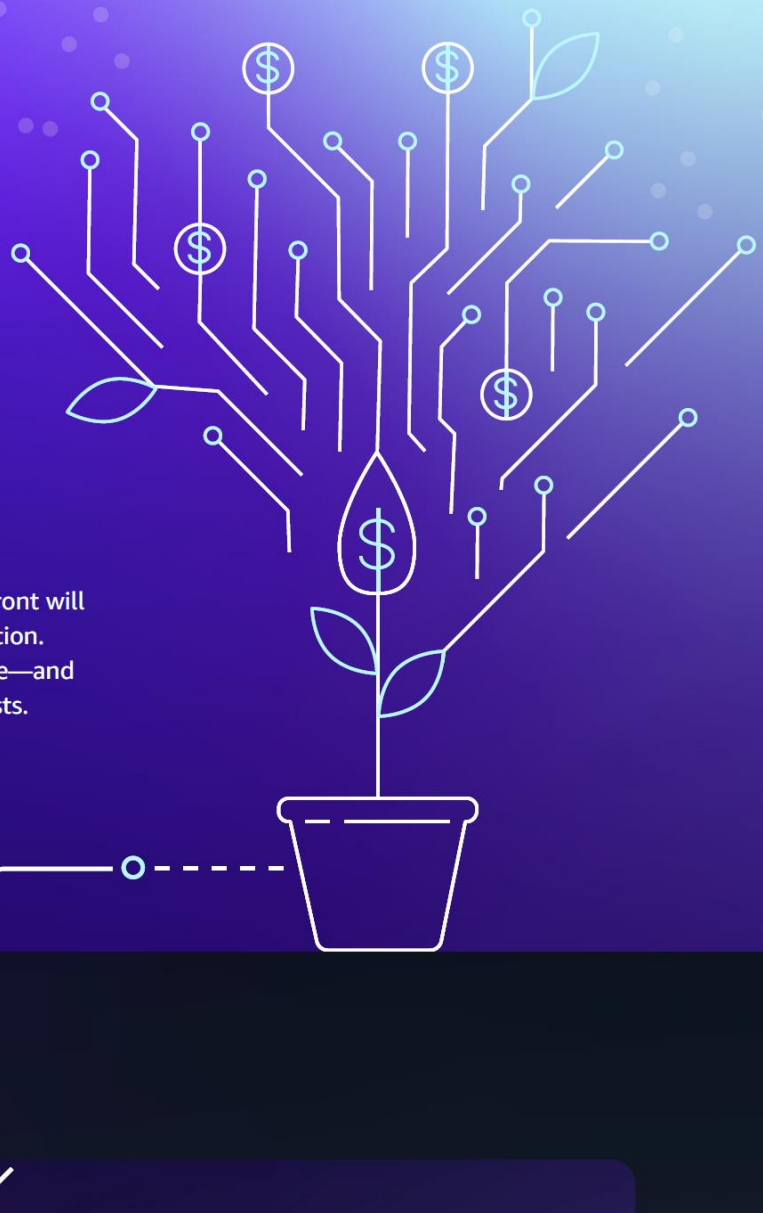
Understanding the costs of generative AI

When building with generative AI, the choices you make upfront will significantly impact the overall costs of your product or solution. Read on to discover which factors have the greatest influence—and how you can optimize them to control your generative AI costs.



What is generative AI?

Generative AI is a type of artificial intelligence (AI) that can create new content and ideas, including conversations, stories, images, videos, and music. It is powered by large models that are pretrained on vast amounts of data, commonly referred to as foundation models (FMs).

Generative AI costs expand over time

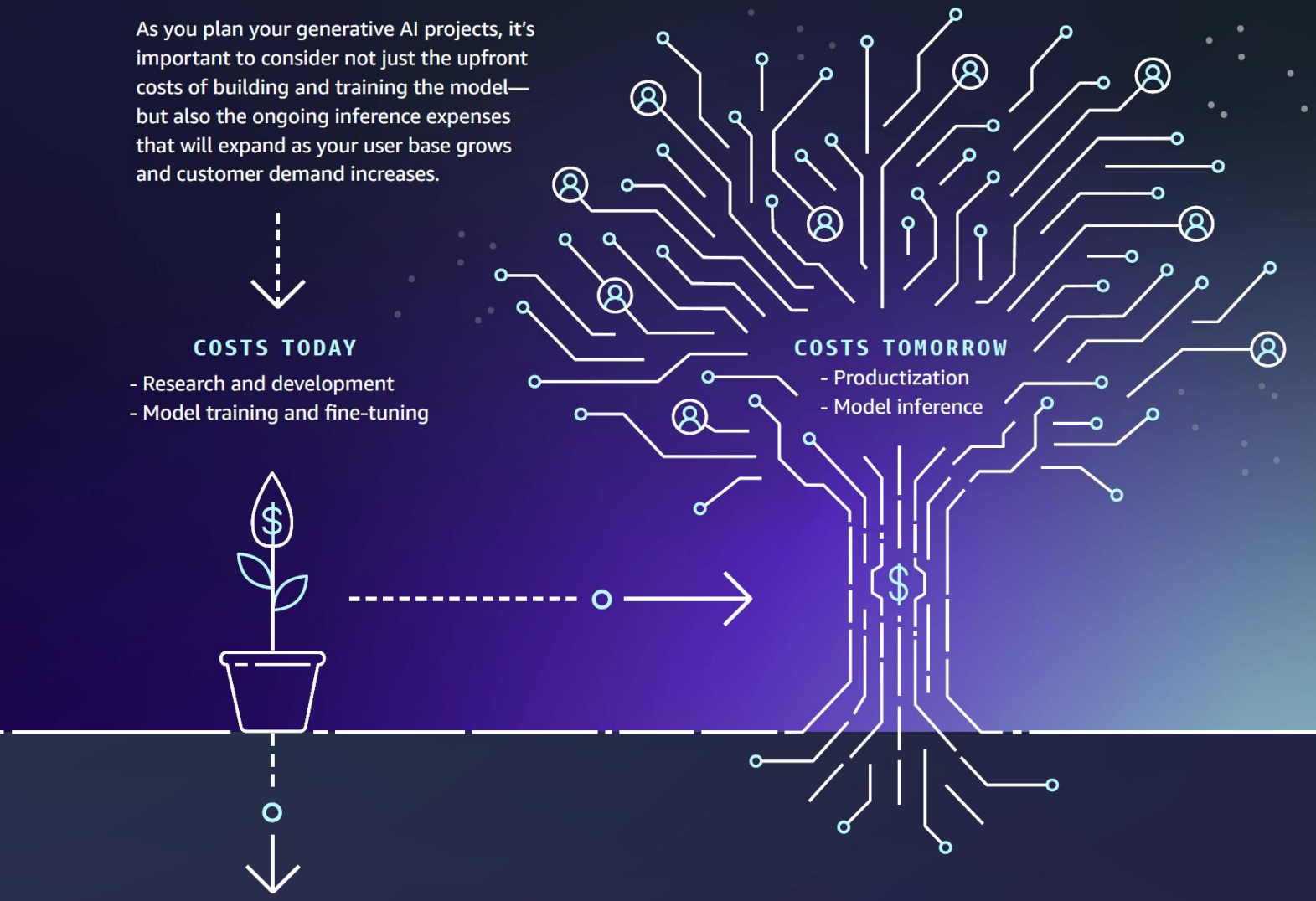
As you plan your generative AI projects, it's important to consider not just the upfront costs of building and training the model—but also the ongoing inference expenses that will expand as your user base grows and customer demand increases.

COSTS TODAY

- Research and development
- Model training and fine-tuning

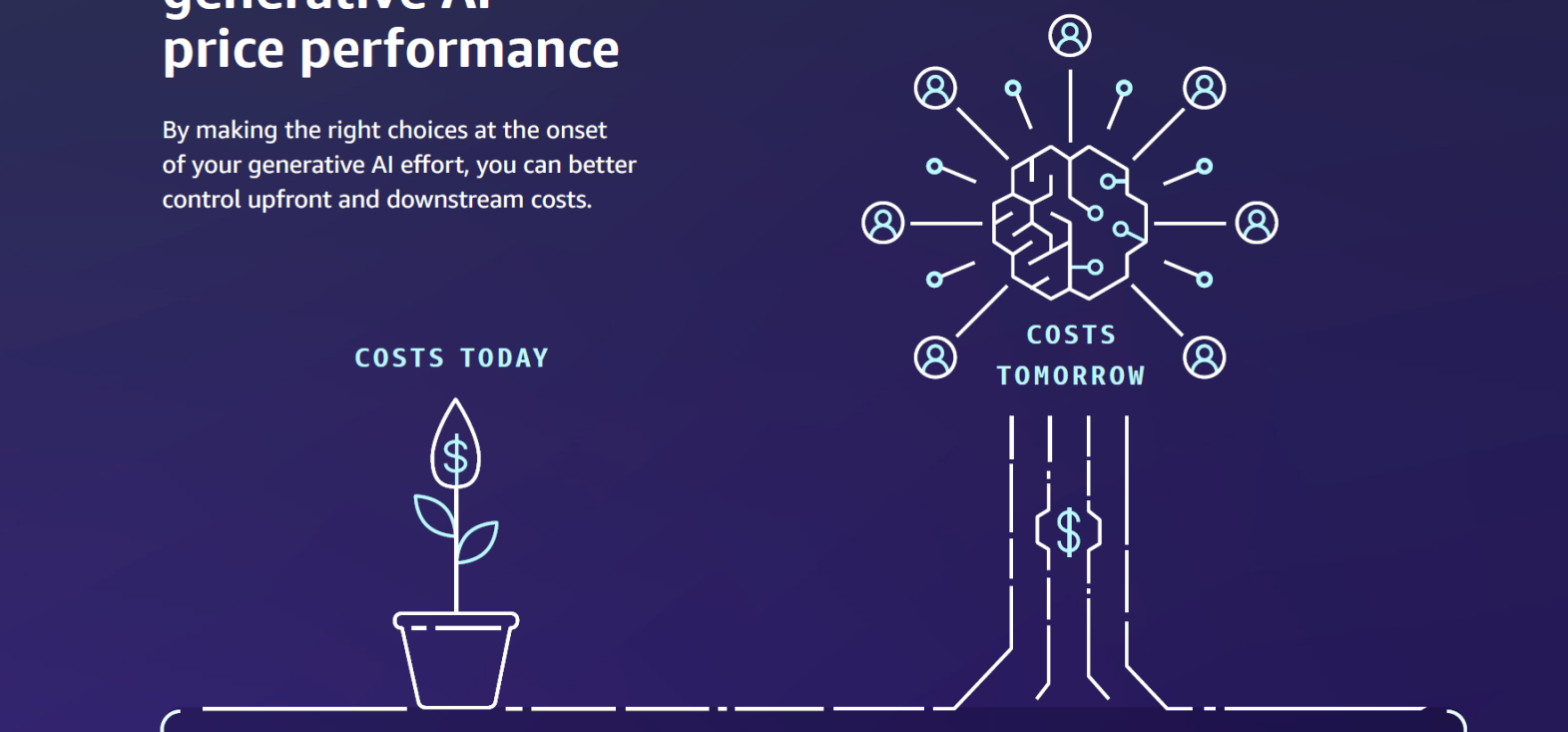
COSTS TOMORROW

- Productization
- Model inference



4 steps to optimizing generative AI price performance

By making the right choices at the onset of your generative AI effort, you can better control upfront and downstream costs.



1 → 2 → 3 → 4

1 Rightsize your model

You may not need the largest model. Pick the right type and size of model for your use case.

2 Choose the optimal infrastructure

Explore a broad set of GPUs and purpose-built accelerators to balance performance and costs.

3 Optimize everything

Keep refining your deployment to maximize utilization of underlying resources.

4 Reduce dev time with better tools

Manage your AI innovation, not your infrastructure.

How AWS can help

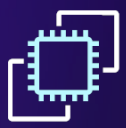
Amazon Web Services (AWS) offers solutions that can help you achieve the four steps outlined above—all with minimal burden on your resources and maximum impact on your generative AI investments.

Amazon Bedrock ›

Bedrock is the easiest way to build and scale generative AI–based applications using FMs.

Amazon SageMaker ›

The service that provides budget-friendly infrastructure, tools, and workflows for building, training, and deploying FMs.



AWS machine learning infrastructure ›

The service that provides high-performance, cost-effective infrastructure, tools, and workflows for building, training, and deploying FMs.

AWS Inferentia ›

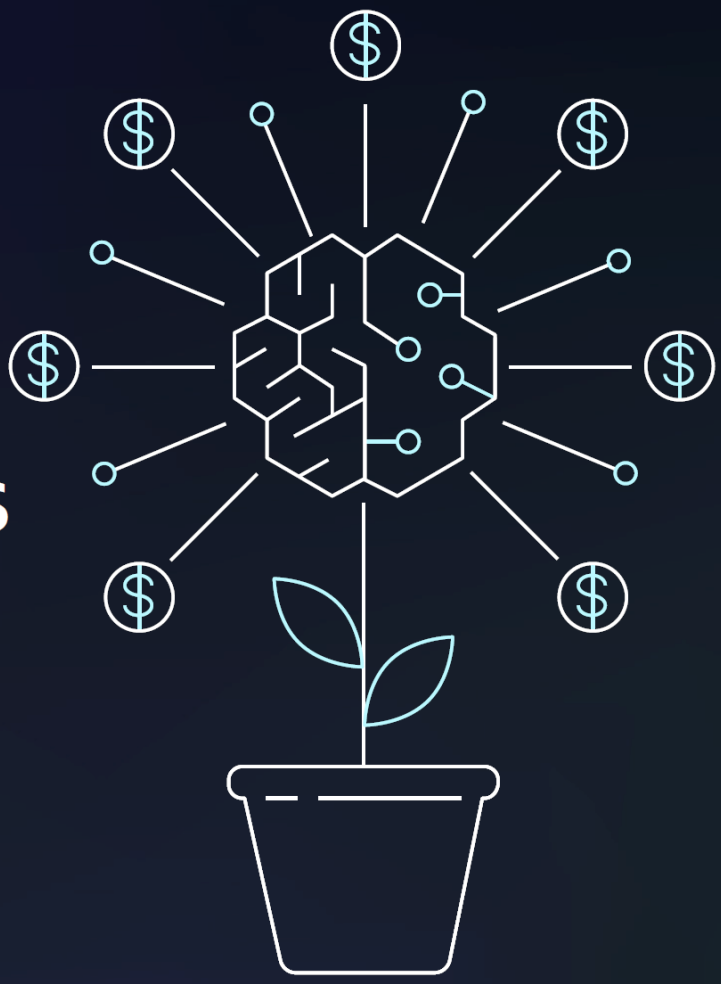
Amazon Elastic Compute Cloud (Amazon EC2) Inf2 instances deliver up to 40% lower cost per inference over comparable Amazon EC2 instances.

AWS Trainium ›

Amazon EC2 Trn1 instances deliver up to 50% in cost-to-train savings over comparable Amazon EC2 instances.

aws

Transform your business with generative AI on AWS



Understanding the costs of FM training and inference can help make generative AI more accessible—and more affordable. Learn how AWS is democratizing the technology for organizations of every size.

Explore generative AI on AWS ›