Understanding the costs of generative AI infrastructure

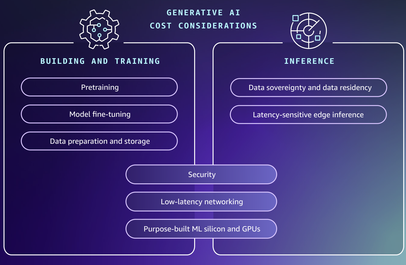
When selecting artificial intelligence (AI) infrastructure, the choices you make upfront can impact everything from cost to performance to sustainability goals to ease of use. Read on to discover factors that play a role in this selection and how you can best optimize for the price and performance you need.

Generative AI is transforming businesses

Powered by large-scale foundation models (FMs) that are trained on up to petabytes of data, generative AI can create new content and ideas, including conversations, stories, images, videos, music, and even software code, in response to a prompt. As these models grow, their parameters also increase—to upwards of trillions of parameters. Even smaller language models can be trained with a few billion parameters and can go up to 15 billion parameters. Organizations need an unprecedented level of infrastructure to build and train these models in a reasonable time and deploy them for inference.

Generative AI infrastructure costs vary over time

As you plan your generative AI projects, it's important to consider not just the upfront costs of building and training the model—but also the ongoing inference expenses that will vary depending on customer demands.



4 steps to optimizing generative AI price performance

By building with specialized AI infrastructure, you can innovate with higher performance while reducing costs.

1 Rightsize your model

You may not need the largest model. Pick the right type and size of model for your use case. View the infographic ***Selecting the Right LLM for Your Use Case*** to learn more.

2 Choose the optimal infrastructure

Develop an infrastructure strategy that includes the right processing power, low-latency networking, and scalable storage without compromising on cost or performance.

3 Optimize everything

Keep refining your deployment to maximize utilization of underlying resources by continuously evaluating models for accuracy, latency, and cost.

4 Reduce dev time with better tools

Utilize purpose-built tools for data labeling, cleansing, enrichment, error detection, and formatting that can help connect, access, and query data from a wide variety of popular data sources.

How AWS can help

Amazon Web Services (AWS) offers solutions that can help you achieve the four steps outlined above—all with minimal burden on your resources and maximum impact on your generative AI investments.

ACCELERATED COMPUTING

From the highest-performance NVIDIA GPU-based Amazon Elastic Compute Cloud (Amazon EC2) to continued investments in our purpose-built machine learning (ML) accelerators, AWS Trainium and AWS Inferentia, AWS delivers the best price performance for training and deploying generative AI models at scale.

AMAZON SAGEMAKER

Amazon SageMaker HyperPod provides a fully managed infrastructure and tools that include high-performance, cost-effective compute alongside integrated, purpose-built ML tools for the full AI workflow. Plus, with SageMaker, you can build, train, and deploy FM models at scale using tools like notebooks, debuggers, profilers, pipelines, MLOps, and more.

NETWORKING

Purpose-built to meet the performance demands for generative AI, AWS features high-throughput and low-latency networking, including Elastic Fabric Adapter (EFA) and Amazon EC2 UltraClusters.

STORAGE

Accelerate compute workloads with Amazon FsX for Lustre, which provides sub-millisecond latencies, up to hundreds of GBs/s of throughput, and millions of IOPS while quickly accessing and processing your datasets on Amazon Simple Storage Service (Amazon S3).

SECURITY

Our accelerated computing Amazon EC2 instances and networking are built on a foundation of the AWS Nitro System, which has been validated by the NCC Group, an independent cybersecurity firm. The level of security protection offered is so critical that we've added it to our AWS Service Terms to provide additional assurance to all of our customers.

Transform your business with generative AI infrastructure on AWS

Accelerate AI innovation with AWS. Learn more about the most comprehensive, secure, and price-performant AI infrastructure.

Explore generative AI infrastructure on AWS ›