PROWESS

Technical Research Study

Cloud Computing: Why You Should Be Looking Under the Hood

Prowess Consulting explains the Amazon® Elastic Compute Cloud™ (Amazon EC2®) architecture nomenclature and how you can select instances that improve cloud performance, efficiency, and total cost of ownership (TCO).

Executive Summary

Cloud users are presented with an abundance of choices when trying to choose Amazon Web Services® (AWS®) cloud services and instances. This report sponsored by Intel demystifies the intricacies of Amazon® Elastic Compute Cloud™ (Amazon EC2®) instance types and helps you select Amazon EC2 instances that deliver the best value.

The challenge of trying to find the right Amazon EC2 instance type for your needs from among a selection of over 600 instance types can lead many users to rely on a recommender system such as AWS Compute Optimizer. However, the convenience of using an automated instance setup may result in higher cloud-computing costs.

Remember the following when choosing cloud instances:

- **Time is money in the cloud** and selecting a higher-performing processor can help lower your operational and cloud licensing costs.
- Your instance type choice can **affect performance per dollar**.
- Manually selecting instances **maximizes your resource choices and your control over costs**.
- Migrating to a newer Intel instance can **deliver a host of performance, functionality, and total cost of ownership (TCO) benefits**.

Our research indicates that the easiest instance selection method may not deliver the best business value. We suggest that a basic understanding of how Amazon EC2 instance types are named can help you make informed instance choices that align with your organization's performance requirements, financial constraints, and cloud strategy goals.

Do your homework before migrating

to an EC2 instance powered with a different processor—
or you could be surprised by higher cloud costs.

Highlights

A processor that combines high-performance cores with built-in accelerators can help lower your cloud-computing TCO:

Reduce per-core licensing costs by running smaller instance sizes.

Shorten time to insights and to market by completing cloud tasks more quickly.

Save money every time you open an instance by increasing workload performance per dollar.

Welcome to the AWS Showroom!

Navigating the world of cloud-based computing is like walking through a vast automobile showroom. This online showroom lets you explore a wide array of Amazon EC2 instances, which are the vehicles that drive your applications and services in the cloud. As you peruse the different offerings, you will notice that each make and model represents a set of standard features and capabilities, as well as a range of optional features.

Selecting the right automobile can keep you safe during your daily commute, save you money on fuel, and provide the performance you desire. Similarly, choosing the right Amazon EC2 instance can optimize your cloud operations, reduce capital and operating costs, and ensure that your applications run smoothly.

We developed this guide to help you decipher the "manufacturer sticker" for an Amazon EC2 instance type and understand what's running under the hood of your cloud instances. Regardless of what AWS pricing plan you choose, these suggestions can help you make the right choice using the best price performance and specific requirements of your AWS services.

Makes and Models

Let's take a closer look at the information you can collect from an Amazon EC2 instance type name. Once you understand how the naming convention works, you can quickly narrow down your Amazon EC2 instance choices simply by knowing what to look for.

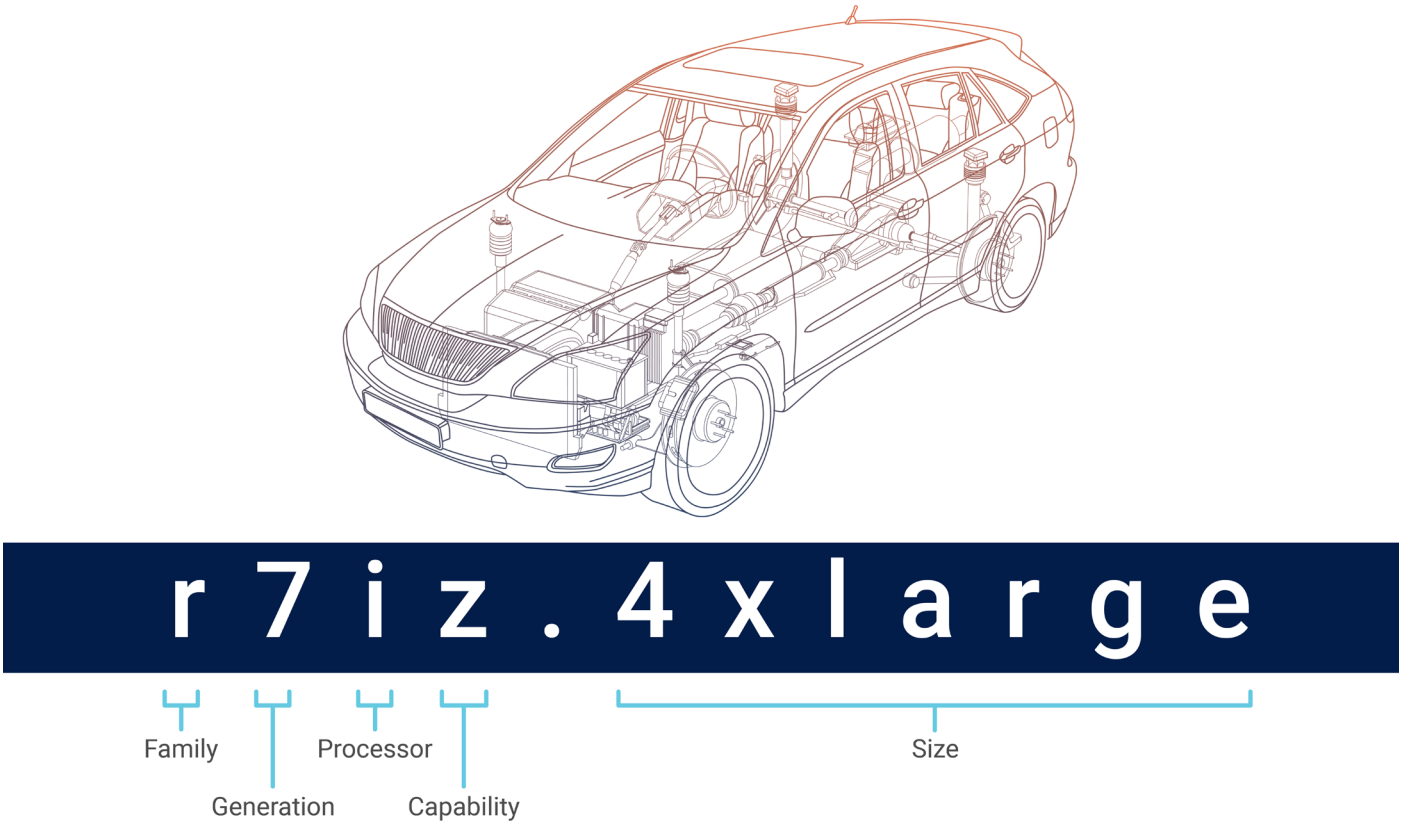

Figure 1 | Amazon EC2® instances follow an architectural naming system that can help you choose the right cloud instance for your needs

Reading from left to right in Figure 1, the first letter of an Amazon EC2 instance name designates its instance family. Next comes a number that designates the instance generation. The next one or two characters are letters. The first letter designates the processor family (CPU) inside the server rack at AWS. The next 1–3 letters are optional and, if present, designate additional capability. The string of numbers and letters following the period describe the instance size in terms of virtual CPU (vCPU), memory, storage, and network bandwidth. To avoid wordy descriptions, this study shortens the positions in the taxonomy to family, generation, processor, capability, and size for each instance and focuses our discussion on families suitable for typical workloads.

Family

Amazon EC2 family letters can be mapped to the following workload categories:

- General purpose: m, t
- Compute optimized: c
- Memory optimized: r, x, z, u (known as "high memory")
- Storage optimized: d, h, u, i, im, is
- High-performance computing (HPC) optimized: hpc
- Accelerated compute:
    - f: field programmable gate array (FPGA)
    - g: graphics acceleration
    - p: HPC with GPU
    - dl: deep learning
    - vt: video transcode
    - trn: deep learning training
    - inf: deep learning inference

Instance families are like car models and are designed to provide the best user experience for the type of driving you're doing. Like passenger sedans and hatchbacks, general purpose instances can handle a variety of computing tasks and deliver a balanced mix of compute power and memory capacity. Compute optimized instances have a higher ratio of compute to memory than general purpose instances for compute-bound applications. Memory optimized instances are like SUVs that can carry more passengers or cargo than sedans and wagons can. These instances are designed for workloads that demand a larger memory capacity. Storage-optimized instances are cargo vans and pickup trucks, built to transport and store large amounts of data. Accelerated compute and HPC-optimized instances are your turbo-charged sports cars. They offer high-performance computing and high-capacity memory, usually at the price of having relatively high energy and cost requirements.

Generation

Generation numbers are issued sequentially. AWS groups together processors from different manufacturers that are contemporaries and that deliver similar performance, assigning them the same instance generation number. This helps users make choices within generations and gives them distinctions between generations. As you would expect, a newer Amazon EC2 instance generation with a higher number offers additional capabilities and improved performance over previous generations.

Starting with AWS EC2 instance generation 6, users can choose from at least three manufacturers within each generation. The latest AWS EC2 generation is 7, which groups together 4th Gen Intel® Xeon® processors, Graviton3 processors, and 4th Generation AMD EPYC™ processors.

Processor

The processor letter can be mapped to a manufacturer:

- AMD: a
- Graviton: g
- Intel: i beginning with generation 6; no letter designation in generation 5 and older

The processor designation determines what engine is running under the hood of your Amazon EC2 instance. What makes this information important is that different manufacturers' processors deliver different levels of overall performance, energy efficiency, and TCO in the cloud.[1] This hardware effect is why we recommend that any cloud implementation strategy should include actively identifying and selecting the processor with the manufacturer in mind.

Capability

If the capability designation is available for an instance type, it's shown as 1–3 letters that appear after the processor letter and before the period. Here are some common categories:

- Enhanced network performance: n
- High frequency: z
- Flex instance: flex

- Enhanced Amazon® Elastic Block Store (Amazon® EBS) performance: b
- Local storage support: d
- Extended memory support: e

Size

The information following the period is the Amazon EC2 instance size designation. The instance size correlates with the provisioning of compute and memory resources. You can get a rough idea of instance size from the descriptor designations, listed here from small to large: nano, micro, small, medium, large, xlarge, and metal.

Numerations in multiples of 2 or 4 added before the designation also indicate increases in instance size. For example, Table 1 compares the sizes of general purpose instances m7i.**large**, m7i.**xlarge**, m7i.**2xlarge**, and m7i.**4xlarge** with Intel processors.[2] Note how the vCPU counts and memory capacities increase as the instance size increases.

Table 1 | Instance size directly correlates with compute and memory capacities

Instance size	vCPUs	Memory (GB)
m7i.large	2	8
m7i.xlarge	4	16
m7i.2xlarge	8	32
m7i.4xlarge	16	64

Why Your Choice of Instance Type Matters

Now that you understand how to read an Amazon EC2 instance type, let's discuss how to use this knowledge to find your best cloud-computing value. As mentioned earlier, selecting the processor type can determine the performance, efficiency, and TCO you get from AWS services and apps.

Remember, Time Is Money

A useful rule of thumb to remember is that most cloud pricing models are calculated from how much time it takes to run a workload to completion. The time spent on the processor is the basis for how you're charged for using on-demand, reserved, or spot-based instances, and for using plans like high-volume savings plans or tiered-discount plans. In general, the less time it takes your workload to run, the less you pay.

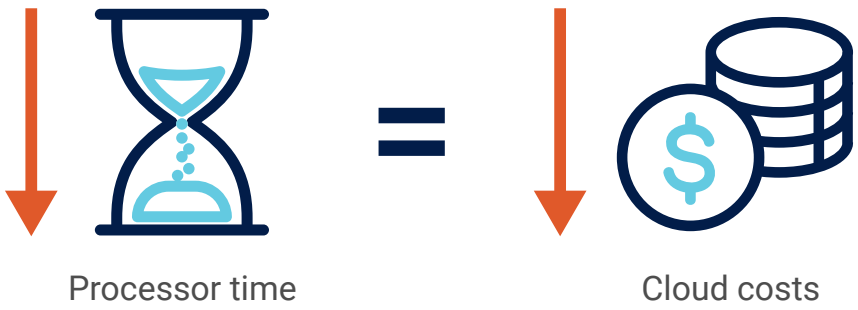

Figure 2 | Time is money in the cloud

Here's an example that illustrates why choosing your instance type matters. In a Prowess Consulting benchmarking study, we compared the performance of Amazon EC2 general purpose sixth-generation instances with Intel Xeon processors and Graviton2 processors. For Elasticsearch workloads, the Intel-based M6i instances ran more than 1.5x faster than the Graviton2-based M6g instances.[3] The shortened amount of time spent on the processor reduces instance operating costs, lowers power consumption, and reduces administrative time—all of which contribute to lowering TCO. Faster cloud performance also shortens your time to solution, which can help reduce production costs and increase profitability.

Reduce Cloud-Licensing Costs

Selecting a higher-performing processor can contribute to lower cloud licensing costs. Many application subscriptions are based on the number of vCPUs in the user's cloud instance. Carefully assessing how many vCPUs are required to meet your service-level agreements (SLAs) can lower the cost of per-core-based licensing. With better performance from each core, you can meet requirements with fewer cores.

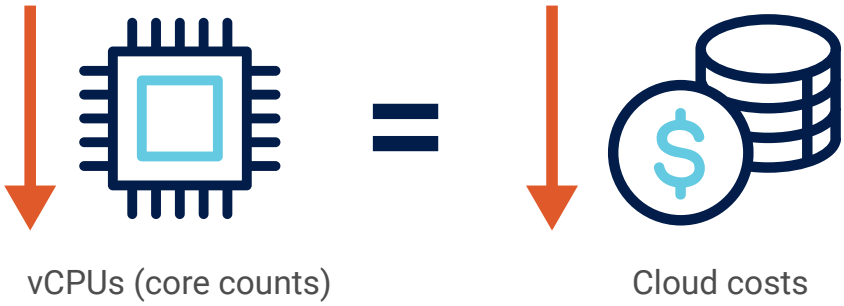

Figure 3 | Reducing core counts can help lower TCO via reduced licensing costs

Intel testing in AWS sixth-generation compute optimized instances with Intel Xeon processors and Graviton2 processors shows that when running the MongoDB database, for example, you can expect up to 1.5x higher throughput in C6i instances than when using C6g instances.[4] This performance boost means that you could run database workloads on a smaller c6i.4xlarge instance with Intel, as opposed to running a c6g.8xlarge instance with Graviton, and still meet your SLA. The smaller instance size halves the number of vCPUs you use, effectively reducing your core-based licensing costs by up to 50 percent. Another way you could benefit from higher performance per instance size would be to support more workloads per instance or complete more workloads in less time.

Boost Performance per Dollar

Your instance type choice can also affect performance per dollar. We conducted a price performance analysis using Intel testing results for WordPress® TPS workloads and the AWS on-demand pricing model. We discovered that M6i instances with Intel let you run up to 1.20x more transactions per second per dollar than do M6g instances with Graviton2.[5]

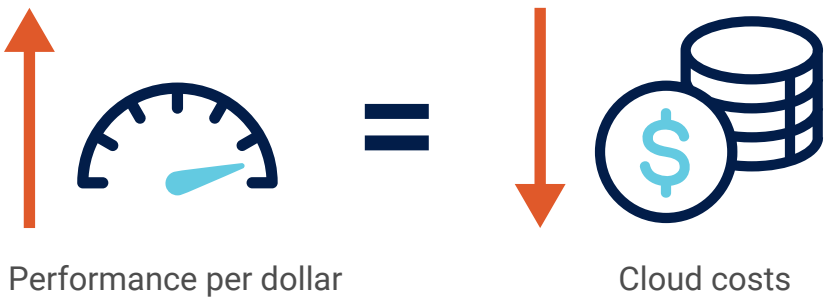

Figure 4 | Instances having higher performance per dollar deliver cost savings every time you open them

Instances that provide higher performance per dollar offer another "time is money" opportunity to lower your cloud TCO. They start lowering instance costs as soon as you open them, and the more you use them the greater your savings.

Selecting Your Instance Type

After you decide which instance type can help you achieve your business or operational goals, the next step is to select and open the instance you want to run. Unfortunately, choosing a cloud instance can be a confusing and time-consuming process. In this section, we introduce methods used for setting up an Amazon EC2 instance, along with pointers for choosing your instance type. Bear in mind that AWS offers online tools (discussed further on) that will make instance choices for you. The goal of this section is to show you where to look for better instance selection options.

Letting Your Provider Choose Your Instance

Returning to our automobile analogy, how you go about choosing a new set of wheels probably depends on your level of technical expertise. Perhaps you're looking for something "fast and sporty with a price tag under $40K," without knowing the difference between torque, horsepower, and compression ratio. Unless you bring along a mechanically inclined friend, you're reliant on the dealer to recommend models that match your preferences.

In the cloud world, AWS offers AWS Compute Optimizer, which recommends Amazon EC2 instances based on attributes such as type of workload, number of vCPUs, and memory capacity. AWS Compute Optimizer greatly simplifies Amazon EC2 instance selection by selecting the cloud instance family, processor, capability, and size for you, based on whatever attributes you need. It also helps analyze pricing plans and discounts based on your usage estimates.

The trade-off for such a streamlined instance selection process is that you're blindly relying on your cloud provider's recommendations. AWS Compute Optimizer doesn't readily let you inspect the processors you're using and see which manufacturers they're from. There are options for viewing and selecting other processors; however, finding them requires a bit of digging through the optimizer documentation. If you spend the extra time and effort to look under the hood, you will discover that the optimizer automatically selects Graviton-based instances by default. By extension, this means its performance and pricing recommendations only include Graviton processors. This process is akin to you asking about red two-door sports coupes and the dealer recommending a red pickup truck because both vehicles can carry passengers, have two doors, and are red.

Determine Your Destiny in the Cloud with Manual Selection

The optimizer's default setting to include only Graviton processors is why we suggest making the extra effort to actively inspect and select the processor powering your instance types. AWS has a wide selection of processor options because they understand that offering choices is critical to enhancing user experiences and customer satisfaction. We just wish they made it a bit easier to exercise those options.

In this section, we introduce additional tools you can use to inspect and select different Amazon EC2 instance types. Manual selection allows you to select Amazon EC2 instances from a robust assortment of families, generations, processor manufacturers, and capabilities. The larger instance pool lets you make accurate comparisons of cloud performance, efficiency, and pricing that reflect live production environments.

Let's revisit our auto showroom as another type of buyer. Perhaps you're the type who knows how to read a third-party vehicle history report, measure brake pad wear, and tell the difference between oil, brake, and transmission fluids. In other words, you're an IT administrator or DevOps team member who can use the optimizer documentation to manually select your instance processor.

At the far end of our car-buyer spectrum is the person who can assemble a classic car kit or rebuild a blown transmission. These are our cloud architects, developers, or engineers who design bare-metal servers using code recipes and GitHub resources. They might find this Amazon EC2 instance selector to be a handy tool.

One important consideration with any instance recommender, whether automated or manual, is that most of them will not help you find the optimal ecosystem settings for a selected instance. A misconfigured cloud environment can rack up recurring extra charges until you can troubleshoot and resolve the problem(s). An analyzer tool, such as the Granulate™ optimizer and migration tool for Intel-based instances, can help ensure that your cloud environment is properly configured for your selected Amazon EC2 instance.[6]

Save More with Spot-Based Instances

If your SLA is flexible, Amazon EC2 Spot Instances Pricing offers an excellent way to lower your cloud-computing costs. The results of a Prowess Consulting price performance analysis concluded that spot-based pricing could save you over 2.5x the price of on-demand pricing for Monte Carlo simulations running on the same instance types.[7]

Migrating to the Next Generation

Migrating to a newer-generation instance can deliver a host of performance, functionality, and TCO benefits. If you're already running Amazon EC2 instances on Intel Xeon processors, we strongly recommend doing your homework before switching to a different processor manufacturer as part of a migration to a newer instance generation. Switching manufacturers means changing platform architectures, which is rarely a seamless or painless process.

Look carefully for any unexpected costs. One cost consideration is that selecting a new manufacturer's processor may require hiring additional IT staff to properly design, test, deploy, monitor, administer, and optimize the infrastructure. Also, if you add instances with a second manufacturer's processor to your existing infrastructure, you may have to maintain a split code base that requires administration by separate IT teams. If your applications were designed to run on one manufacturer's processor and you move to another manufacturer's processor, you'll spend additional time analyzing and optimizing your workloads post-migration. If the processor you selected doesn't deliver the performance you need, you could end up provisioning more vCPUs and memory resources to meet SLAs, which would of course increase the cloud costs.

Another reason we recommend choosing to refresh your infrastructure on Intel is the flexibility and scalability afforded to you by having multiple instance options to choose from. AWS offers over 400 Intel-based Amazon EC2 instance types, which collectively represent nearly two-thirds of the more than 600 total Amazon EC2 instance types that are available. To help plan your migration to a newer generation of Intel-based instances, Table 2 shows the past and current generations of some popular Amazon EC2 instance types.

Table 2 | Intel®-based Amazon EC2 instance types

Intel® processor	General purpose	Compute optimized	Memory optimized	Accelerated compute	Storage optimized	HPC
4th Gen Intel® Xeon® Scalable processor	M7i*, M7i-flex*	C7i*	R7i*, R7iz*			
Intel Gaudi® processor				DL1		
3rd Gen Intel Xeon Scalable processor	M6id, M6idn	C6id, C6in	X2idn, X2iedn, R6id, R6idn		I4i	HPC6id
2nd Gen Intel Xeon Scalable processor	M5zn		R5dn, R5b, X2iezn	G4dn, P4d	D3, D3en	
1st and 2nd Gen Intel Xeon Scalable processors	T3, M5dn	C5d	R5d, HMI			
Intel Xeon Scalable processor	M5d	C5n	Z1d	P3dn	I3en	
Intel Xeon v4 processor	T2, M4		X1, R4	P2, P3, G3, F1	H1, I3	
Intel Xeon v3 processor		C4	X1e		D2	

* Newly available

Conclusion

Choosing AWS cloud services and instances from a myriad of available options can be a monumentally confusing and time-consuming task. While it may seem simpler to rely on system defaults or provider auto-selectors, taking the easy route can lead to overpaying for cloud licensing and services or having to correct platform configurations.

You can use this technical research study to understand your options and be a more informed, active, and cost-effective cloud user. A good start is understanding how Amazon EC2 architecture nomenclature is organized. This basic knowledge allows you to select the best instance type for your needs, whether you are looking for compute capability, memory capacity, overall performance, or TCO savings.

Learn More

Discover more about Intel on AWS by visiting "When to select Intel over Graviton."
Give your developers the tools they need to develop in the cloud.
See more research reports by Prowess Consulting.

[1] Prowess Consulting. "Java-Based Benchmarking Shines a Light on How Underlying Architecture Impacts Cloud Performance," Sponsored by Intel. 2022.

[2] Amazon. Amazon EC2 M7i Instances.

[3] Normalized performance results based on latency testing. Prowess Consulting. "Java-Based Benchmarking Shines a Light on How Underlying Architecture Impacts Cloud Performance." Commissioned by Intel. 2022. See endnote 4 and Table 2 for details.

[4] Intel. See claim [143] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[5] Intel. See claim [132] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary. AWS pricing information for US-East-1c as of January 5, 2023. https://aws.amazon.com/ec2/pricing/on-demand/.

[6] Intel. "IT@Intel: Granulate™ Optimizes Memory and CPU Utilization in Intel IT's Cloudera Platform." Accessed October 2023.

[7] Prowess Consulting. "Improving Price and Performance with AWS® Spot-Based Instances." Sponsored by Intel. 2023.

1123/230190