# Amazon EC2 FAQs - Amazon Web Services

aws.amazon.com (https://aws.amazon.com/ec2/faqs/)

## General

Longer EC2, EBS, and Storage Gateway resource IDs | Overview | Service level agreement (SLA)

### Longer EC2, EBS, and Storage Gateway resource IDs

Q: What is changing?

Starting July 2018, all newly created EC2 resources will receive longer format IDs. The new format will only apply to newly created resources; your existing resources won't be affected. Instances and volumes already use this ID format. Through the end of June 2018, customers will have the ability to opt-in to use longer IDs. During this time, you can choose which ID format resources are assigned and update your management tools and scripts to add support for the longer format. Please visit this (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/resource-ids.html) documentation for instructions.

Q: Why is this necessary?

Given how fast AWS continues to grow, we will start to run low on IDs for certain resources in 2018. In order to enable the long-term, uninterrupted creation of new resources, we need to introduce a longer ID format. All Amazon EC2 resource IDs will change to the longer format in July 2018.

Q: I already opted in for longer IDs last year. Why do I need to opt-in again?

In 2016, we moved to the longer ID format for Amazon EC2 instances, reservations, volumes, and snapshots only. This opt-in changes the ID format for all remaining EC2 resource types

Q: What will the new identifier format look like?

The new identifier format will follow the pattern of the current identifier format, but it will be longer. The new format will be -<17 characters>, e.g. "vpc-1234567890abcdef0" for VPCs or "subnet-1234567890abcdef0" for subnets.

Q: Which IDs are changing?

- bundle

- conversion-task
- customer-gateway
- dhcp-options
- elastic-ip-allocation
- elastic-ip-association
- export-task
- flow-log
- image
- import-task
- internet-gateway
- network-acl
- network-acl-association
- network-interface
- network-interface-attachment
- prefix-list
- route-table
- route-table-association
- security-group
- subnet
- subnet-cidr-block-association
- vpc
- vpc-cidr-block-assocation
- vpc-endpoint
- vpc-peering-connection
- vpn-connection
- vpn-gateway

Q: How does this impact me?

There is a good chance that you won't need to make any system changes to handle the new format. If you only use the console to manage AWS resources, you might not be impacted at all, but you should still update your settings to use the longer ID format as soon as possible. If you interact with AWS resources via APIs, SDKs, or the AWS CLI, you might be impacted, depending on whether your software makes assumptions about the ID format when validating or persisting resource IDs. If this is the case, you might need to update your systems to handle the new format.
Some failure modes could include:

- If your systems use regular expressions to validate the ID format, you might error if a longer format is encountered.
- If there are expectations about the ID length in your database schemas, you might be unable to store a longer ID.

Q: Will this affect existing resources?

No. Only resources that are created after you opt-in to the longer format will be affected. Once a resource has been assigned an ID (long or short), that ID will never change. Each ID is unique and will never be reused. Any resource created with the old ID format will always retain its shorter ID. Any resource created with the new format will retain its longer ID, even if you opt back out.

Q: When will this happen?

Through the end of June 2018, longer IDs will be available for opt-in via APIs and the EC2 Console. All accounts can opt-in and out of longer IDs as needed for testing. Starting on July 1, 2018, the option to switch formats will no longer be available, and newly created EC2 resources to receive longer IDs. All regions launching in July 2018 and onward will only support longer IDs.

Q: Why is there an opt-in period?

We want to give you as much time as possible to test your systems with the new format. This transition time offers maximum flexibility to test and update your systems incrementally and will help minimize interrupts as you add support for the new format, if necessary.

Q: How do I opt in and out of receiving longer IDs?

Throughout the transition period (Now through the end of June 2018), you can opt to receive longer or shorter IDs by using the APIs or the EC2 Console. Instructions are provided in this (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/resource-ids.html) documentation.

Q: What will happen if I take no action?

If you do not opt-in to the new format during the transition period, you will be automatically begin receiving the longer format IDs after July 1, 2018. We do not recommend this approach. It is better to add support for the new format during the transition window, which offers the opportunity for controlled testing.

Q: What if I prefer to keep receiving the shorter ID format after the end of June 2018?

This is not possible regardless of your user settings specified.

Q: When will the longer IDs' final transition happen?

In July 2018, your newly created resources will start to receive longer IDs. You can check the scheduled transition date for your each region by using the AWS CLI describe-id-format (https://docs.aws.amazon.com/cli/latest/reference/ec2/describe-id-format.html).

Q: If I opt in to longer IDs and then opt back out during the transition period, what will happen to resources that were

created with longer IDs?

Once a resource has been assigned an ID it will not change, so resources that are created with longer IDs will retain the longer IDs regardless of later actions. If you opt in to the longer format, create resources, and then opt out, you will see a mix of long and short resource IDs, even after opting out. The only way to get rid of long IDs will be to delete or terminate the respective resources. For this reason, exercise caution and avoid creating critical resources with the new format until you have tested your tools and automation.

Q: What should I do if my systems are not working as expected before the transition period ends?

If your systems are not working as expected during the transition period, you can temporarily opt out of longer format IDs and remediate your systems, however your account will automatically be transitioned back to using longer IDs after the end of June 2018. Regardless of your account settings, all new resources will receive the longer format IDs, so it is important for you to test your systems with longer format IDs before the transition period ends. By testing and opting in earlier, you give yourself valuable time to make modifications to your resources with short resource IDs and you minimize the risk of any impact to your systems.

Q: What will happen if I launch resources in multiple regions during the transition period?

Your resources' ID length will depend upon the region you launch your resources. If the region has already transitioned to using longer IDs, resources launched in that region will have longer format IDs; if not, they will have shorter resource IDs. Therefore, during the transition window, you may see a mix of shorter and longer resource IDs.

Q: If AWS adds new regions during the transition period, will new regions support longer IDs?

Yes. All new regions launching after July 2018 will issue longer format IDs by default for both new and existing accounts.

Q: What will be the default ID type for new accounts?

Accounts created on March 15, 2018 or later will be configured to receive the longer ID format by default in every AWS region except AWS GovCloud (US). If you are a new customer, this will make the transition to longer IDs really simple. If you would like your new account to assign the shorter ID format to your resources, then simply reconfigure your account for shorter IDs as described above. This workflow will be necessary until you are ready for your accounts to receive longer IDs.

Q: Will I need to upgrade to a new version of the AWS SDKs or CLI?

The following AWS CLI and SDKs are fully compatible with longer IDs: PHP v2.8.27+, PHP v3.15.0+, AWS CLI v1.10.2+, Boto3v1.2.1+, Botocorev1.3.24+, PHP v1, Boto v1, Boto v2, Ruby v1, Ruby v2, JavaScript, Java, .NET, AWS Tools for Windows PowerShell, and Go.

Q: How can I test my systems with longer IDs?

Amazon Machine Images (AMIs) with longer format IDs have been published for testing purposes. Instruction on how to access these AMIs are provided here (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/resource-ids.html).

## Overview

Q: What is Amazon Elastic Compute Cloud (Amazon EC2)?

Amazon Elastic Compute Cloud (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. Amazon EC2’s 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, click the “Sign up for This Web Service” button on the Amazon EC2 detail page. You must have an Amazon Web Services 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 sign-up process. After signing up, please refer to the Amazon EC2 documentation (http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=87), 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 number 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 enables any developer to leverage Amazon’s own benefits of massive scale with no up-front 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 20 On-Demand instances, complete the Amazon EC2 instance request form (https://aws.amazon.com/contact-us/ec2-request/).

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 up to a total of 20 On-Demand instances across the instance family, purchasing 20 Reserved Instances, and requesting Spot Instances per your dynamic Spot limit (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-limits.html) per region. New AWS accounts may start with limits that are lower than the limits described here. Certain instance types are further limited per region as follows:

Instance Type On-Demand Limit Reserved Limit Spot Limit m5.large 20 20 Dynamic Spot Limit m5.xlarge 20 20 Dynamic Spot Limit m5.2xlarge 20 20 Dyanmic Spot Limit m5.4xlarge 10 20 Dynamic Spot Limit m5.12xlarge 5 20 Dynamic Spot Limit m5.24xlarge 5 20 Dynamic Spot Limit m4.4xlarge 10 20 Dynamic Spot Limit m4.10xlarge 5 20 Dynamic Spot Limit m4.16xlarge 5 20 Dynamic Spot Limit c5.large 20 20 Dynamic Spot Limit c5.xlarge 20 20 Dynamic Spot Limit c5.2xlarge 20 20 Dynamic Spot Limit c5.4xlarge 10 20 Dynamic Spot Limit c5.9xlarge 5 20 Dynamic Spot Limit c5.18xlarge 5 20 Dynamic Spot Limit c4.4xlarge 10 20 Dynamic Spot Limit c4.8xlarge 5 20 Dynamic Spot Limit hs1.8xlarge 2 20 Not offered cr1.8xlarge 2 20 Dynamic Spot Limit p3.2xlarge 1 20 Dynamic Spot Limit p3.8xlarge 1 20 Dynamic Spot Limit p3.16xlarge 1 20 Dynamic Spot Limit p2.xlarge 1 20 Dynamic Spot Limit p2.8xlarge 1 20 Dynamic Spot Limit p2.16xlarge 1 20 Dynamic Spot Limit g3.4xlarge 1 20 Dynamic Spot Limit g3.8xlarge 1 20 Dynamic Spot Limit g3.16xlarge 1 20 Dynamic Spot Limit r4.large 20 20 Dynamic Spot Limit r4.xlarge 20 20 Dynamic Spot Limit r4.2xlarge 20 20 Dynamic Spot Limit r4.4xlarge 10 20 Dynamic Spot Limit r4.8xlarge 5 20 Dynamic Spot Limit r4.16xlarge 1 20 Dynamic Spot Limit r3.4xlarge 10 20 Dynamic Spot Limit r3.8xlarge 5 20 Dynamic Spot Limit h1.8xlarge 10 20 Dynamic Spot Limit h1.16xlarge 5 20 Dynamic Spot Limit i3.large 2 20 Dynamic Spot limit i3.xlarge 2 20 Dynamic Spot limit i3.2xlarge 2 20 Dynamic Spot limit i3.4xlarge 2 20 Dynamic Spot limit i3.8xlarge 2 20 Dynamic Spot limit i3.8xlarge 2 20 Dynamic Spot limit i3.16xlarge 2 20 Dynamic Spot limit i2.2xlarge 8 20 Dynamic Spot Limit i2.4xlarge 4 20 Dynamic Spot Limit i2.8xlarge 2 20 Dynamic Spot Limit d2.4xlarge 10 20 Dynamic Spot Limit d2.8xlarge 5 20 Dynamic Spot Limit t2.nano 20 20 Dynamic Spot Limit t2.micro 20 20 Dynamic Spot Limit t2.small 20 20 Dynamic Spot Limit t2.medium 20 20 Dynamic Spot Limit t2.large 20 20 Dynamic Spot Limit t2.xlarge 20 20 Dynamic Spot Limit t2.2xlarge 20 20 Dynamic Spot Limit All Other Instance Types 20 20 Dynamic Spot Limit

*Note that cc2.8xlarge, hs1.8xlarge, cr1.8xlarge, G2, D2, and I2 instances are not available in all regions.*

If you need more instances, complete the Amazon EC2 instance request form (https://aws.amazon.com/support/createCase?type=service_limit_increase&serviceLimitIncreaseType=ec2-instances) with your use case and your instance 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 (https://portal.aws.amazon.com/gp/aws/html-forms-controller/contactus/ec2-email-limit-rdns-request).

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, Fedora, 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, up-front 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.

### 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 a SLA Service Credit?

You are eligible for a 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 99.95% 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, please see http://aws.amazon.com/ec2/sla/

## Instance types

Accelerated Computing instances | Compute 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?

Accelerated Computing instance family is a family of instances which 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 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: 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 includes 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, academia, among many others.

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

P3 instance 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, meteorology, among many others. Research indicates that over 70% of the most popular HPC applications provide built-in support for GPUs.

Q: Will P3 instances support EC2 Classic networking and Amazon VPC?

P3 instances will support VPC only.

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 (https://aws.amazon.com/marketplace/search/results/?searchTerms=GPU) which 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 (https://forums.aws.amazon.com/forum.jspa?forumID=30#).

Q. Does the use of G2 and G3 instances require third-party licenses?

Aside from the NVIDIA drivers and GRID SDK, the use of G2 and G3 instances does not necessarily require any third-party licenses. However, you are responsible for determining whether your content or technology used on G2 and 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 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 a FPGA expert to use an F1 instance?

AWS customers subscribing to a 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 a 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 publically 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 and f1.16 xlarge.

### 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. 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/processor_state_control.html).

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

C5 instances: C5 instances are the latest generation of EC2 Compute Optimized instances. C5 instances are based on Intel Xeon Platinum processors, part of the Intel Xeon Scalable (codenamed Skylake-SP) processor family, and are available in 6 sizes and offer up to 72 vCPUs and 144 GiB memory. C5 instances deliver 25% improvement in price/performance compared to C4 instances.

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. Should I move my workloads from C3 or C4 instances to C5 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 operating systems/AMIs are supported on C5 Instances?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe-based storage can be used with C5 instances. The following AMIs are supported on C5:

- Amazon Linux 2014.03 or newer
- Ubuntu 14.04 or newer
- SUSE Linux Enterprise Server 12 or newer
- Red Hat Enterprise Linux 7.4 or newer
- CentOS 7 or newer
- Windows Server 2008 R2
- Windows Server 2012
- Windows Server 2012 R2
- Windows Server 2016
- FreeBSD 11.1-RELEASE

For optimal local NVMe-based SSD storage performance on C5d, Linux kernel version 4.9+ is recommended.

Q. What are the storage options available to C5 customers?

C5 instances use EBS volumes for storage, are EBS-optimized by default, and offer up to 9 Gbps throughput to both encrypted and unencrypted EBS volumes. C5 instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces. NVMe is an efficient and scalable storage interface commonly used for flash based SSDs such as local NVMe storage provided with I3 and I3en instances. Though the NVMe interface may provide lower latency compared to Xen paravirtualized block devices, when used to access EBS volumes the volume type, size, and provisioned IOPS (if applicable) will determine the overall latency and throughput characteristics of the volume. When NVMe is used to provide EBS volumes, they are attached and detached by PCI hotplug.

Q. What network interface is supported on C5 instances?

C5 instances use the Elastic Network Adapter (ENA) for networking and enable Enhanced Networking by default. With ENA, C5 instances can utilize up to 25 Gbps of network bandwidth.

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. How many EBS volumes can be attached to C5 instances?

C5 instances support a maximum for 27 EBS volumes for all Operating systems. The limit is shared with ENI attachments which can be found here http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-eni.html. For example: since every instance has at least 1 ENI, if you have 3 additional ENI attachments on the c4.2xlarge, you can attach 24 EBS volumes to that instance.

Q. What is the underlying hypervisor on C5 instances?

C5 instances use a new EC2 hypervisor that is based on core KVM technology.

Q: Why does the total memory reported by the operating system not match the advertised memory of the C5 instance type?

In C5, portions of the total memory for an instance are reserved from use by the Operating System including areas used by the virtual BIOS for things like ACPI tables and for devices like the virtual video RAM.

### General Purpose instances

Q: What are Amazon EC2 A1 instances?

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

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

AWS Graviton processors are a new line of processors that 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 customer workloads that are supported by the extensive Arm ecosystem and can fit within the available memory footprint. A1 instances are ideal for scale-out applications such as web servers, containerized microservices, caching fleets, and distributed data stores. These instances will also appeal to developers, enthusiasts, and educators across the Arm developer community. Most applications that make use of open source software like Apache HTTP Server, Perl, PHP, Ruby, Python, NodeJS, and Java easily run on multiple processor architectures due to the support of Linux based operating systems. We encourage customers running such applications to give A1 instances a try.

Applications that require higher compute and network performance, require higher memory, or have dependencies on x86 architecture will be better suited for existing instances like the M5, C5, or R5 instances. Applications with variable CPU usage that experience occasional spikes in demand will get the most cost savings from the burstable performance T3 instances.

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 A1 instances?

You will want to ensure that you use the “arm64” AMIs with the A1 instances. x86 AMIs are not compatible with A1 instances.

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 (https://aws.amazon.com/ec2/nitro/), a combination of dedicated hardware and Nitro hypervisor.

Q: Why does the total memory reported by Linux not match the advertised memory of the A1 instance type?

In A1 instances, portions of the total memory for an instance are reserved from use by the operating system including areas used by the virtual UEFI for things like ACPI tables.

Q: What are the key use cases for Amazon EC2 M5 Instances?

M5 instances offer a good choice for running development and test environments, web, mobile and gaming applications, analytics applications, and business critical applications including ERP, HR, CRM, and collaboration apps. Customers who are interested in running their data intensive workloads (e.g. HPC, or SOLR clusters) on instances with a higher memory footprint will also find M5 to be a good fit. Workloads that heavily use single and double precision floating point operations and vector processing such as video processing workloads and need higher memory can benefit substantially from the AVX-512 instructions that M5 supports.

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 (aka Skylake), 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: How does support for Intel AVX-512 benefit EC2 M5 and M5d Instance customers?

Intel Advanced Vector Extension 512 (AVX-512) is a set of new CPU instructions available on the latest Intel Xeon Scalable processor family, 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 EC2 M5 and M5d customers can deploy more secure data and services into distributed environments without compromising performance.

Q: What are the various processor options available to M5 customers?

The M5 and M5d instance types use a 3.1 GHz Intel Xeon Platinum 8000 series processor. The M5a and M5ad instance types use a 2.5 GHz AMD EPYC 7000 series processor.

Q: What are the various storage options available to M5 customers?

The M5 and M5a instance types leverage EBS volumes for storage. The M5d and M5ad instance types support up to 3.6TB (4 x 900GB) of local NVMe storage.

Q: When should I use the different M5 instance types?

Customers should consider using the M5a and M5ad instance types if they are looking to save money on price when their workloads do not fully utilize the compute resources of their chosen instance, resulting in them paying for performance that they don’t actually need. For workloads that require the highest processor performance or high floating-point performance capabilities, including vectorized computing with AVX-512 instructions, then we suggest you use the M5 or M5d instance types.

Q: Which network interface is supported on M5 instances?

M5, M5a, M5d, and M5ad instances support only ENA based Enhanced Networking and will not support netback. With ENA, M5 and M5d instances can deliver up to 25 Gbps of network bandwidth between instances and the M5a and M5ad instance types can support up to 20Gbps of network bandwidth between instances.

Q. Which operating systems/AMIs are supported on M5 Instances?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe-based storage can be used with M5 instances. The following AMIs are supported on M5, M5a, M5ad, and M5d:

- Amazon Linux 2014.03 or newer
- Ubuntu 14.04 or newer
- SUSE Linux Enterprise Server 12 or newer
- Red Hat Enterprise Linux 7.4 or newer
- CentOS 7 or newer
- Windows Server 2008 R2
- Windows Server 2012
- Windows Server 2012 R2
- Windows Server 2016
- FreeBSD 11.1-RELEASE

For optimal local NVMe-based SSD storage performance on M5d, Linux kernel version 4.9+ is recommended.

Q. What interface connects EBS storage to my M5 insatnces?
M5, M5a, M5ad, and M5d instances use EBS volumes for storage, are EBS-optimized by default, and offer up to 10 Gbps throughput to both encrypted and unencrypted EBS volumes. M5 instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces. NVMe is an efficient and scalable storage interface commonly used for flash based SSDs such as local NVMe storage provided with I3 and I3en instances. Though the NVMe interface may provide lower latency compared to Xen paravirtualized block devices, when used to access EBS volumes the volume type, size, and provisioned IOPS (if applicable) will determine the overall latency and throughput characteristics of the volume. When NVMe is used to provide EBS volumes, they are attached and detached by PCI hotplug.
Q. How many EBS volumes can be attached to M5 instances?
M5 and M5a instances support a maximum for 27 EBS volumes for all Operating systems. The limit is shared with ENI attachments which can be found here http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-eni.html. For example: since every instance has at least 1 ENI, if you have 3 additional ENI attachments on the m5.2xlarge, you can attach 24 EBS volumes to that instance.
Q. What is the underlying hypervisor on M5 instances?
M5, M5a, M5ad, and M5d instances use a new lightweight Nitro (https://aws.amazon.com/ec2/nitro/) Hypervisor that is based on core KVM technology.
Q: Why does the total memory reported by the operating system not match the advertised memory of the M5 instance type?
In M5, M5a, M5ad, and M5d, portions of the total memory for an instance are reserved from use by the operating system including areas used by the virtual BIOS for things like ACPI tables and for devices like the virtual video RAM.
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.large, 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 (e.g. 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 (https://aws.amazon.com/marketplace/search?page=1&instance_types=t2.nano&filters=instance_types). Windows customers who do not need the GUI can use the Microsoft Windows Server 2012 R2 Core AMI (https://aws.amazon.com/marketplace/pp/B00KQOWEPO/ref=srh_res_product_title?ie=UTF8&sr=0-2&qid=1448487264646).

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 instance family 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.

### High Memory instances

Q. What are EC2 High Memory instances?

Amazon EC2 High Memory instances offer 6 TB, 9 TB, or 12 TB 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 are the first Amazon EC2 instances powered by an 8-socket platform with latest generation Intel® Xeon® Platinum 8176M (Skylake) processors that are optimized for mission-critical enterprise workloads. EC2 High Memory instances deliver high networking throughput and low-latency with 25 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.

Q. Which instances are available within High Memory instance category?

Three High Memory instances are available. u-6tb1.metal offers 6 TB memory; u-9tb1.metal offers 9 TB memory; and u-12tb1.metal offers 12 TB memory. Each High Memory instance offers 448 logical processors, where each logical processor is a hyperthread on the 8-socket platform with total of 224 CPU cores.

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 14 Gbps of storage bandwidth to both encrypted and unencrypted EBS volumes.

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

High Memory instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/nvme-ebs-volumes.html). 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 (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/enhanced-networking.html) by default. With ENA, High Memory instances can utilize up to 25 Gbps of network bandwidth.

Q. Can I run High Memory instances in my existing Amazon Virtual Private Cloud (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 are EC2 bare metal instances, and do not run on a hypervisor. These instances allow the operating systems to run directly on the underlying hardware, while still providing access to the benefits of the cloud.

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 3.8 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?

High Memory instances are available on EC2 Dedicated Hosts on a 3-year Reservation. After the 3-year reservation expires, you can continue using the host at an hourly rate or release it anytime.

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. The Dedicated Host will be allocated to your account for the period of 3-year reservation. After the 3-year reservation expires, you can continue using the host or release it anytime.

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

You can launch, stop/start, and terminate instances on your EC2 Dedicated Hosts 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, and Red Hat Enterprise Linux 7.5 for SAP.

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 HANA (https://docs.aws.amazon.com/quickstart/latest/sap-hana/welcome.html) deployments to rapidly deploy all the necessary 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.

## 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. Your C1, C3, CC2, CR1, G2, HS1, M1, M2, M3, R3 and T1 instances are still fully functional and will not be deleted because of this change.

Q: Are Previous Generation instances being discontinued soon?

Currently, there are no plans to end of life Previous Generation instances. However, 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.

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.

## 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, High Performance Computing (HPC), scientific computing, and other memory-intensive applications.

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. When should I use X1e instances?

X1e instances are ideal for running in-memory databases like SAP HANA, high-performance databases and other memory optimized enterprise applications. X1e instances offer twice the memory per vCPU compared to the X1 instances. The x1e.32xlarge instance is 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. How do X1 and X1e instances differ?

X1e instances offer 32GB of memory per vCPU whereas X1 instances offer 16GB of memory per vCPU. X1e instance sizes enable six instance configurations starting from 4 vCPUs and 122 GiB memory up to 128 vCPUs and 3,904 GiB of memory. X1 instances enable two instance configurations, 64 vCPUs with 976 GiB memory and 128 vCPUs with 1,952 GiB memory.

Q. What are the key specifications of Intel E7 (codenamed Haswell) processors that power X1 and X1e instances?

The E7 processors have a high core count to support workloads that scale efficiently on large number of cores. The Intel E7 processors also feature high memory bandwidth and larger L3 caches to boost the performance of in-memory applications. In addition, the Intel E7 processor:

- Enables increased cryptographic performance via the latest Intel AES-NI feature.
- Supports Transactional Synchronization Extensions (TSX) to boost the performance of in-memory transactional data processing.
- Supports Advanced Vector Extensions 2 (Intel AVX2) processor instructions to expand most integer commands to 256 bits.

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.

Q: What operating systems are supported on X1 and X1e instances?

X1 and X1e instances provide high number of vCPUs, which might cause launch issues in some Linux operating systems that have a lower vCPU limit. We strongly recommend that you use the latest AMIs when you launch these instances.

AMI support for SAP HANA workloads include: SUSE Linux 12, SUSE Linux 12 SP1, SLES for SAP 12 SP1, SLES for SAP 12 SP2, and RHEL 7.2 for SAP HANA.

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. What storage options are available for X1 customers?

X1 instances offer SSD based instance store, which is ideal for temporary storage of information such as logs, buffers, caches, temporary tables, temporary computational data, and other temporary content. X1 instance store provides the best I/O performance when you use a Linux kernel that supports persistent grants (https://blog.xenproject.org/2012/11/23/improving-block-protocol-scalability-with-persistent-grants/), an extension to the Xen block ring protocol.

X1 instances are EBS-optimized by default and offer up to 14 Gbps of dedicated bandwidth to EBS volumes. EBS offers multiple volume types to support a wide variety of workloads. For more information see the EC2 User Guide (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/x1-instances.html).

Q. How do I build cost-effective failover solution on X1 and X1e instances?

You can design simple and cost-effective failover solutions on X1 instances using Amazon EC2 Auto Recovery (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instance-recover.html), an Amazon EC2 feature that is designed to better manage failover upon instance impairment. You can enable Auto Recovery for X1 instances by creating an AWS CloudWatch alarm. Choose the “EC2 Status Check Failed (System)” metric and select the “Recover this instance” action. Instance recovery is subject to underlying limitations, including those reflected in the Instance Recovery Troubleshooting documentation (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/TroubleshootingInstanceRecovery.html). For more information visit Auto Recovery documentation (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instance-recover.html) and Creating Amazon CloudWatch Alarms (http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/AlarmThatSendsEmail.html) respectively.

Q. Are there standard SAP HANA reference deployment frameworks available for the X1 instance and the AWS Cloud?

You can use the AWS Quick Start reference HANA deployments to rapidly deploy all the necessary HANA building blocks on X1 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. For additional information on deploying HANA on AWS, please refer to SAP HANA on AWS Cloud: Quick Start Reference Deployment Guide (https://s3.amazonaws.com/quickstart-reference/sap/hana/latest/doc/SAP+HANA+Quick+Start.pdf).

### 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 and HDD-storage instances compare to High I/O instances?

High I/O instances (I2) 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 (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-storage instances, d2.8xlarge, can deliver up to 3.5 GBps read and 3.1 GBps write disk throughput with a 2 MiB block size. The largest H1 instances size, h1.16xlarge, can deliver up to 1.15 GBps read and write. To ensure the best disk throughput performance from your D2 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?

The primary data storage for Dense-storage instances is HDD-based instance storage. Like all instance storage, these 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 durable data storage solutions such as Amazon Simple Storage Service (S3) for additional data durability. Please refer to Amazon S3 for reference.

Q. How do Dense-storage and 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 storage provides directly attached non-persistent, 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 H1 instances as Amazon EBS-optimized instances?

Each H1 instance type is EBS-optimized by default. H1 instances offer 1,750 Mbps to 14,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on H1 instances, launching a H1 instance explicitly as EBS-optimized will not affect the instance's behavior.

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. Are HDD-storage instances offered in EC2 Classic?

The current generation of HDD-storage instances (H1 instances) can only be launched in Amazon VPC. With Amazon VPC, you can leverage a number of features that are available only on the Amazon VPC platform – such as enabling enhanced networking, assigning multiple private IP addresses to your instances, or changing your instances' security groups. For more information about the benefits of using a VPC, see Amazon EC2 and Amazon Virtual Private Cloud (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-vpc.html) (Amazon VPC).

Q. Are Dense-storage instances offered in EC2 Classic?

The current generation of Dense-storage instances (D2 instances) can be launched in both EC2-Classic and Amazon VPC. However, by launching a Dense-storage instance into a VPC, you can leverage a number of features that are available only on the Amazon VPC platform – such as enabling enhanced networking, assigning multiple private IP addresses to your instances, or changing your instances' security groups. For more information about the benefits of using a VPC, see Amazon EC2 and Amazon Virtual Private Cloud (http://docs.aws.amazon.com /AWSEC2/latest/UserGuide/using-vpc.html) (Amazon VPC). You can take steps to migrate your resources from EC2-Classic to Amazon VPC. For more information, see Migrating a Linux Instance from EC2-Classic to a VPC (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/vpc-migrate.html).

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. I3 and I3en instances offer NVMe only storage, while previous generation I2 instances allow legacy blkfront storage access. Currently you can only purchase High I/O instances as On-Demand, Reserved Instances or as Spot instances.

Q. Is there a limit on the number of High I/O instances I can use?

Currently, you can launch 2 i3.16xlarge instances by default. If you wish to run more than 2 On-Demand instances, please complete the Amazon EC2 instance request form.

Q. How many IOPS can i3.16.xlarge instances deliver?

Using HVM AMIs, High I/O I3 instances can deliver up to 3.3 million IOPS measured at 100% random reads using 4KB block size, and up to 300,000 100% random write IOPs, measured at 4KB block sizes to applications across 8 x 1.9 TB NVMe devices.

Q. What is the sequential throughput of i3 instances?

The maximum sequential throughput, measured at 128K block sizes is 16 GB/s read throughput and 6.4 GB/s write throughput.

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 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. I3 and I3en instances support TRIM.

Q. How many IOPS can I3en.24xlarge instances deliver?

Using HVM AMIs, high I/O I3en instances can deliver up to 2 million IOPS measured at 100% random reads using 4KB block sizes, and up to 1.6 million 100% random write IOPs, measured at 4KB block sizes to applications across 8 x 7.5 TB NVMe devices.

Q. What is the sequential throughput of I3en instances?

The maximum sequential throughput, measured at 128K block sizes is 16 GB/s read throughput and 8 GB/s write throughput.

## Storage

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

### Amazon Elastic Block Store (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 and 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 on see the EBS product details page, and for additional information on performance, see the Amazon EC2 User Guide's EBS Performance section (http://docs.aws.amazon.com /AWSEC2/latest/UserGuide/EBSPerformance.html).

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 hard disk drives (HDDs) and provides 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 General Purpose SSD (gp2). 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 on Amazon EBS see the EBS product details page.

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

While you are able to attach multiple volumes to a single instance, attaching multiple instances to one volume is not supported at this time.

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 (http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=243). 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?

You can visit the Amazon EBS FAQ page.

### Amazon Elastic File System (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 (http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html).

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 (http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html).

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, and Amazon EC2 Classic instances outside your VPC can mount a file system via ClassicLink (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/vpc-classiclink.html). 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: Which instance types offer NVMe instance storage?

Today, I3en, I3, C5d, M5d, M5ad, R5d, R5ad, z1d, and F1 instances offer NVMe instance storage.

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

Yes, all data is encrypted in an AWS Nitro (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 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 Fabric Adapter (EFA) | Elastic IP | Elastic Load Balancing | Enhanced networking | Security

### 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?

High Performance Computing (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 which 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 Message Passing Interface (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?

EFA is currently available on the c5n.18xlarge, c5n.metal, p3dn.24xlarge, i3en.24xlarge, and i3en.metal instance sizes. Support for more instance types and sizes being added in the coming months.

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.

### 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 the 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 address here (https://aws.amazon.com/contact-us/eip_limit_request/). 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 to 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 (https://portal.aws.amazon.com/gp/aws/html-forms-controller/contactus/ec2-email-limit-rdns-request). 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.

## 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.

## 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. C5, C5d, F1, G3, H1, I3, I3en, m4.16xlarge, M5, M5a, M5ad, M5d, P2, P3, R4, R5, R5a, R5ad, R5d, T3, T3a, X1, X1e, and z1d instances use the Elastic Network Adapter (which uses the “ena” Linux driver) for Enhanced Networking. C3, C4, D2, I2, M4 (excluding m4.16xlarge), and R3 instances use Intel® 82599g Virtual Function Interface (which uses the “ixgbevf” Linux driver). 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: Why is Enhanced Networking only supported in Amazon VPC?

Amazon VPC allows us to deliver many advanced networking features to you that are not possible in EC2-Classic. Enhanced Networking is another example of a capability enabled by Amazon VPC.

Q: Which instance types support Enhanced Networking?

Depending on your instance type, enhanced networking can be enabled 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 25 Gbps for supported instance types. C5, C5d, F1, G3, H1, I3, I3en, m4.16xlarge, M5, M5a, M5ad, M5d, P2, P3, R4, R5, R5a, R5ad, R5d, T3, X1, X1e, and z1d instances use the Elastic Network Adapter for enhanced networking.

Q. Which instance types offer NVMe instance storage?

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 high bandwidth networking.

### 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 (https://console.aws.amazon.com/cloudtrail/home). 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 (https://d1.awsstatic.com/whitepapers/aws-security-whitepaper.pdf) white paper and to our Amazon EC2 running Windows Security Guide (http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1767).

## 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 (https://aws.amazon.com/ec2/autoscaling/) 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 determines 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 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 shutdown 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 (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Hibernate.html).

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 “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?

Hibernation is currently supported across M3, M4, M5, C3, C4, C5, R3, R4, and R5 instances with less than 150 GB of RAM running Amazon Linux 1 and Ubuntu. To review the list of supported OS versions, refer to the user guide (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Hibernate.html).

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 hibernate?

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 enough space, hibernation will fail and the instance will get shutdown 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 Windows Server 2003, Windows Server 2003 R2, Windows Server 2008, Windows Server 2012 R1, Red Hat Enterprise Linux (RHEL) 5.1-6.5 (using Cloud Access), Centos 5.1-6.5, Ubuntu 12.04, 12.10, 13.04, 13.10, and Debian 6.0.0-6.0.8, 7.0.0-7.2.0. For more details on VM Import, including supported file formats, architectures, and operating system configurations, please see the VM Import/Export section of the Amazon EC2 User Guide (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/VMImportPrerequisites.html).

Q. What virtual machine 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 support 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 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 and 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 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 virtual machine?

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-premise 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-premise 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 (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 | Convertible Reserved Instances | EC2 Fleet | On-Demand Capacity Reservation | Reserved Instances | Reserved Instance Marketplace | 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 will be charged as Internet Data Transfer on both sides of the transfer. 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Stop_Start.html).

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.

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 Internet Data Transfer rates. Therefore, if data is transferred between these two instances, it is charged at Internet Data Transfer Out for the first instance and at Internet 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. An example would be an instance running for 1 hour 10 minutes and 4 seconds would look like 1.1677. Read this (https://aws.amazon.com/blogs/aws/new-per-second-billing-for-ec2-instances-and-ebs-volumes/) 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.

### 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. Click here to learn more about exchanging Convertible RIs.

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

No, a 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 (http://docs.aws.amazon.com/AWSEC2/latest/APIReference/API_GetReservedInstancesExchangeQuote.html). 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 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 launches 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 (https://aws.amazon.com/blogs/aws/ec2-fleet-manage-thousands-of-on-demand-and-spot-instances-with-one-request/) 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 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 a 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.

### On-Demand Capacity Reservation

On-Demand Capacity Reservation is an EC2 offering that lets you create and manage reserved capacity on Amazon EC2. You can create a Capacity Reservation by choosing an Availability Zone 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 Capacity Reservations cost?

When the 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 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 a Capacity Reservation for 20 c5.2xlarge instances and you run 15 c5.2xlarge instances, you will be charged for 15 instances and 5 unused spots in the reservation (effectively charged for 20 instances).

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

Yes. Regional RI (RI scoped to a region) discounts apply to Capacity Reservations. AWS Billing automatically applies your RI discount when the attributes of a Capacity Reservation match the attributes of an active Regional RI. When a Capacity Reservation is used by an instance, you are only charged for the instance (with RI discounts

applied). Regional RI discounts are preferentially applied to running instances before covering unused Capacity Reservations.

For example, if you have a Regional RI for 50 c5.2xlarge instances and a Capacity Reservation for 50 c5.2xlarge instances in the same region, the RI discount will apply to the unused portion of the reservation. Note that discounts will first apply to any c5 instance usage (across instances sizes and Availability Zones) within that region before applying to unused reservations.

Note: Zonal RIs (RIs scoped to an Availability Zone) do not apply to Capacity Reservations, as Zonal RIs already come with a capacity reservation.

Q. When should I use RIs and when should I use Capacity Reservations?

Use Regional RIs for their discount benefit while committing to a one or three year term. Regional RIs automatically apply your discount to usage across Availability Zones and instance sizes, making it easier for you to take advantage of the RI’s discounted rate.

Use Capacity Reservations if you need the additional confidence in your ability to launch instances. Capacity Reservations can be created for any duration and can be managed independently of your RIs.

If you have Regional RI discounts, they will automatically apply to matching Capacity Reservations. This gives you the flexibility to selectively add Capacity Reservations to a portion of your instance footprint and still get the Regional RI discounts for that usage.

Q. I have a Zonal RI (RI scoped to an Availability Zone) that also provides a capacity reservation? How does this compare with a Capacity Reservation?

A Zonal RI provides both a discount and a capacity reservation in a specific Availability Zone in return for a 1-to-3 year commitment. Capacity Reservation allows you to create and manage reserved capacity independently of your RI commitment and term length.

A Regional RI can be combined with an On-Demand Capacity Reservation to get, at the minimum, the exact same benefits of a Zonal RI for no additional cost. You also get the enhanced flexibility of Regional RI discounts and the features of Capacity Reservation: the ability to add or subtract from the reservation at any time, view utilization in real-time, and the ability to target a Capacity Reservation for specific workloads.

Re-scoping your Zonal RIs to a region immediately gives you the Availability Zone 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 Availability Zone to a region using the EC2 management console or the ModifyReservedInstances API.

Q. I created a Capacity Reservation. How can I use it?

A Capacity Reservation is tied to a specific Availability Zone and, by default automatically utilized by running instances in that Availability Zone. 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 (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-reservations.html) or windows (https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/ec2-capacity-reservations.html) technical documentation to learn more about the targeting option.

Q. How many instances am I allowed to reserve?

The number of instances 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 (https://aws.amazon.com/support/createCase?type=service_limit_increase&serviceLimitIncreaseType=ec2-instances) 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 a Capacity Reservation after it has started?

Yes. You can reduce the number of instances 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 a Capacity Reservation that has ended or has been deleted.

Q. Can I end a Capacity Reservation after it has started?

Yes. You can end a 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 Capacity Reservation including deletion or expiration of a reservation.

Q. Where can I find more information about using Capacity Reservations?

Refer to Linux (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-reservations.html) or windows (https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/ec2-capacity-reservations.html) technical documentation to learn about creating and using a Capacity Reservation.

Q. Can I share a Capacity Reservation with another AWS Account?

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

Note that sharing of 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 master (payer) account or through an AWS Organization are exempt from this restriction.

Q. What happens when I share a Capacity Reservation with another AWS account?

When a Capacity Reservation is shared with other accounts, those accounts can consume the reserved capacity to run their EC2 Instances. The exact behavior depends by the preferences set on the Capacity Reservation. By default, Capacity Reservations automatically match existing and new instances from other accounts that have shared access to the reservation. You can also target a Capacity Reservation for specific workloads/instances. Individual accounts can control which of their instances consume Capacity Reservations. Refer to Linux (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-reservations.html) or windows (https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/ec2-capacity-reservations.html) technical documentation to learn more about the instance matching options.

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

There is no additional charge for sharing a reservation.

Q. Who gets charged when a Capacity Reservation is shared across multiple accounts?

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

Q. Can I prioritize access to Capacity Reservation among the AWS accounts that have shared access?

No. Instance spots in a Capacity Reservation are available on a first-come-first-serve basis to any account that has shared access.

Q. How can I communicate the Availability Zone (AZ) of a CR with another account, given AZ name mappings could be different across AWS accounts?

You can now use Availability Zone ID (AZ ID) instead of AZ name. Availability Zone 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 Capacity Reservation once I have shared it?

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

Q. Where can I find more information about sharing Capacity Reservations?

Refer to Linux (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/capacity-reservation-sharing.html) or windows (https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/capacity-reservation-sharing.html) technical documentation to learn about sharing Capacity Reservations.

### 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 forego 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 Availability Zones 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 Availability Zone.

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 offers broader applicability of the RI’s discounted rate.

Q: What are Availability Zone and instance size flexibility?

Availability Zone and instance size flexibility make it easier for you to take advantage of your regional RI’s discounted rate. Availability Zone flexibility applies your RI’s discounted rate to usage in any Availability Zone 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.

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

Regional RIs do not require any action to take advantage of Availability Zone 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 Availability Zone to a region from the EC2 management 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 Management Console or by using the AWS CLI. Simply specify the instance type, platform, tenancy, term, payment option, and region or Availability Zone.

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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html#apply_ri) 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html).

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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html#apply_ri) to learn more about how instance size flexibility of RIs applies to your EC2 usage. And click here (http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/billing-reports.html#enhanced-RI) 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 Availability Zone of the RI, change the scope of the RI from Availability Zone to region (and vice-versa), change the network platform from EC2-VPC to EC2-Classic (and vice versa) 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 Availability Zone, 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 alongside.

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 (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html#reserved-instances-discounts) portion of the Amazon EC2 User Guide (http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/Welcome.html).

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, however 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 (http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-tiers.html) portion of the Amazon EC2 User Guide (http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/Welcome.html).

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 Marketplace?

The Reserved Instance Marketplace is an online marketplace that provides AWS customers the flexibility to sell their Amazon Elastic Compute Cloud (Amazon EC2) Reserved Instances to other businesses and organizations. Customers can also browse the Reserved Instance Marketplace to find an even wider selection of Reserved Instance term lengths and pricing options sold by other AWS customers.

Q. When can I list a Reserved Instance on the Reserved Instance Marketplace?

You can list a Reserved Instance when:

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

Q. How will I register as a seller for the Reserved Instance Marketplace?

To register for the Reserved Instance Marketplace, you can enter the registration workflow by selling a Reserved Instance from the EC2 Management Console (https://console.aws.amazon.com/) 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 Reserved Instances, or plan to sell 50 or more Reserved Instances, you will need to provide tax information before you can list your Reserved Instances. 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 Reserved Instances 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 Reserved Instance Marketplace?

You can start selling on the Reserved Instance 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 a Reserved Instance for sale?

To list a Reserved Instance, simply complete these steps in the Amazon EC2 Console:

1. Select the Reserved Instances 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 Reserved Instance 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 Reserved Instance 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 Reserved Instances can I list for sale?

You can list any Reserved Instances 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 Reserved Instance can be in the active state prior to AWS receiving payment. In this case, your Reserved Instance will not be listed until we have received your payment.

Q. How are listed Reserved Instances displayed to buyers?

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

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

Q. How much of my Reserved Instance term can I list?

You can sell a Reserved Instance 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 Reserved Instance.

Q. Can I remove my Reserved Instance after I've listed it for sale?

Yes, you can remove your Reserved Instance listings at any point until a sale is pending (meaning a buyer has bought your Reserved Instance and confirmation of payment is pending).

Q. Which pricing dimensions can I set for the Reserved Instances I want to list?

Using the Reserved Instance 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 Reserved Instance), 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 Reserved Instance 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 a Reserved Instance that I purchased from the Reserved Instance Marketplace?

Yes, you can resell Reserved Instances purchased from the Reserved Instance Marketplace just like any other Reserved Instance.

Q. Are there any restrictions when selling Reserved Instances?

Yes, you must have a US bank account to sell Reserved Instances in the Reserved Instance Marketplace. Support for non-US bank accounts will be coming soon. Also, you may not sell Reserved Instances in the US GovCloud region.

Q. Can I sell Reserved Instances purchased from the public volume pricing tiers?

No, this capability is not yet available.

Q. Is there a charge for selling Reserved Instances on the Reserved Instance Marketplace?

Yes, AWS charges a service fee of 12% of the total upfront price of each Reserved Instance you sell in the Reserved Instance Marketplace.

Q. Can AWS sell subsets of my listed Reserved Instances?

Yes, AWS may potentially sell a subset of the quantity of Reserved Instances that you have listed. For example, if you list 100 Reserved instances, 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 Reserved Instances until and unless you decide not to list them any longer.

Q. How do buyers pay for Reserved Instances that they've purchased?

Payment for completed Reserved Instance 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 Reserved Instance 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 Reserved Instance was been sold.

Q. If I sell my Reserved Instance in the Reserved Instance Marketplace, will I get refunded for the Premium Support I was charged too?

No, you will not receive a pro-rated refund for the upfront portion of the AWS Premium Support Fee.

Q. Will I be notified about Reserved Instance Marketplace activities?

Yes, you will receive a single email once a day that details your Reserved Instance Marketplace activity whenever you create or cancel Reserved Instance 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 Reserved Instances?

Yes, you cannot purchase your own listed Reserved Instances, including those in any of your linked accounts (via Consolidated Billing).

Q. Do I have to pay for Premium Support when purchasing Reserved Instances from the Reserved Instance Marketplace?

Yes, if you are a Premium Support customer, you will be charged for Premium Support when you purchase a Reserved Instance through the Reserved Instance Marketplace.

## Spot Instances

Q. What is a Spot Instance?

Spot Instances are spare EC2 capacity that can save you up 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 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-limits.html).

If you would like a higher limit, complete the Amazon EC2 instance request form (https://aws.amazon.com/support/createCase?serviceLimitIncreaseType=ec2-instances&type=service_limit_increase&isauthcode=true&code=dzkwDcZ52oii6pr9iPYylIvuM78ObPVkQug42ntPetVgfanCcFjKifdSEPXRx2_tIQp6hLDGGq_GSvJ7t1wKy_3x97YfW6IheoEZJhc9I_rifBs7i_76ZG8f5qZ8EU10kPjq1Bm6L-vOdh3iTfdFaObJ_irgWqaWTsMcZrt0Je9t9HV3272554QNYYjLS0PdDnBFbZlalCUf4p37KTyoifrZPgMGdVVZZuRE0KKJIOvEZo3GIw-MpYkcA0MlUqT3jenFQ9V8CYtkM0T7BCfdWmdXN59U4fmS9NKp75gpAvub4PTi1J5GqzS6QRE_UsKy6BXqguDhyQs57Unu93COwLZdwEmqMqo6WBhAi90ZY10XfleVigCgLvcp1Nh7_dPZ2W6Dn4oldC3odsTS4ZrX7tfsa13LsysDgEdgI6TzNeZ3rV1UpRlUwQe8W7tVYLmKQNPHz4DdEKynfj7Zas-RSTAUOJcbGmylTsraI4Vx9mpnZgs8vnFf-3sAzgmuchdoZLx0qmTrJSZWNm7qOLwxUVbNX2kaGocCzioUn1wxcN2AFcyanjFUHuVlxRoSYc4WYGncCZGuCS-bnoNZqYmTxLZniybzAixVS3UecOgOFG8QX8LJkVFkB-I2LUdfp4Cvq1UqT9rB00VKRNW2NPauDVxb2zYpcOabNGylt9dZiitfTsCy3mrULMRPOZQiBRqHyCIQ6Th-UZfT9uelnfpCXG5WD2lEEIN0W7ZZEpkoYbmVrFv23y2ahqXkzXLs3_YqOwDw9ODQt8jpc9_v-lPXaHLwvwOJNEYLzbFDdWo07E6w6dBiXoOBjBgOtFLC8lUU4UmrrMEahS32I-WdShTpBrQAPfx1TVGDAFhPbhEG-DG199LowkOlVAPWVWXZrHmKHxapnpflHWOm6ZNrA_dh4a5xho-ca101el-3AG2CEwIPZw8s0L5f2t9clLrwzuVH6OD5NbH23PptJs3iGVN20bjwPgpzlxThzxKMHziT4wXAcrTOAzFUa7t0v5wtbL

NJSPURrXYKWYX7SJmzZELL7z9YKqRYUGCyiVmvyIhOHHjSLhsmvqfWO3lgjeBkNBTeKEbsLzMHwGNlFZKGqny D8SCGwdrIBvLxanjR2XsAmPX-CrWAJU9KaRfXDZli-OusDX-T1DWcUKoLtJe1hnOiqZgzCgrb7wABkRRpCG8hgmvVaeKPSp-_YvlFbg) 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, Availability Zone, and network platform (EC2-Classic or EC2-VPC). 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 (https://aws.amazon.com/ec2/spot/getting-started/).

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 (https://aws.amazon.com /documentation/ec2/).

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 are 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. 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). Secondarily, 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-hibernation.html) 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 shutdown 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Stop_Start.html). 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 enough space, hibernation will fail and the instance will get shutdown 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-hibernation.html) 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 restart a stopped instance or resume a hibernated instance?

No, you will not be able to re-start a stopped instance or resume a hibernated instance directly. Stop-start and hibernate-resume cycles are controlled by Amazon EC2. If an instance is stopped or hibernated by Spot, it will be restarted or 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-hibernation.html).

Q. How will I be charged if my Spot instance is 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 terminate the 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 terminate the instance yourself, you will be charged for an entire hour.

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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-interruptions.html#spot-instance-termination-notices)), 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet.html).

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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-limits.html#spot-fleet-limitations) 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-requests.html#manage-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 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-examples.html) 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet.html#spot-fleet-allocation-strategy) 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-requests.html#manage-spot-fleet) 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 re-launch 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 (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/set-time.html) and Windows (http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/windows-set-time.html).

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 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 is 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 a high performance networking for High Performance Compute (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 use of this instance type for HPC applications.

Q. What kind of network performance can I expect when I launch instances in 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 multi-flow 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 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, power and network maintenance without disrupting customer instances. To achieve this we employ a combination of tools and methods across the entire AWS Global infrastructure, such as redundant and concurrently maintainable systems, as well as live system updates and migration. For example, in these cases - Example 1, Example 2 - EC2 used live system updates to perform the required security maintenance non-disruptively for over 90% of EC2 Instances, with each maintenance completing in less than two seconds. AWS continuously invests in technology and processes to complete routine maintenance ever more safely and quickly, often with no disruption to customer instances.

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 (M5, M4) or burstable performance (T2); Compute Optimized instances (C5, C4) 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 (X1e, X1, R4) offer larger memory sizes for memory-intensive applications, including database and memory caching applications; Accelerating Computing instances (P3, P2, G3, F1) take advantage of the parallel processing capabilities of NVIDIA Tesla GPUs for high performance computing and machine/deep learning; GPU Graphics instances (G3) offer high-performance 3D graphics capabilities for applications using OpenGL and DirectX; F1 instances deliver Xilinx FPGA-based reconfigurable computing; Storage Optimized Instances (H1, I3, I3en, D2) that provide very high, low latency, I/O capacity using SSD-based local instance storage for I/O-intensive applications, with D2 or H1, the dense-storage and HDD-storage instances, 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 (http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide /concepts_micro_instances.html) about use of 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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/nitro/) Hypervisor will remain the same. For example, the “hypervisor” field of the DescribeInstances response, which 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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 (https://aws.amazon.com/ec2/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 high-performance computing (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 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 instance.

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 (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-optimize-cpu.html).

## Workloads

Amazon EC2 running IBM (https://aws.amazon.com/partners/find/partnerdetails/?n=IBM-&id=001E000000UfakGIAR) | Amazon EC2 running Microsoft Windows and other third-party software

### 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 (i.e. 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 (https://aws.amazon.com/partners/find/partnerdetails/?n=IBM-&id=001E000000UfakGIAR).

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.

Check out additional product-related resources

There are many resources to help you learn how to build with Amazon EC2.

Learn more

Sign up for a free account

Instantly get access to the AWS Free Tier.

Start building in the console

Get started building with Amazon EC2 in the AWS Console.

Register for re:Invent Bootcamps

Learn from AWS experts at exam prep bootcamps - space is limited!

(https://reinvent.awsevents.com/learn/bootcamps/?sc_icampaign=aware_reinvent_bootcamps2019_aws&sc_ichannel=ha&sc_icontent=awssm-2448&sc_iplace=2up&trk=ha_awssm-2448)

Get AWS Certified

AWS Certifications can help you advance your career and boost your earning power

(https://pages.awscloud.com/tc_get-aws-certified.html?sc_icampaign=aware_getcertified_evergreen2019&sc_ichannel=ha&sc_icontent=awssm-2655&sc_iplace=2up&trk=ha_awssm-2655)

AWS re:Invent | December 2 – 6, 2019 | Las Vegas, Nevada

Invite your friends to register for re:Invent, and compete to win a VIP experience on-site plus free passes to Intersect. Learn more ≫

AWS re:Invent

(https://reinvent.awsevents.com/vip/?sc_icampaign=Event_reInvent_2019_1up_DG3_VIP&sc_ichannel=ha&sc_icontent=awssm-2785&sc_ioutcome=Strategic_Events&sc_iplace=1up&trk=ha_a131L0000057yyxQAA~ha_awssm-2785&trkCampaign=AWS_reInvent_2019)

Page Content

General Instance types Storage Networking and security Management Billing and purchase options Platform Workloads

aws.amazon.com (https://aws.amazon.com/ec2/faqs/)