aws

Q&A: Choosing the right compute infrastructure for machine learning

An expert viewpoint on infrastructure considerations for machine learning

Accelerate machine learning with low-cost, high-performance, ML-optimized infrastructure

Seemingly overnight, machine learning (ML) exited the world of aspirational technology and entered the mainstream. Organizations of every size and across nearly every industry want in on the action—and ML is realistically within reach for all because of the cloud. The cloud brings together data, low-cost storage, security, and ML services along with high-performance CPU- and GPU-based instances for model training and deployment. Now, organizations can store as much data and have as much high-performance compute as they need elastically, so realizing the value of ML can happen much faster.

However, with the emergence of a wide breadth and depth of cloud infrastructure options and services, it can be difficult to make the right selection for your use case. Many executives are asking, "What factors should I consider when choosing the right cloud compute infrastructure and services for my ML objectives?"

For the answers to that question and more, we turned to Dr. Bratin Saha, vice president of machine learning services at Amazon AI. Read on to discover guidance and best practices for evaluating the infrastructure requirements of your ML workloads—and for ensuring you make the right choices to meet those needs.

About the author

Dr. Bratin Saha is the vice president of machine learning services at Amazon AI. He is responsible for the ML and AI services (e.g., Amazon SageMaker, ML Engines, Amazon Personalize, Amazon Forecast, Health AI, AWS Panorama, Amazon SageMaker Edge Manager, Amazon CodeGuru, and others) that power AWS machine learning and built one of the fastest-growing services in AWS history. Dr. Saha is an alumnus of Harvard Business School (General Management Program), Yale University (PhD Computer Science), and Indian Institute of Technology (BS Computer Science). He has more than 70 patents granted (with another 50+ pending) and more than 30 papers in conferences/journals. Prior to Amazon, Dr. Saha worked at NVIDIA and Intel, leading different product groups spanning imaging, analytics, media processing, high-performance computing, machine learning, and software infrastructure.

Q

What are the biggest challenges developers and organizations are encountering in their ML workflows?

Dr. Bratin Saha: Let's take a step back and define ML. ML uses computational algorithms that learn from existing data through a process called training to make decisions about new data through a process called inference.

During training, patterns and relationships in the data are identified to build a model. The model allows you to make intelligent decisions about data that it hasn't encountered before.

Unfortunately, this process can be costly. There are two significant cost drivers. First, there's training the model. Training models is a data- and compute-intensive process. It requires high-performance CPU (central processing unit) and GPU (graphics processing unit) compute, as well as specialized tools to process and store terabytes of data. And companies that frequently train models typically have dedicated teams to manage the training environments because the volume of data needed for training grows significantly over time. These requirements have prevented some organizations from leveraging ML due to the limitations of their own data centers and resources.

Second, there's inference. In deep learning applications, such as computer vision and natural language processing (NLP), inference happens continuously and takes up a significant portion of the overall operational budget.[1] This is due to two factors. First, stand-alone GPU instances are typically designed for model training, not for inference, and stand-alone CPU instances are often too slow for deep learning inference. Second, different models have different CPU, GPU, and memory requirements, and optimizing for one resource can lead to underutilization of other resources and higher costs.

The good news is that many of these barriers can be overcome through cloud services.

[1] Amazon EC2 Inf1 Instances

Q

How has infrastructure evolved to meet today's ML needs?

Dr. Bratin Saha: Today's hardware uses accelerators to help meet the needs of compute-intensive applications, like high performance computing (HPC) and ML. But the use of hardware accelerators to speed up computations is not new.

In the early days of computing, FPUs (floating point units) were paired with CPUs to offload complex floating-point computations to a specially designed chip, freeing up CPU cycles for executing the applications. In the 1990s, CPUs shipped CPUs and FPUs together, making computers much faster and more efficient. Also, during this time, engineers and scientists demanded faster computers for data processing, modeling, and simulations. This meant there was a need for high-performance processors that could accelerate these workloads much faster than a CPU. GPUs emerged to fulfill that need. In the 2010s, deep learning and ML practitioners started using GPUs to accelerate deep learning training and inference.

Today, we have access to specialized hardware, known as ML accelerators, with custom-designed silicon built just for deep learning inference. These specialized processors—also called application-specific integrated circuits or ASICs—can be far more performant and cost-effective compared to general purpose processors (for workloads that are supported by the processor).

ML accelerators can be used to accelerate basic ML computations, reduce latency or time to train, and reduce the cost of training and deploying ML-based applications. They include GPUs that have thousands of processing cores, are highly effective at parallel execution, and have an architecture that is tolerant of memory latency.

A great example of such specialized processors is **AWS Inferentia**, an ASIC custom designed by AWS for accelerating deep learning inference. What's also great about our current time is you can access any amount or type of CPU or GPU in the cloud.

Q

What factors should organizations consider when selecting infrastructure services?

Dr. Bratin Saha: Choosing the right type of compute infrastructure for your ML workload depends on your application's performance and throughput needs, the type and size of your model, your choice of framework, and your budget.

Because ML use cases are so broad and expansive—spanning from forecasting to recommendations to computer vision—it's nearly impossible to find a single infrastructure option that delivers the best performance and lowest cost to address all of your needs. Further complicating matters is the fact that 70 percent of ML projects rely on ML frameworks such as TensorFlow, PyTorch, and Apache MXNet,[2] so infrastructure that supports and optimizes these frameworks is a must. An ML framework is any tool, interface, or library that lets you develop ML models easily, without the need to understand the underlying algorithms. Therefore, when selecting infrastructure, organizations should consider cost, performance, and support for ML frameworks.

[2] Amazon Machine Learning Infrastructure

Here are some specific questions to consider when evaluating infrastructure:

What is your target performance?

For training, this is the amount of time it takes to train your model; that is, how many computations per second you can deliver. For inference, this is both throughput—how many inference requests you are able to process per second—and round-trip latency—how quickly you are able to deliver results.

What are your cost barriers?

As model complexity increases, the cost to frequently retrain your model for greater accuracy in predictions will become prohibitive. When you choose the infrastructure to train or deploy your model, the amount of underlying resource that's being optimally utilized also ties into your cost evaluation.

What type of ML framework support do you need?

Most organizations rely on ML frameworks for machine learning, so picking infrastructure optimized for your unique model built on these frameworks is essential.

How easy is it to port code between different environments?

How easy is it to go from training the model to deploying it in production—and scaling it as time goes on? You don't want to be locked into infrastructure as your needs change and your business grows. Furthermore, it's a good idea to pick infrastructure solutions with large, active developer communities so you can benefit from the expertise and best practices of others.

How do you want to manage infrastructure?

As you build the right infrastructure for your ML project, you also need to consider how you want to manage it. ML infrastructure, as well as the gigabytes or terabytes often needed for training, must be managed in a secure, scalable, and cost effective way. The cloud offers almost limitless storage and scalable networking in a cost effective manner. You can use fully managed services, which automatically manage your infrastructure so you don't need to worry about hardware and software maintenance or applying security upgrades or do-it-yourself infrastructure management so you can still get the scale and security of the cloud while managing infrastructure in a more hands-on way.

How do you manage capital expenditures?

Oftentimes, resources used for ML model training are used sporadically; therefore, you will use more powerful compute during the training periods, but not every day. Using the cloud, you don't need to invest in all possible infrastructure options upfront. Instead, you can take advantage of elastic compute, storage, and networking, and the cloud is always up to date with the newest infrastructure innovations.

Q

How does AWS help address these customer questions?

Dr. Bratin Saha: **Amazon EC2 P3** and **Amazon EC2 P4d** instances are the most powerful NVIDIA GPU-powered instances for ML training that you'll find anywhere in the world.

They have networking throughput that's 3x as fast as anything else out there, twice as much GPU memory as anything out there, and a hundred plus gigabytes more of system memory than anything out there. Amazon EC2 P4d instances are also deployed in hyperscale clusters called Amazon EC2 UltraClusters that are comprised of thousands of GPUs, high-speed networking, and high-throughput, low-latency storage that enable you to run the most complex multi-node ML training with high efficiency.

We are rapidly investing in infrastructure to keep up with the growth of machine learning sophistication, introducing new chips and instances that help our customers keep the cost of training and inference down while speeding up their innovation. Let me share the latest innovations.

For training, we have **Amazon EC2 DL1** instances powered by Gaudi accelerators from Habana Labs, an Intel company. DL1 instances will offer 40 percent better price performance over current GPU-based EC2 instances for training deep learning models.

AWS also provides high-performance GPU instances such as **Amazon EC2 G4dn** and the industry's first instances featuring custom-built silicon for ML inference. We also launched AWS Inferentia-based instances. They provide the lowest cost per inference in the cloud, up to 70 percent lower cost and 2.3x higher throughput than comparable GPU-based instances. After migrating the vast majority of inferences to **Amazon EC2 Inf1**, the Amazon Alexa team saw 25 percent lower end-to-end latency for their speech-to-text workloads. And customers such as Snap, Autodesk, and Condé Nast use **Amazon EC2 Inf1** instances to get high-performance and low-cost ML inference. Condé Nast, for instance, observed a 72 percent reduction in cost than the previously deployed GPU instances. Software for the Amazon EC2 Inf1 instances includes AWS Neuron SDK, which is integrated with popular machine learning frameworks such as TensorFlow, PyTorch, and Apache MXNet.

You can continue to use the same ML frameworks you use today and migrate your software onto Amazon EC2 Inf1 with minimal code changes and without tie-in to vendor-specific solutions. Additionally, support for Hugging Face model repository provides customers the ability to compile and run inference using the pre-trained models—or even fine-tuned ones, easily, by changing just a single line of code.

Q

How does AWS optimize infrastructure for ML frameworks?

Dr. Bratin Saha: There are two options: **AWS Deep Learning Containers** (AWS DL Containers) and **AWS Deep Learning AMIs**, which are do-it-yourself services in the cloud.

AWS DL Containers are Docker images preinstalled with deep learning frameworks to make it easy to deploy custom ML environments quickly by letting you skip the complicated process of building and optimizing your environments from scratch. AWS DL Containers support TensorFlow, PyTorch, and Apache MXNet.

AWS Deep Learning AMIs provide ML practitioners and researchers with the infrastructure and tools to accelerate deep learning in the cloud at any scale. You can quickly launch Amazon EC2 instances preinstalled with popular deep learning frameworks and interfaces such as TensorFlow, PyTorch, Apache MXNet, Chainer, Gluon, Horovod, and Keras to train sophisticated, custom AI models, experiment with new algorithms, or learn new skills and techniques.

Whether you need Amazon EC2 GPU or CPU instances, there is no additional charge for the AWS Deep Learning AMIs—you pay for only the AWS resources needed to store and run your applications.

Q

What is the easiest way to access AWS infrastructures?

Dr. Bratin Saha: The easiest way to use any of the infrastructures we've described is through **Amazon SageMaker**, a fully managed service that helps you build, train, and deploy ML models.

When you're ready to train in Amazon SageMaker, simply specify the location of your data in (Amazon Simple Storage Service) Amazon S3, indicate the type and quantity of instance you need, and get started with a single click. Amazon SageMaker sets up a distributed compute cluster, performs the training, outputs the result to Amazon S3, and tears down the cluster when complete.

Amazon SageMaker makes it easy to deploy your trained model into production with a single click—so you can start generating predictions for real-time or batch data quickly. You can one-click deploy your model onto autoscaling Amazon ML instances across multiple availability zones for high redundancy. Amazon SageMaker will launch the instances, deploy your model, and set up the secure HTTPS endpoint for your application.

To help you get the most out of your ML infrastructure, Amazon SageMaker also offers software innovations. Many of the most common use cases for ML, such as personalization, require you to manage anywhere from a few hundred to hundreds of thousands of models. For example, taxi services train custom models based on each city's traffic patterns to predict rider wait times. While this approach leads to higher prediction accuracy, the downside is that the cost to deploy the models increases significantly because you have to use one endpoint per model. Amazon SageMaker multi-model endpoints allow you to deploy thousands of models behind a single endpoint, reducing cost by orders of magnitude.

If you want to get up to speed quickly on Amazon SageMaker, check out the new **Practical Data Science Specialization on Coursera**.

Q

How can organizations get started with ML?

Dr. Bratin Saha: With the broadest and deepest set of ML services, AWS can partner with you to provide all the support you need on your ML journey.

Through the **Amazon Machine Learning (ML) Solutions Lab**, we can help you identify your highest return-on-investment ML opportunities. We'll then "work backwards" from your business problems to create a prioritized roadmap of ML use cases and an implementation plan to address them.

We also offer a wide variety of **training courses and educational resources** that can help your business stakeholders develop a greater understanding of ML and help technical stakeholders build their ML skills. The **AWS Ramp-Up Guide: Machine Learning** provides a hierarchical catalog of all our ML training resources in a single, convenient document. There are options for free self-paced digital training, or your teams can dive deeper with public or private instructor-led classroom training, which is available virtually or in person.

For more comprehensive guidance on your ML journey, the **AWS Machine Learning Embark** program combines the training, coaching, and implementation support needed to launch your company's ML initiatives and transform your teams into ML practitioners.

In addition, our portfolio of Deep Devices—**AWS DeepRacer**, **AWS DeepComposer**, and **AWS DeepLens**—was designed to give developers hands-on experience with a fun and engaging way to learn ML.

Find the right infrastructure for your machine learning needs

Organizations have found new ways to leverage ML for recommendation engines, object detection, voice assistants, fraud detection, and more. Although the use of ML is gaining traction, training and deploying ML models can be expensive, model development can be time-consuming, and procuring the right amount of infrastructure to meet changing business conditions poses an ongoing challenge.

AWS ML software and infrastructure services accelerate your ML projects by offering high-performing, cost-effective, and highly flexible infrastructure—along with easy ways to use this hardware through fully managed services.

Discover why more than 100,000 customers choose AWS to accelerate their ML journeys. With the broadest and deepest set of ML services and supporting cloud infrastructure, the possibilities you can unlock with AWS machine learning are virtually endless. Start exploring them today.

Get started ›

aws