices.

Q: Can I still use my reservation while it is listed on the RI Marketplace?

Yes, you will continue to receive the capacity and billing benefit of your reservation until it is sold. Once sold, any running instance that was being charged at the discounted rate will be charged at the On-Demand rate until and unless you purchase a new reservation, or terminate the instance.

Q: Can I resell an RI that I purchased from the RI Marketplace?

Yes, you can resell RIs purchased from the RI Marketplace just like any other RI.

Q. Are there any restrictions when selling RIs?

Yes, you must have a US bank account to sell RIs in the RI Marketplace. Support for non-US bank accounts will be coming soon. Also, you may not sell RIs in the US GovCloud Region.

Q: Can I sell RIs purchased from the public volume pricing tiers?

No, this capability is not yet available.

Q: Is there a charge for selling RIs in the RI Marketplace?

Yes, AWS charges a service fee of 12% of the total upfront price of each RI that you sell in the RI Marketplace.

Q: Can AWS sell subsets of my listed RIs?

Yes, AWS may potentially sell a subset of the quantity of RIs that you have listed. For example, if you list 100 RIs, we may only have a buyer interested in purchasing 50 of them. We will sell those 50 instances and continue to list your remaining 50 RIs until and unless you decide not to list them any longer.

Q: How do buyers pay for RIs that they've purchased?

Payment for completed RI sales are done via ACH wire transfers to a US bank account.

Q: When will I receive my money?

Once AWS has received funds from the customer that has bought your reservation, we will disburse funds via wire transfer to the bank account you specified when you registered for the RI Marketplace.

Then, we will send you an email notification letting you know that we've wired you the funds. Typically, funds will appear in your account within 3-5 days of when your RI was sold.

Q: If I sell my RI in the RI Marketplace, will I get refunded for the Premium Support I was charged, too?

No, you will not receive a prorated refund for the upfront portion of the AWS Premium Support Fee.

Q: Will I be notified about RI Marketplace activities?

Yes, you will receive a single email once a day that details your RI Marketplace activity whenever you create or cancel RI listings, buyers purchase your listings, or AWS disburses funds to your bank account.

Q: What information is exchanged between the buyer and seller to help with the transaction tax calculation?

The buyer's city, state, zip+4, and country information will be provided to the seller via a disbursement report. This information will enable sellers to calculate any necessary transaction taxes they need to remit to the government (e.g., sales tax, value-added tax, etc.). The legal entity name of the seller will also be provided on the purchase invoice.

Q: Are there any restrictions on the customers when purchasing third-party RIs?

Yes, you cannot purchase your own listed RIs, including those in any of your linked accounts (via Consolidated Billing).

Q: Do I have to pay for Premium Support when purchasing RIs from the RI Marketplace?

Yes, if you are a Premium Support customer, you will be charged for Premium Support when you purchase an RI through the RI Marketplace.

Savings Plans

Q: What is Savings Plans?

Savings Plans is a flexible pricing model that offers low prices on EC2, Lambda and Fargate usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a one- or three-year term. When you sign up for Savings Plans, you will be charged the discounted Savings Plans price for your usage up to your commitment. For example, if you commit to $10 of compute usage an hour, you will get the Savings Plans prices on that usage up to $10 and any usage beyond the commitment will be charged On Demand rates.

Q: What types of Savings Plans does AWS offer?

AWS offers two types of Savings Plans:

1. Compute Savings Plans provide the most flexibility and help to reduce your costs by up to 66%. These plans automatically apply to EC2 instance usage regardless of instance family, size, AZ, region, OS or tenancy, and also apply to AWS Fargate and Lambda usage. For example, with Compute Savings Plans, you can change from C4 to M5 instances, shift a workload from EU (Ireland) to EU (London), or move a workload from EC2 to Fargate or Lambda at any time and automatically continue to pay the Savings Plans price.
2. EC2 Instance Savings Plans provides the lowest prices, offering savings up to 72% in exchange for commitment to usage of individual instance families in a Region (e.g., M5 usage in N. Virginia). This automatically reduces your cost on the selected instance family in that region regardless of AZ, size, OS or tenancy. EC2 Instance Savings Plans give you the flexibility to change your usage between instances

within a family in that region. For example, you can move from c5.xlarge running Windows to c5.2xlarge running Linux and automatically benefit from the Savings Plan prices.

Q: How do Savings Plans compare to EC2 RIs?

Savings Plans offers significant savings over On Demand, just like EC2 RIs, but automatically reduce your bills on compute usage across any AWS region, even as usage changes. This provides you the flexibility to use the compute option that best suits your needs and continue to save money, all without having to perform exchanges or modifications.

Compute Savings Plans, which provide savings up to 66% (just like Convertible RIs), automatically reduce your cost on any EC2 instance usage regardless of region, instance family, size, OS, tenancy and even on AWS Fargate and Lambda. EC2 Instance Savings Plans, which provide savings up to 72% (just like Standard RIs), automatically save you money on any instance usage within a given EC2 instance family in a chosen region (e.g., M5 in N. Virginia) regardless of size, OS, or tenancy.

Q: Do Savings Plans provide capacity reservations for EC2 instances?

No, Savings Plans do not provide a capacity reservation. You can however reserve capacity with On-Demand Capacity Reservations and pay lower prices on them with Savings Plans.

Q: How do I get started with Savings Plans?

You can get started with Savings Plans from AWS Cost Explorer in the AwS Management Console or by using the API/CLI. You can easily make a commitment to a Savings Plan by using the recommendations provided in AWS Cost Explorer, to realize the biggest savings. The recommended hourly commitment is based on your historical On Demand usage and your choice of plan type, term length, and payment option. Once you sign up for a Savings Plan, your compute usage will automatically be charged at the discounted Savings Plan prices and any usage beyond your commitment will be charged at regular On Demand rates.

Q: Can I continue to purchase EC2 RIs?

Yes. You can continue purchasing RIs to maintain compatibility with your existing cost management processes, and your RIs will work alongside Savings Plans to reduce your overall bill. However as your RIs expire we encourage you to sign up for Savings Plans as they offer the same savings as RIs, but with additional flexibility.

Spot Instances

Q. What is a Spot Instance?

Spot Instances are spare EC2 capacity that can save you up to 90% off of On-Demand prices that AWS can interrupt with a 2-minute notification. Spot uses the same underlying EC2 instances as On-Demand and Reserved Instances, and is best suited for fault-tolerant, flexible workloads. Spot Instances provides an additional option for obtaining compute capacity and can be used along with On-Demand and Reserved Instances.

Q. How is a Spot Instance different than an On-Demand instance or Reserved Instance?

While running, Spot Instances are exactly the same as On-Demand or Reserved instances. The main differences are that Spot Instances typically offer a significant discount off the On-Demand prices, your instances can be interrupted by Amazon EC2 for capacity requirements with a 2-minute notification, and Spot prices adjust gradually based on long term supply and demand for spare EC2 capacity.

See here for more details on Spot Instances.

Q. How do I purchase and start up a Spot instance?

Spot instances can be launched using the same tools you use to launch instances today, including AWS Management Console, Auto-Scaling Groups, Run Instances and Spot Fleet. In addition many AWS services support launching Spot instances such as EMR, ECS, Datapipeline, CloudFormation and Batch.

To start up a Spot Instance, you simply need to choose a Launch Template and the number of instances you would like to request.

See here for more details on how to request Spot Instances.

Q. How many Spot Instances can I request?

You can request Spot Instances up to your Spot limit for each region. Note that customers new to AWS might start with a lower limit. To learn more about Spot Instance limits, please refer to the Amazon EC2 User Guide.

If you would like a higher limit, complete the Amazon EC2 instance request form with your use case and your instance increase will be considered. Limit increases are tied to the region they were requested for.

Q. What price will I pay for a Spot Instance?

You pay the Spot price that's in effect at the beginning of each instance-hour for your running instance. If Spot price changes after you launch the instance, the new price is charged against the instance usage for the subsequent hour.

Q. What is a Spot capacity pool?

A Spot capacity pool is a set of unused EC2 instances with the same instance type, operating system, and Availability Zone. Each spot capacity pool can have a different price based on supply and demand.

Q. What are the best practices to use Spot Instances?

We highly recommend using multiple Spot capacity pools to maximize the amount of Spot capacity available to you. EC2 provides built-in automation to find the most cost-effective capacity across multiple Spot capacity pools using EC2 Auto Scaling, EC2 Fleet or Spot Fleet. For more information, please see Spot Best Practices.

Q. How can I determine the status of my Spot request?

You can determine the status of your Spot request via Spot Request Status code and message. You can access Spot Request Status information on the Spot Instance page of the EC2 console of the AWS Management Console, API and CLI. For more information, please visit

the Amazon EC2 Developer guide.

Q. Are Spot Instances available for all instance families and sizes and in all regions?

Spot Instances are available in all public AWS regions. Spot is available for nearly all EC2 instance families and sizes, including the newest compute-optimized instances, accelerated graphics, and FPGA instance types. A full list of instance types supported in each region is listed here.

Q. Which operating systems are available as Spot Instances?

Linux/UNIX, Windows Server and Red Hat Enterprise Linux (RHEL) are available. Windows Server with SQL Server is not currently available.

Q. Can I use a Spot Instance with a paid AMI for third-party software (such as IBM's software packages)?

Not at this time.

Q. Can I stop my running Spot Instances?

Yes, you can "stop" your running Spot Instances when they are not needed and keep these stopped instances for later use, instead of terminating instances or cancelling the Spot request. Stop is available for persistent Spot requests.

Q. How can I stop the Spot Instances?

You can stop your Spot Instances by calling the StopInstances API and providing Instance Ids of the Spot Instances similar to stopping your On-Demand Instances. You can also do this through the AWS Management Console by selecting your instance, then clicking Actions > Instance State > Stop.

Q. How can I start the stopped Spot Instances?

You can start the stopped Spot Instances by calling the StartInstances API and providing Instance Ids of the Spot Instances similar to starting your On-Demand Instances. You can also do this through the AWS Management Console by selecting your instance, then clicking Actions > Instance State > Start.

Note: The Spot Instances will start only if Spot capacity is still available within your maximum price. Spot evaluates capacity availability every time whenever you will start the stopped Spot instances.

Q: How can I tell whether I have stopped my Spot Instance or it has been interrupted?

You can tell that the Spot Instance has been stopped by you or interrupted by looking at the Spot Request Status code. This is visible as Spot Request Status on the Spot Requests page of the AWS Management Console or in the DescribeSpotInstanceRequests API response as "status-code" field.

If the Spot request status code is "instance-stopped-by-user", it means that you have stopped your spot instance.

Q. How will I be charged if my Spot instance is stopped or interrupted?

If your Spot instance is terminated or stopped by Amazon EC2 in the first instance hour, you will not be charged for that usage. However, if you stop or terminate the Spot instance yourself, you will be charged to the nearest second. If the Spot instance is terminated or stopped by Amazon EC2 in any subsequent hour, you will be charged for your usage to the nearest second. If you are running on Windows or Red Hat Enterprise Linux (RHEL) and you stop or terminate the Spot instance yourself, you will be charged for an entire hour.

Q. When would my Spot Instance get interrupted?

Over the last 3 months, 92% of Spot Instance interruptions were from a customer manually terminating the instance because the application had completed its work.

In the circumstance EC2 needs to reclaim your Spot Instance it can be for two possible reasons, with the primary one being Amazon EC2 capacity requirements (e.g. On Demand or Reserved Instance usage). Alternatively, if you have chosen to set a "maximum Spot price" and the

Spot price rises above this, your instance will be reclaimed with a two-minute notification. This parameter determines the maximum price you would be willing to pay for a Spot instance hour, and by default, is set at the On-Demand price. As before, you continue to pay the Spot market price, not your maximum price, at the time your instance was running, charged in per-second increments.

Q. What happens to my Spot instance when it gets interrupted?

You can choose to have your Spot instances terminated, stopped or hibernated upon interruption. Stop and hibernate options are available for persistent Spot requests and Spot Fleets with the "maintain" option enabled. By default, your instances are terminated.

Refer to Spot Hibernation to learn more about handling interruptions.

Q. What is the difference between Stop and Hibernate interruption behaviors?

In the case of Hibernate, your instance gets hibernated and the RAM data persisted. In the case of Stop, your instance gets shut down and RAM is cleared.

In both the cases, data from your EBS root volume and any attached EBS data volumes is persisted. Your private IP address remains the same, as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow. Stop and Hibernate are available for Amazon EBS backed instances only. Local instance storage is not persisted.

Q. What if my EBS root volume is not large enough to store memory state (RAM) for Hibernate?

You should have sufficient space available on your EBS root volume to write data from memory. If the EBS root volume does not have enough space, hibernation will fail and the instance will get shut down instead. Ensure that your EBS volume is large enough to persist memory data before choosing the hibernate option.

Q. What is the benefit if Spot hibernates my instance on interruption?

With hibernate, Spot instances will pause and resume around any interruptions so your workloads can pick up from exactly where they left off. You can use hibernation when your instance(s) need to retain instance state across shutdown-startup cycles, i.e. when your applications

running on Spot depend on contextual, business, or session data stored in RAM.

Q. What do I need to do to enable hibernation for my Spot instances?

Refer to Spot Hibernation to learn about enabling hibernation for your Spot instances.

Q. Do I have to pay for hibernating my Spot instance?

There is no additional charge for hibernating your instance beyond the EBS storage costs and any other EC2 resources you may be using. You are not charged instance usage fees once your instance is hibernated.

Q. Can I resume a hibernated instance?

No, you will not be able to resume a hibernated instance directly. Hibernate-resume cycle is controlled by Amazon EC2. If an instance is hibernated by Spot, it will be resumed by Amazon EC2 when the capacity becomes available.

Q. Which instances and operating systems support hibernation?

Spot Hibernation is currently supported for Amazon Linux AMIs, Ubuntu and Microsoft Windows operating systems running on any instance type across C3, C4, C5, M4, M5, R3, R4 instances with memory (RAM) size less than 100 GiB.

To review the list of supported OS versions, refer to Spot Hibernation.

Q. How am I charged if Spot price changes while my instance is running?

You will pay the price per instance-hour set at the beginning of each instance-hour for the entire hour, billed to the nearest second.

Q. Where can I see my usage history for Spot instances and see how much I was billed?

The AWS Management Console makes a detailed billing report available which shows Spot instance start and termination/stop times for all instances. Customers can check the billing report against historical Spot prices via the API to verify that the Spot price they were billed is correct.

Q: Are Spot blocks (Fixed Duration Spot instances) ever interrupted?

Spot blocks are designed not to be interrupted and will run continuously for the duration you select, independent of Spot market price. In rare situations, Spot blocks may be interrupted due to AWS capacity needs. In these cases, we will provide a two-minute warning before we terminate your instance (termination notice), and you will not be charged for the affected instance(s).

Q. What is a Spot fleet?

A Spot Fleet allows you to automatically request and manage multiple Spot instances that provide the lowest price per unit of capacity for your cluster or application, like a batch processing job, a Hadoop workflow, or an HPC grid computing job. You can include the instance types that your application can use. You define a target capacity based on your application needs (in units including instances, vCPUs, memory, storage, or network throughput) and update the target capacity after the fleet is launched. Spot fleets enable you to launch and maintain the target capacity, and to automatically request resources to replace any that are disrupted or manually terminated. Learn more about Spot fleets.

Q. Is there any additional charge for making Spot Fleet requests?

No, there is no additional charge for Spot Fleet requests.

Q. What limits apply to a Spot Fleet request?

Visit the Spot Fleet Limits section of the Amazon EC2 User Guide to learn about the limits that apply to your Spot Fleet request.

Q. What happens if my Spot Fleet request tries to launch Spot instances but exceeds my regional Spot request limit?

If your Spot Fleet request exceeds your regional Spot instance request limit, individual Spot instance requests will fail with a "Spot request limit exceeded request status". Your Spot Fleet request's history will show any Spot request limit errors that the Fleet request received. Visit the Monitoring Your Spot Fleet section of the Amazon EC2 User Guide to learn how to describe your Spot Fleet request's history.

Q. Are Spot fleet requests guaranteed to be fulfilled?

No. Spot fleet requests allow you to place multiple Spot Instance requests simultaneously, and are subject to the same availability and prices as a single Spot Instance request. For example, if no resources are available for the instance types listed in your Spot Fleet request, we may be unable to fulfill your request partially or in full. We recommend that you to include all the possible instance types and availability zones that are suitable for your workloads in the Spot Fleet.

Q. Can I submit a multi-Availability Zone Spot Fleet request?

Yes, visit the Spot Fleet Examples section of the Amazon EC2 User Guide to learn how to submit a multi-Availability Zone Spot Fleet request.

Q. Can I submit a multi-region Spot Fleet request?

No, we do not support multi-region Fleet requests.

Q. How does Spot Fleet allocate resources across the various Spot Instance pools specified in the launch specifications?

The RequestSpotFleet API provides three allocation strategies: capacity-optimized, lowestPrice and diversified. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing.

The lowestPrice strategy allows you to provision your Spot Fleet resources in instance pools that provide the lowest price per unit of capacity at the time of the request. The diversified strategy allows you to provision your Spot Fleet resources across multiple Spot Instance pools. This enables you to maintain your fleet's target capacity and increase your application's availability as Spot capacity fluctuates.

Running your application's resources across diverse Spot Instance pools also allows you to further reduce your fleet's operating costs over time. Visit the Amazon EC2 User Guide to learn more.

Q. Can I tag a Spot Fleet request?

You can request to launch Spot Instances with tags via Spot Fleet. The Fleet by itself cannot be tagged.

Q. How can I see which Spot fleet owns my Spot Instances?

You can identify the Spot Instances associated with your Spot Fleet by describing your fleet request. Fleet requests are available for 48 hours after all its Spot Instances have been terminated. See the Amazon EC2 User Guide to learn how to describe your Spot Fleet request.

Q. Can I modify my Spot Fleet request?

Yes, you can modify the target capacity of your Spot Fleet request. You may need to cancel the request and submit a new one to change other request configuration parameters.

Q. Can I specify a different AMI for each instance type that I want to use?

Yes, simply specify the AMI you'd like to use in each launch specification you provide in your Spot Fleet request.

Q. Can I use Spot Fleet with Elastic Load Balancing, Auto Scaling, or Elastic MapReduce?

You can use Auto Scaling features with Spot Fleet such as target tracking, health checks, CloudWatch metrics, etc., and can attach instances to your Elastic load balancers (both classic and application load balancers). Elastic MapReduce has a feature named "Instance fleets" that provides capabilities similar to Spot Fleet.

Q. Does a Spot Fleet request terminate Spot Instances when they are no longer running in the lowest priced or capacity-optimized Spot pools and relaunch them?

No, Spot Fleet requests do not automatically terminate and relaunch instances while they are running. However, if you terminate a Spot Instance, Spot Fleet will replenish it with a new Spot Instance in the new lowest priced pool or capacity-optimized pool based on your allocation strategy.

Q: Can I use stop or Hibernation interruption behaviors with Spot Fleet?

Yes, stop-start and hibernate-resume are supported with Spot Fleet with “maintain” fleet option enabled.

Platform

Amazon Time Sync Service | Availability zones | Cluster instances | Hardware information | Micro instances | Nitro Hypervisor | Optimize CPUs

Amazon Time Sync Service

Q. How do I use this service?

The service provides an NTP endpoint at a link-local IP address (169.254.169.123) accessible from any instance running in a VPC. Instructions for configuring NTP clients are available for Linux and Windows.

Q. What are the key benefits of using this service?

A consistent and accurate reference time source is crucial for many applications and services. The Amazon Time Sync Service provides a time reference that can be securely accessed from an instance without requiring VPC configuration changes and updates. It is built on Amazon’s proven network infrastructure and uses redundant reference time sources to ensure high accuracy and availability.

Q. Which instance types are supported for this service?

All instances running in a VPC can access the service.

Availability Zones

Q: How isolated are Availability Zones from one another?

Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone.

Q: Is Amazon EC2 running in more than one AWS Region?

Yes. Please refer to Regional Products and Services for more details of our product and service availability by Region.

Q: How can I make sure that I am in the same Availability Zone as another developer?

We do not currently support the ability to coordinate launches into the same Availability Zone across AWS developer accounts. One Availability Zone name (for example, us-east-1a) in two AWS customer accounts may relate to different physical Availability Zones.

Q: If I transfer data between Availability Zones using public IP addresses, will I be charged twice for Regional Data Transfer (once because it's across zones, and a second time because I'm using public IP addresses)?

No. Regional Data Transfer rates apply if at least one of the following is true, but you are only charged once for a given instance even if both are true:

- The other instance is in a different Availability Zone, regardless of which type of address is used.
- Public or Elastic IP addresses are used, regardless of which Availability Zone the other instance is in.

Cluster instances

Q. What is a Cluster Compute Instance?

Cluster Compute Instances combine high compute resources with high performance networking for HPC applications and other demanding network-bound applications. Cluster Compute Instances provide similar functionality to other Amazon EC2 instances but have been specifically engineered to provide high performance networking.

Amazon EC2 cluster placement group functionality allows users to group Cluster Compute Instances in clusters—allowing applications to get the low-latency network performance necessary for tightly coupled node-to-node communication typical of many HPC applications. Cluster Compute Instances also provide significantly increased network throughput both within the Amazon EC2 environment and to the Internet. As a result, these instances are also well suited for customer applications that need to perform network-intensive operations.

Learn more about using this instance type for HPC applications.

Q. What kind of network performance can I expect when I launch instances in a cluster placement group?

The bandwidth an EC2 instance can utilize in a cluster placement group depends on the instance type and its networking performance specification. Inter-instance traffic within the same region can utilize 5 Gbps for single-flow and up to 25 Gbps for multiflow traffic. When launched in a placement group, select EC2 instances can utilize up to 10 Gbps for single-flow traffic.

Q. What is a Cluster GPU Instance?

Cluster GPU Instances provide general-purpose graphics processing units (GPUs) with proportionally high CPU and increased network performance for applications benefiting from highly parallelized processing that can be accelerated by GPUs using the CUDA and OpenCL programming models. Common applications include modeling and simulation, rendering and media processing.

Cluster GPU Instances give customers with HPC workloads an option beyond Cluster Compute Instances to further customize their high performance clusters in the cloud for applications that can benefit from the parallel computing power of GPUs.

Cluster GPU Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters—allowing applications to get the low-latency, high bandwidth network performance required for tightly coupled node-to-node communication

typical of many HPC applications.

Learn more about HPC on AWS.

Q. What is a High Memory Cluster Instance?

High Memory Cluster Instances provide customers with large amounts of memory and CPU capabilities per instance in addition to high network capabilities. These instance types are ideal for memory intensive workloads including in-memory analytics systems, graph analysis and many science and engineering applications.

High Memory Cluster Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters—allowing applications to get the low-latency, high bandwidth network performance required for tightly coupled node-to-node communication typical of many HPC and other network intensive applications.

Q. Does use of Cluster Compute and Cluster GPU Instances differ from other Amazon EC2 instance types?

Cluster Compute and Cluster GPU Instances use differs from other Amazon EC2 instance types in two ways.

First, Cluster Compute and Cluster GPU Instances use Hardware Virtual Machine (HVM) based virtualization and run only Amazon Machine Images (AMIs) based on HVM virtualization. Paravirtual Machine (PVM) based AMIs used with other Amazon EC2 instance types cannot be used with Cluster Compute or Cluster GPU Instances.

Second, in order to fully benefit from the available low latency, full bisection bandwidth between instances, Cluster Compute and Cluster GPU Instances must be launched into a cluster placement group through the Amazon EC2 API or AWS Management Console.

Q. What is a cluster placement group?

A cluster placement group is a logical entity that enables creating a cluster of instances by launching instances as part of a group. The cluster of instances then provides low latency connectivity between instances in the group. Cluster placement groups are created through the Amazon EC2 API or AWS Management Console.

Q. Are all features of Amazon EC2 available for Cluster Compute and Cluster GPU Instances?

Currently, Amazon DevPay is not available for Cluster Compute or Cluster GPU Instances.

Q. Is there a limit on the number of Cluster Compute or Cluster GPU Instances I can use and/or the size of cluster I can create by launching Cluster Compute Instances or Cluster GPU into a cluster placement group?

There is no limit specific for Cluster Compute Instances. For Cluster GPU Instances, you can launch 2 Instances on your own. If you need more capacity, please complete the Amazon EC2 instance request form (selecting the appropriate primary instance type).

Q. Are there any ways to optimize the likelihood that I receive the full number of instances I request for my cluster via a cluster placement group?

We recommend that you launch the minimum number of instances required to participate in a cluster in a single launch. For very large clusters, you should launch multiple placement groups, e.g. two placement groups of 128 instances, and combine them to create a larger, 256 instance cluster.

Q. Can Cluster GPU and Cluster Compute Instances be launched into a single cluster placement group?

While it may be possible to launch different cluster instance types into a single placement group, at this time we only support homogenous placement groups.

Q. If an instance in a cluster placement group is stopped then started again, will it maintain its presence in the cluster placement group?

Yes. A stopped instance will be started as part of the cluster placement group it was in when it stopped. If capacity is not available for it to start within its cluster placement group, the start will fail.

Hardware information

Q: What CPU options are available on EC2 instances?

EC2 instances offer a variety of CPU options to help customers balance performance and cost requirements. Depending on the instance type, EC2 offers a choice in CPU including AWS Graviton/Graviton2 processors (Arm), AMD processors (x86), and Intel processors (x86).

Q: What kind of hardware will my application stack run on?

Visit Amazon EC2 Instance Type for a list of EC2 instances available by region.

Q: How does EC2 perform maintenance?

AWS regularly performs routine hardware, software, power, and network maintenance with minimal disruption across all EC2 instance types. This is achieved by a combination of technologies and methods across the entire AWS Global infrastructure, such as live update and live migration as well as redundant and concurrently maintainable systems. Non-intrusive maintenance technologies such as live update and live migration do not require instances to be stopped or rebooted. Customers are not required to take any action prior to, during or after live migration or live update. These technologies help improve application uptime and reduce your operational effort. Amazon EC2 uses live update to deploy software to servers quickly with minimal impact to customer instances. Live update ensures that customers' workloads run on servers with software that is up-to-date with security patches, new instance features and performance improvements. Amazon EC2 uses live migration when running instances need to be moved from one server to another for hardware maintenance or to optimize placement of instances or to dynamically manage CPU resources. Amazon EC2 has been expanding the scope and coverage of non-intrusive maintenance technologies over the years so that scheduled maintenance events are a fallback option rather than the primary means of enabling routine maintenance.

Q: How do I select the right instance type?

Amazon EC2 instances are grouped into 5 families: General Purpose, Compute Optimized, Memory Optimized, Storage Optimized and Accelerated Computing instances. General Purpose Instances have memory to CPU ratios suitable for most general purpose applications and come with fixed performance or burstable performance; Compute Optimized instances have proportionally more CPU resources than memory (RAM) and are well suited for scale out compute-intensive applications and High Performance Computing (HPC) workloads; Memory

Optimized Instances offer larger memory sizes for memory-intensive applications, including database and memory caching applications; Accelerated Computing instances use hardware accelerators, or co-processors, to perform functions such as floating point number calculations, graphics processing, or data pattern matching, more efficiently than is possible in software running on CPUs; Storage Optimized Instances provide low latency, I/O capacity using SSD-based local instance storage for I/O-intensive applications, as well as dense HDD-storage instances, which provide local high storage density and sequential I/O performance for data warehousing, Hadoop and other data-intensive applications. When choosing instance types, you should consider the characteristics of your application with regards to resource utilization (i.e. CPU, Memory, Storage) and select the optimal instance family and instance size.

Q: What is an "EC2 Compute Unit" and why did you introduce it?

Transitioning to a utility computing model fundamentally changes how developers have been trained to think about CPU resources. Instead of purchasing or leasing a particular processor to use for several months or years, you are renting capacity by the hour. Because Amazon EC2 is built on commodity hardware, over time there may be several different types of physical hardware underlying EC2 instances. Our goal is to provide a consistent amount of CPU capacity no matter what the actual underlying hardware.

Amazon EC2 uses a variety of measures to provide each instance with a consistent and predictable amount of CPU capacity. In order to make it easy for developers to compare CPU capacity between different instance types, we have defined an Amazon EC2 Compute Unit. The amount of CPU that is allocated to a particular instance is expressed in terms of these EC2 Compute Units. We use several benchmarks and tests to manage the consistency and predictability of the performance from an EC2 Compute Unit. The EC2 Compute Unit (ECU) provides the relative measure of the integer processing power of an Amazon EC2 instance. Over time, we may add or substitute measures that go into the definition of an EC2 Compute Unit, if we find metrics that will give you a clearer picture of compute capacity.

Q: How does EC2 ensure consistent performance of instance types over time?

AWS conducts yearly performance benchmarking of Linux and Windows compute performance on EC2 instance types. Benchmarking results, a test suite that customers can use to conduct independent testing, and guidance on expected performance variance is available under NDA for M,C,R, T and z1d instances; please contact your sales representative to request them.

Q: What is the regional availability of Amazon EC2 instance types?

For a list of all instances and regional availability, visit Amazon EC2 Pricing.

Micro instances

Q. How much compute power do Micro instances provide?

Micro instances provide a small amount of consistent CPU resources and allow you to burst CPU capacity up to 2 ECUs when additional cycles are available. They are well suited for lower throughput applications and web sites that consume significant compute cycles periodically but very little CPU at other times for background processes, daemons, etc. Learn more about using this instance type.

Q. How does a Micro instance compare in compute power to a Standard Small instance?

At steady state, Micro instances receive a fraction of the compute resources that Small instances do. Therefore, if your application has compute-intensive or steady state needs we recommend using a Small instance (or larger, depending on your needs). However, Micro instances can periodically burst up to 2 ECUs (for short periods of time). This is double the number of ECUs available from a Standard Small instance. Therefore, if you have a relatively low throughput application or web site with an occasional need to consume significant compute cycles, we recommend using Micro instances.

Q. How can I tell if an application needs more CPU resources than a Micro instance is providing?

The CloudWatch metric for CPU utilization will report 100% utilization if the instance bursts so much that it exceeds its available CPU resources during that CloudWatch monitored minute. CloudWatch reporting 100% CPU utilization is your signal that you should consider scaling – manually or via Auto Scaling – up to a larger instance type or scale out to multiple Micro instances.

Q. Are all features of Amazon EC2 available for Micro instances?

Currently Amazon DevPay is not available for Micro instances.

Nitro Hypervisor

Q. What is the Nitro Hypervisor?

The launch of C5 instances introduced a new hypervisor for Amazon EC2, the Nitro Hypervisor. As a component of the Nitro system, the Nitro Hypervisor primarily provides CPU and memory isolation for EC2 instances. VPC networking and EBS storage resources are implemented by dedicated hardware components, Nitro Cards that are part of all current generation EC2 instance families. The Nitro Hypervisor is built on core Linux Kernel-based Virtual Machine (KVM) technology, but does not include general-purpose operating system components.

Q. How does the Nitro Hypervisor benefit customers?

The Nitro Hypervisor provides consistent performance and increased compute and memory resources for EC2 virtualized instances by removing host system software components. It allows AWS to offer larger instance sizes (like c5.18xlarge) that provide practically all of the resources from the server to customers. Previously, C3 and C4 instances each eliminated software components by moving VPC and EBS functionality to hardware designed and built by AWS. This hardware enables the Nitro Hypervisor to be very small and uninvolved in data processing tasks for networking and storage.

Q. Will all EC2 instances use the Nitro Hypervisor?

Eventually all new instance types will use the Nitro Hypervisor, but in the near term, some new instance types will use Xen depending on the requirements of the platform.

Q. Will AWS continue to invest in its Xen-based hypervisor?

Yes. As AWS expands its global cloud infrastructure, EC2's use of its Xen-based hypervisor will also continue to grow. Xen will remain a core component of EC2 instances for the foreseeable future. AWS is a founding member of the Xen Project since its establishment as a Linux Foundation Collaborative Project and remains an active participant on its Advisory Board. As AWS expands its global cloud infrastructure, EC2's Xen-based hypervisor also continues to grow. Therefore EC2's investment in Xen continues to grow, not shrink.

Q. How many EBS volumes and Elastic Network Interfaces (ENIs) can be attached to instances running on the Nitro Hypervisor?

Instances running on the Nitro Hypervisor support a maximum of 27 additional PCI devices for EBS volumes and VPC ENIs. Each EBS volume or VPC ENI uses a PCI device. For example, if you attach 3 additional network interfaces to an instance that uses the Nitro Hypervisor, you can attach up to 24 EBS volumes to that instance.

Q. Will the Nitro Hypervisor change the APIs used to interact with EC2 instances?

No, all the public facing APIs for interacting with EC2 instances that run using the Nitro Hypervisor will remain the same. For example, the "hypervisor" field of the DescribeInstances response will continue to report "xen" for all EC2 instances, even those running under the Nitro Hypervisor. This field may be removed in a future revision of the EC2 API.

Q. Which AMIs are supported on instances that use the Nitro Hypervisor?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe storage can be used with instances that run under the Nitro Hypervisor. The latest Amazon Linux AMI and Windows AMIs provided by Amazon are supported, as are the latest AMI of Ubuntu, Debian, Red Hat Enterprise Linux, SUSE Enterprise Linux, CentOS, and FreeBSD.

Q. Will I notice any difference between instances using Xen hypervisor and those using the Nitro Hypervisor?

Yes. For example, instances running under the Nitro Hypervisor boot from EBS volumes using an NVMe interface. Instances running under Xen boot from an emulated IDE hard drive, and switch to the Xen paravirtualized block device drivers.

Operating systems can identify when they are running under a hypervisor. Some software assumes that EC2 instances will run under the Xen hypervisor and rely on this detection. Operating systems will detect they are running under KVM when an instance uses the Nitro Hypervisor, so the process to identify EC2 instances should be used to identify EC2 instances that run under both hypervisors.

All the features of EC2 such as Instance Metadata Service work the same way on instances running under both Xen and the Nitro Hypervisor. The majority of applications will function the same way under both Xen and the Nitro Hypervisor as long as the operating system has the

needed support for ENA networking and NVMe storage.

Q. How are instance reboot and termination EC2 API requests implemented by the Nitro Hypervisor?

The Nitro Hypervisor signals the operating system running in the instance that it should shut down cleanly by industry standard ACPI methods. For Linux instances, this requires that acpid be installed and functioning correctly. If acpid is not functioning in the instance, termination events will be delayed by multiple minutes and will then execute as a hard reset or power off.

Q. How do EBS volumes behave when accessed by NVMe interfaces?

There are some important differences in how operating system NVMe drivers behave compared to Xen paravirtual (PV) block drivers.

First, the NVMe device names used by Linux based operating systems will be different than the parameters for EBS volume attachment requests and block device mapping entries such as /dev/xvda and /dev/xvdf. NVMe devices are enumerated by the operating system as /dev/nvme0n1, /dev/nvme1n1, and so on. The NVMe device names are not persistent mappings to volumes, therefore other methods like file system UUIDs or labels should be used when configuring the automatic mounting of file systems or other startup activities. When EBS volumes are accessed via the NVMe interface, the EBS volume ID is available via the controller serial number and the device name specified in EC2 API requests is provided by an NVMe vendor extension to the Identify Controller command. This enables backward compatible symbolic links to be created by a utility script. For more information see the EC2 documentation on device naming and NVMe based EBS volumes.

Second, by default the NVMe drivers included in most operating systems implement an I/O timeout. If an I/O does not complete in an implementation specific amount of time, usually tens of seconds, the driver will attempt to cancel the I/O, retry it, or return an error to the component that issued the I/O. The Xen PV block device interface does not time out I/O, which can result in processes that cannot be terminated if it is waiting for I/O. The Linux NVMe driver behavior can be modified by specifying a higher value for the nvme.io timeout kernel module parameter.

Third, the NVMe interface can transfer much larger amounts of data per I/O, and in some cases may be able to support more outstanding I/O requests, compared to the Xen PV block interface. This can cause higher I/O latency if very large I/Os or a large number of I/O requests are issued to volumes designed to support throughput workloads like EBS Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes. This I/O

latency is normal for throughput optimized volumes in these scenarios, but may cause I/O timeouts in NVMe drivers. The I/O timeout can be adjusted in the Linux driver by specifying a larger value for the nvme_core.io_timeout kernel module parameter.

Optimize CPUs

Q: What is Optimize CPUs?

Optimize CPUs gives you greater control of your EC2 instances on two fronts. First, you can specify a custom number of vCPUs when launching new instances to save on vCPU-based licensing costs. Second, you can disable Intel Hyper-Threading Technology (Intel HT Technology) for workloads that perform well with single-threaded CPUs, such as certain HPC applications.

Q: Why should I use Optimize CPUs feature?

You should use Optimize CPUs if:

- You are running EC2 workloads that are not compute bound and are incurring vCPU-based licensing costs. By launching instances with a custom number of vCPUs you may be able to optimize your licensing spend.
- You are running workloads that will benefit from disabling hyper-threading on EC2 instances.

Q: How will the CPU optimized instances be priced?

CPU optimized instances will be priced the same as equivalent full-sized instances.

Q: How will my application performance change when using Optimize CPUs on EC2?

Your application performance change with Optimize CPUs will be largely dependent on the workloads you are running on EC2. We encourage you to benchmark your application performance with Optimize CPUs to arrive at the right number of vCPUs and optimal hyper-threading behavior for your application.

Q: Can I use Optimize CPUs on EC2 Bare Metal instance types (such as i3.metal)?

No. You can use Optimize CPUs with only virtualized EC2 instances.

Q: How can I get started with using Optimize CPUs for EC2 Instances?

For more information on how to get started with Optimize CPUs and supported instance types, please visit the Optimize CPUs documentation page here.

Workloads

Amazon EC2 running IBM | Amazon EC2 running Microsoft Windows and other third-party software | macOS workloads

Amazon EC2 running IBM

Q: How am I billed for my use of Amazon EC2 running IBM?

You pay only for what you use and there is no minimum fee. Pricing is per instance-hour consumed for each instance type. Partial instance-hours consumed are billed as full hours. Data transfer for Amazon EC2 running IBM is billed and tiered separately from Amazon EC2. There is no Data Transfer charge between two Amazon Web Services within the same Region (for example, between Amazon EC2 US West and another AWS service in the US West). Data transferred between AWS services in different regions will be charged as Internet Data Transfer on both sides of the transfer.

For Amazon EC2 running IBM pricing information, please visit the pricing section on the Amazon EC2 running IBM detail page.

Q: Can I use Amazon DevPay with Amazon EC2 running IBM?

No, you cannot use DevPay to bundle products on top of Amazon EC2 running IBM at this time.

Amazon EC2 running Microsoft Windows and other third-party software

Q: Can I use my existing Windows Server license with EC2?

Yes you can. After you've imported your own Windows Server machine images using the ImportImage tool, you can launch instances from these machine images on EC2 Dedicated Hosts and effectively manage instances and report usage. Microsoft typically requires that you track usage of your licenses against physical resources such as sockets and cores and Dedicated Hosts helps you to do this. Visit the Dedicated Hosts detail page for more information on how to use your own Windows Server licenses on Amazon EC2 Dedicated Hosts.

Q: What software licenses can I bring to the Windows environment?

Specific software license terms vary from vendor to vendor. Therefore, we recommend that you check the licensing terms of your software vendor to determine if your existing licenses are authorized for use in Amazon EC2.

macOS workloads

Q: What are Amazon EC2 Mac instances?

Amazon EC2 Mac instances allow customers to run on-demand macOS workloads in the cloud for the first time, extending the flexibility, scalability, and cost benefits of AWS to all Apple developers. With EC2 Mac instances, developers creating apps for iPhone, iPad, Mac, Apple Watch, Apple TV, and Safari can provision and access macOS environments within minutes, dynamically scale capacity as needed, and benefit from AWS's pay-as-you-go pricing.

Q: What workloads should you run on EC2 Mac instances?

Amazon EC2 Mac instances are designed to build, test, sign, and publish applications for Apple platforms such as iOS, iPadOS, watchOS, tvOS, macOS, and Safari. Customers such as Pinterest, Intuit, FlipBoard, Twitch, and Goldman Sachs have seen up to 75% better build performance,

up to 80% lower build failure rates, and up to 5x the number of parallel builds compared to running macOS on premises.

Q: What are EC2 x86 Mac instances?

x86-based EC2 Mac instances are built on Apple Mac mini computers featuring Intel Core i7 processors and are powered by the AWS Nitro System. They offer customers a choice of macOS Mojave (10.14), macOS Catalina (10.15), macOS Big Sur (11), and macOS Monterey (12) as Amazon Machine Images (AMIs). x86-based EC2 Instances are available in 12 Regions: US East (Ohio, N. Virginia), US West (Oregon), Europe (Stockholm, Frankfurt, Ireland, London), and Asia Pacific (Mumbai, Seoul, Singapore, Sydney, Tokyo). Learn more and get started with x86-based EC2 Mac instances here.

Q: What are EC2 M1 Mac instances?

EC2 M1 Mac instances are built on Apple M1 Mac mini computers and are powered by the AWS Nitro System. They deliver up to 60 percent better price performance over x86-based EC2 Mac instances for iOS and macOS application build workloads. EC2 M1 Mac instances enable ARM64 macOS environments for the first time in AWS, and support macOS Big Sur (11) and macOS Monterey (12) as Amazon Machine Images (AMIs). EC2 M1 Mac instances are available in 4 Regions: US East (N. Virginia), US West (Oregon), Europe (Ireland), and Asia Pacific (Singapore). Learn more and get started with EC2 M1 Mac instances here.

Q: What are EC2 M2 Mac instances?

EC2 M2 Mac instances are built on Apple M2 Mac mini computers and powered by the AWS Nitro System. They are up to 10% more performant than EC2 M1 Mac instances for iOS and macOS application build workloads. EC2 M2 Mac instances enable ARM64 macOS environments on AWS and support macOS Ventura (version 13.2 and later) as Amazon Machine Images (AMIs). EC2 M2 Mac instances are available in 5 Regions: US East (N. Virginia, Ohio), US West (Oregon), Europe (Frankfurt), and Asia Pacific (Sydney). Learn more and get started with EC2 M2 Mac instances here

Q: What are EC2 M2 Pro Mac instances?

EC2 M2 Pro Mac instances are built on Apple M2 Pro Mac mini computers and powered by the AWS Nitro System. They are up to 35% more

performant than EC2 M1 Mac instances for iOS and macOS application build workloads. EC2 M2 Pro Mac instances enable ARM64 macOS environments on AWS and support macOS Ventura (version 13.2 and later) as Amazon Machine Images (AMIs). EC2 M2 Pro Mac instances are available in 4 Regions: US East (N. Virginia, Ohio), US West (Oregon), and Asia Pacific (Sydney). Learn more and get started with EC2 M2 Pro Mac instances here.

Q: What pricing models are available for EC2 Mac instances?

Amazon EC2 Mac instances are available as Dedicated Hosts through both On-Demand and Savings Plans pricing models. The Dedicated Host is the unit of billing for EC2 Mac instances. Billing is per second, with a 24-hour minimum allocation period for the Dedicated Host to comply with the Apple macOS Software License Agreement. At the end of the 24-hour minimum allocation period, the host can be released at any time with no further commitment. Both Compute and Instance Savings Plans are available for EC2 Mac instances and offer up to 44 percent off On-Demand pricing. Visit the Dedicated Host pricing page for more information. (Note: Please select “Dedicated Host” tenancy and “Linux” operating system to view details.) You can also access EC2 Mac instances pricing on the AWS Pricing Calculator for Dedicated Hosts.

Q: How do you release a Dedicated Host?

The minimum allocation period for an EC2 Mac instance Dedicated Host is 24 hours. After the allocation period has exceeded 24 hours, first stop or terminate the instance running on the host, then release the host using the aws ec2 release-hosts CLI command or the AWS Management Console.

Q: Can you share EC2 Mac Dedicated Hosts with other AWS accounts in your organization?

Yes. You can share EC2 Mac Dedicated Hosts with AWS accounts inside your AWS organization, an organizational unit inside your AWS organization, or your entire AWS organization via AWS Resource Access Manager. For more information, please refer to the AWS Resource Access Manager documentation.

Q: How many EC2 Mac instances can you run on an EC2 Mac Dedicated Host?

EC2 Mac instances leverage the full power of the underlying Mac mini hardware. You can run 1 EC2 Mac instance on each EC2 Mac Dedicated Host.

Q: Can you update the EFI NVRAM variables on an EC2 Mac instance?

Yes, you can update certain EFI NVRAM variables on an EC2 Mac instance that will persist across reboots. However, EFI NVRAM variables will be reset if the instance is stopped or terminated. Please see the EC2 Mac instances documentation for more information.

Q: Can you use FileVault to encrypt the Amazon Elastic Block Store (Amazon EBS) boot volume on EC2 Mac instances?

FileVault requires a login before booting into macOS and before remote access can be enabled. If FileVault is enabled, you will lose access to your data on the boot volume at instance reboot, stop, or terminate. We strongly recommend you do not enable FileVault. Instead, we recommend using Amazon EBS encryption for both boot and data EBS volumes on EC2 Mac instances.

Q: Can you access to the microphone input or audio output on an EC2 Mac instance?

There is no access to the microphone input on an EC2 Mac instance. The built-in Apple Remote Desktop VNC server does not support audio output. Third party remote desktop software, such as Teradici CAS, supports remote audio on macOS.

Q: What macOS-based Amazon Machine Images (AMIs) are available for EC2 Mac instances?

EC2 Mac instances use physical Mac mini hardware to run macOS. Apple hardware only supports the macOS version shipped with the hardware (or later). x86-based EC2 Mac instances use the 2018 Intel Core i7 Mac mini, which means macOS Mojave (10.14.x) is as 'far back' as you can go, since the 2018 Mac mini shipped with Mojave. EC2 M1 Mac instances use 2020 M1 Mac mini, which shipped with macOS Big Sur (11.x). EC2 M2 and M2 Pro Mac instances use the 2023 M2 and M2 Pro Mac Minis respectively, which shipped with macOS Ventura (13.2). To see which latest versions of macOS are available as EC2 Mac AMIs, please visit the documentation.

Q: How can you run older versions of macOS on EC2 Mac instances?

EC2 Mac instances are bare metal instances and do not use the Nitro hypervisor. You can install and run a type-2 virtualization layer on x86-based EC2 Mac instances to get access to macOS High Sierra, Sierra, or older macOS versions. On EC2 M1 Mac instances, as macOS Big Sur is the first macOS version to support Apple Silicon, older macOS versions will not run even under virtualization.

Q: How can I run beta or preview versions of macOS on EC2 Mac instances?

Installation of beta or preview macOS versions is only available on Apple Silicon-based EC2 Mac Instances . Amazon EC2 doesn't qualify beta or preview macOS versions and doesn't ensure instances will remain functional after an update to a pre-production macOS version.

Q: How can you use EC2 user data with EC2 Mac instances?

As with EC2 Linux and Windows instances, you can pass custom user data to EC2 Mac instances. Instead of using cloud-init, EC2 Mac instances use an open-source launch daemon: ec2-macos-init. You can pass this data into the EC2 Launch Wizard as plain-text, as a file, or as base64-encoded-text.

Q: How do you install Xcode on an EC2 Mac instance?

AWS provides base macOS AMIs without any prior Xcode IDE installation. You can install Xcode (and accept the EULA) just like you would on any other macOS system. You can install the latest Xcode IDE from the App Store, or earlier Xcode versions from the Apple Developer website. Once you have Xcode installed, we recommend creating a snapshot of your AMI for future use.

Q: What is the release cadence of macOS AMIs?

We make new macOS AMIs available on a best effort basis. You can subscribe to SNS notifications for updates. We are targeting 30-60 days after a macOS minor version update and 90-120 days after a macOS major version update to release official macOS AMIs.

Q: What agents and packages are included in EC2 macOS AMIs?

The following agents and packages are included by default in EC2 macOS AMIs:

- ENA Driver for macOS
- AWS CLI
- EC2-macos-init
- Amazon CloudWatch Agent
- Chrony
- Homebrew
- AWS Systems Manger Agent

Q: Can you update the agents and packages included in macOS AMIs?

There is a public GitHub repository of the Homebrew tap for all agents and packages added to the base macOS image. You can use Homebrew to install the latest versions of agents and packages on macOS instances.

Q: Can you apply OS and software updates to your Mac instances directly from Apple Update Servers?

Automatic macOS software updates are disabled on EC2 Mac instances. We recommend using our officially vended macOS AMIs to launch the version of macOS you need. On x86-based and all Apple Silicon EC2 Mac instances, you can update the version of macOS via the Software Update preferences pane, or via the software update CLI command. On both EC2 Mac instances, you can install and update applications and any other user-space software.

Q: How do you connect to an EC2 Mac instance over SSH?

After launching your instance and receiving an instance id, you can use the following command to poll the instance and determine when it is ready for SSH access. Connecting over SSH to EC2 Mac instances follows the same process as connecting to other EC2 instances, such as those running Linux or Windows. To support connecting to your instance using SSH, launch the instance using a key pair and a security group that allows SSH access. Provide the .pem file for the key pair when you connect to the instance. For more information, please see the documentation.

Q: How do you connect to an EC2 Mac instance over VNC?

macOS has built-in Screen Sharing functionality that is disabled by default, but can be enabled and used to connect to a Graphical (Desktop) session of your EC2 Mac instance. For more information on how to enable the built-in Screen Sharing, please see the documentation.

Q: How do you connect to an EC2 Mac instance using AWS Systems Manager Session Manager?

You can connect to your EC2 Mac instances with AWS Systems Manager Session Manager (SSM). Session Manager is a fully managed AWS Systems Manager feature that provides secure and auditable instance management. It removes the need to keep open inbound ports, maintain bastion hosts, or manage SSH keys. The SSM Agent is pre-installed by default on all EC2 macOS AMIs. For more information, please see this blog.

Q: How many Amazon EBS volumes and Elastic Network Interfaces (ENIs) are supported by EC2 Mac instances?

x86-based EC2 Mac instances support 16 EBS volumes and 8 ENI attachments, and EC2 M1 Mac instances support up to 10 EBS volumes and 8 ENI attachments.

Q: Do EC2 Mac instances support EBS?

EC2 Mac instances are EBS optimized by default and offer up to 8 Gbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes.

Q: Do EC2 Mac instances support booting from local storage?

EC2 Mac instances can only boot from EBS-backed macOS AMIs. The internal SSD of the Mac mini is present in Disk Utility, but is not bootable.

Q: Do EC2 Mac instances support Amazon FSx?

Yes. EC2 Mac instances support FSx using the SMB protocol. You will need to enroll the EC2 Mac instance into a supported directory service (such as Active Directory or the AWS Directory Service) to enable FSx on EC2 Mac instances. For more information on FSx, visit the product page.

Q: Do EC2 Mac instances support Amazon Elastic File System (Amazon EFS)?

Yes, EC2 Mac instances support EFS over the NFSv4 protocol. For more information on EFS, visit the product page.

Nitro System support for previous generation

Q: What is Nitro System support for Older Generation instances?

The AWS Nitro System now will provide its modern hardware and software components for previous generation EC2 instances to extend the length of service beyond the typical lifetime of the underlying hardware. With the Nitro System support, customers can continue running their workloads and applications on the instance families they were built on.

Q: Which previous-generation instances will receive Nitro System support and within what time frame?

We have enabled Nitro system support for Amazon EC2 C1, M1, M2, C3, M3, R3, C4, and M4 instances. Customers of these instances will receive a maintenance notification of migration to Nitro System. We will add support for additional instance types in 2023.

Q: What actions do I need to take to migrate my existing previous generation instances?

Customers do not need to take any action for migrating active previous generation instances running on older generation hardware. For instances that are on older generation hardware, each customer account ID mapped to instance(s) will receive an email notification 2 weeks prior to the scheduled maintenance.

Similar to our typical maintenance events, customers will have the option to reschedule their maintenance as many times as needed within 2 additional weeks from the original scheduled maintenance time.

Q: What will happen if instance is stopped and started before or during the scheduled maintenance window?

Stop/Start of an instance during the scheduled maintenance window will migrate the instance to a new host and the instance will not have to undergo the scheduled maintenance. The stop/start may result in migration of the customer instance to be supported by the AWS Nitro System. Please note that the data on any local instance-store volumes will not be preserved when you stop and start your instance. Click here for more information about stop/start.

Q: What will happen to my instance during this maintenance event?

We will work in conjunction with the customer as a part of our standard AWS maintenance process. Several AWS teams have already migrated and are running previous generation instances on Nitro hardware. During maintenance, the instance will be rebooted which can take up to 30 minutes depending upon the instance size and attributes. For example: Instances with local disk take longer to migrate than instances without local disk. After the reboot, your instance retains its IP address, DNS name, and any data on local instance-store volumes.

Q: Do I need to rebuild/recertify workloads to run on previous generation instances migrated to AWS Nitro System?

No, customers don't need to rebuild/recertify workloads on previous generation instances migrated to AWS Nitro System.

Q: Will there be any changes to my instance specifications once migrated to AWS Nitro System?

There will be no change to instance specification of previous generation instances when instances are migrated to AWS Nitro System.

Q: Will all features and AMIs on my previous generation instances be supported as a part of this migration?

Yes, all existing features and AMIs supported on previous generation instances will be supported as we migrate these instances to AWS Nitro System.