Increase your application performance at lower costs

7 ways to maximize performance and minimize costs

Table of contents

INTRODUCTION

Optimize costs with AWS

For any organization, moving to the cloud offers nearly unlimited opportunities for accelerating innovation, streamlining operations, and improving user experience. Among its many benefits—scalability, reliability, and security, to name a few—cloud migration also allows organizations to yield substantial cost savings in various ways. This is great news for infrastructure and operations leaders, who rank cost savings among their top priorities.[1]

Amazon Web Services (AWS) pioneered cloud computing in 2006, years before any other cloud provider. Since the beginning, we have helped customers migrate their workloads to the cloud offering more performance at lower costs than any other cloud. In fact, organizations that moved to AWS from on premises increased administrator productivity by an average of 66 percent and achieved an average of 20 percent cost savings on infrastructure.[2]

Helping you maximize savings so you can focus on innovating is at the core of what we do. For example, AWS relentlessly optimizes the efficiency of AWS services and has reduced priced 115 times since inception as to lower your costs of experimentation and accelerate your innovation. And we want to continue to help you optimize costs.

This eBook guides you through seven ways you can improve performance and reduce costs and shows you how AWS can help you optimize your infrastructure investment. You will also learn how leading organizations, such as Salesforce and Siemens, are using our differentiated solutions to achieve high performance at a lower cost for their workloads.

[1] "Gartner Leadership Vision for 2022: Top Strategic Priorities for IT Leaders," Gartner, 2021
[2] "The Business Value of Migration to Amazon Web Services," The Hackett Group, January 2022

METHOD #1

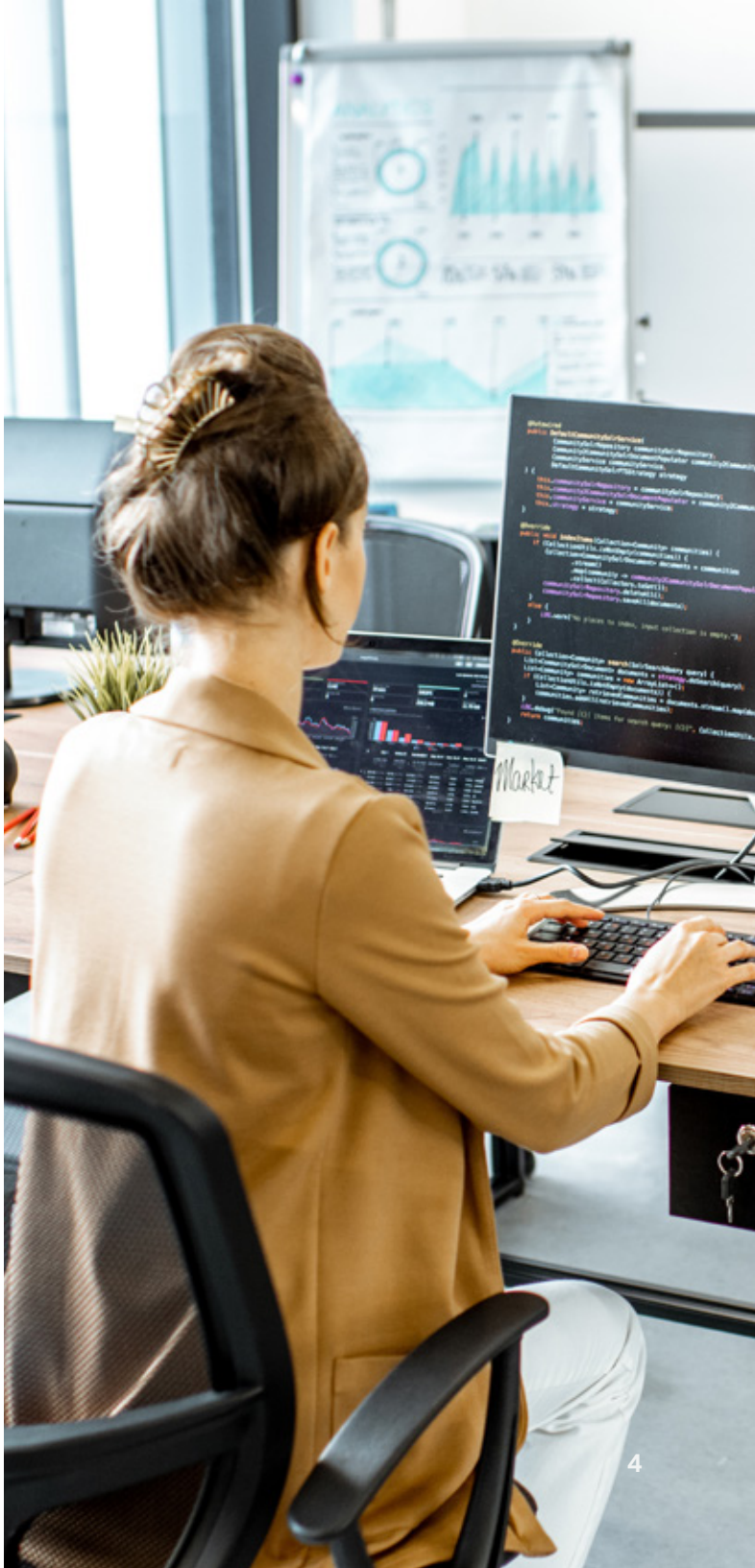
Build and run your applications on AWS

Where high performance meets lower cost

You can take advantage of better performance at lower costs just by using AWS. That's because AWS partners not only with major processor manufacturers to offer you the latest generation processors, but also, for nearly a decade, AWS has invested in designing and producing silicon optimized for the cloud. Our custom silicon enables us to offer you industry-leading performance, enhanced security, and faster innovation, all at lower costs. It includes our virtualization platform (AWS Nitro System), AWS-designed processor (AWS Graviton), and machine learning (ML) accelerators (AWS Inferentia and AWS Trainium).

Latest generation processors

AWS delivers state-of-the-art compute, storage, and networking technologies at scale, including the latest generation processors from partners such as Intel, AMD, Nvidia, and Apple, enabling you to innovate faster with higher performance at a lower cost. AWS has over 16 years of collaboration with Intel, including more than 400 Intel-based instances. AWS was also the first cloud provider to bring AMD-based processors to the cloud and now offers over 100 AMD-based instances. This partnership includes the latest, third generation processor technologies.



AWS-designed virtualization platform

The **AWS Nitro System** is the underlying platform for our modern **Amazon Elastic Compute Cloud** (Amazon EC2) instances. The AWS Nitro System enables us to deliver more performance, further reduce costs for you, and provide added benefits such as increased security and new instance types. The AWS Nitro System and Amazon EC2 instances can deliver more than 15 percent higher throughput performance on some workloads as compared to other major cloud providers running the same CPU.

Custom-built processor

AWS Graviton processors are designed by AWS to deliver the best price performance for your cloud workloads running on Amazon EC2—they deliver up to 40 percent better price performance over comparable current-generation x86-based instances for a broad spectrum of workloads.

Purpose-built machine learning accelerators

AWS Inferentia is the first ML accelerator designed and purpose-built by AWS to accelerate deep learning inference. Amazon EC2 Inf1 instances, powered by AWS Inferentia, deliver up to 2.3 times higher throughput and up to 70 percent lower cost per inference than comparable GPU-based instances.

AWS Trainium is an ML accelerator that AWS purpose-built for high-performance, low-cost deep learning training. Amazon EC2 Trn1 instances, powered by AWS Trainium chips, offer up to 50 percent cost-to-train savings over comparable GPU-based EC2 instances. Trn1 instances are the first EC2 instances with up to 800 Gbps of Elastic Fabric Adapter (EFA) network bandwidth. They are deployed in EC2 UltraClusters that enable scaling up to 30,000 AWS Trainium accelerators, which are interconnected with a non-blocking petabit-scale network to provide up to 6.3 exaflops of compute.

AWS “firsts” – innovation for your digital business

- First to offer AMD processors in the cloud
- First to offer Intel’s latest, third-generation Intel Xeon Scalable processors
- First to offer Arm-based processors
- First and only major cloud provider to offer on-demand macOS-based instances
- First to offer DDR5 memory in the cloud

METHOD #2

Optimize costs and accelerate innovation with serverless computing

Serverless computing allows you to build and run applications and services without thinking about servers—so you can focus on building applications instead of configuring them. It eliminates infrastructure management tasks such as server or cluster provisioning, patching, operating-system maintenance, and capacity provisioning.

Serverless computing on AWS is another way to significantly reduce your overall infrastructure costs. With our pay-for-value billing model, resource utilization is automatically optimized, and you never pay for over-provisioning. And with faster-than-ever deployments and updates, you can deliver better-than-ever user experiences.

You can get started on serverless by migrating your workloads to AWS Lambda or AWS Fargate:

AWS Lambda is an event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers. You can trigger Lambda from more than 200 AWS services and software-as-a-service (SaaS) applications and only pay for what you use, resulting in 34 percent better price performance for your application compared to x86-based EC2 instances.

AWS Fargate is a pay-as-you-go compute engine that lets you focus on building applications without managing servers. Fargate is compatible with both **Amazon Elastic Container Service** (Amazon ECS) and **Amazon Elastic Kubernetes Service** (Amazon EKS).

If serverless computing is not right for your workload, consider Amazon EC2 for secure, reliable, and cost-saving compute capacity.

With serverless computing on AWS, you can:

- Build and iterate quickly so you can go to market faster
- Release features and updates fast—in hours instead of days—to keep users engaged
- Scale automatically to accommodate changes in demand
- Automate security and compliance processes
- Lower total cost of ownership (TCO) by requiring fewer resources
- Minimize unplanned downtime and security risks

METHOD #3

Choose the instance type that matches your application needs and budget

AWS has more than 600 instance types, exceeding any other cloud provider. Each instance type provides a choice of processor, storage, networking, and operating system, so you can choose the instance configuration that best fits your specific workload. And each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

AWS gives you the flexibility to change your instance type as quickly as your needs change, eliminating overhead costs for unused resources. Amazon EC2 instances fall into six categories:

General purpose

Our most popular instances provide a balance of CPU, memory, and network resources and are ideal for running web servers, containerized microservices, caching fleets, and development environments. One of the main distinctions within this class is between instances with fixed (e.g., **M5a**) versus burstable (e.g., **T4g**) performance.

Compute-optimized

Good for compute-intensive, CPU-bound, demanding applications such as frontend fleets for high-traffic websites, on-demand batch processing, distributed analytics, video encoding, dedicated gaming servers, and high-performance science and engineering applications. These instances offer the highest ratio of virtual CPUs to memory than the other families and the lowest cost per virtual CPU of all the EC2 instance types.

Memory-optimized

These instances are ideal for memory-intensive applications, such as real-time big data analytics, in-memory databases, enterprise-class applications that require significant memory resources, or general analytics, such as Hadoop or Spark.

Accelerated computing

Instances in this category include additional accelerators, as well as GPUs, FPGAs, and ML chips that provide massive amounts of parallel processing for tasks such as graphics processing, ML training, ML inference, and high performance computing (HPC).

Storage-optimized

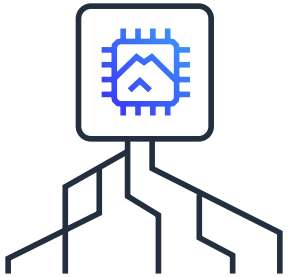
Ideal for tasks that require local access to **very large amounts of storage**, **extreme storage performance**, or **both**. Instances are available that include both large-capacity HDD and extreme low-latency local NVMe SSDs. Choose from the industry's broadest portfolio of storage solutions, optimized for your block, file, and object data.

HPC-optimized

HPC instances are purpose built to offer the best price performance for running HPC workloads at scale on AWS. HPC instances are ideal for applications that benefit from high-performance processors such as large, complex simulations and deep learning workloads.

METHOD #4

Migrate to AWS Graviton for the best price performance for a broad set of applications



AWS Graviton processors are designed by AWS to deliver the best performance at the lowest cost for your cloud workloads running on Amazon EC2. AWS Graviton-based instances deliver up to 40 percent better performance at lower costs versus comparable x86-based EC2 instances. They are also highly energy efficient, using up to 60 percent less energy for the same performance than comparable Amazon EC2 instances. AWS Graviton is available in more than 25 regions, and migrating to AWS Graviton can help you increase performance, reduce costs, lower latency, and achieve better scalability.

The **AWS Graviton Fast Start** program helps you quickly and easily move your workloads to AWS Graviton in as few as four hours. Or, accelerate your adoption of AWS Graviton with the help of **AWS Graviton Partners**.

Graviton processors

AWS Graviton2 processors give customers up to 40 percent better price performance for a broad range of workloads versus comparable x86 processors from other providers.

The latest in the AWS Graviton processor family, AWS Graviton3 processors provide up to 25 percent better compute performance, up to two times higher floating-point performance, and up to two times faster cryptographic workload performance compared to AWS Graviton2 processors. AWS Graviton3 processors deliver up to three times better performance compared to AWS Graviton2 processors for ML workloads, including support for bfloat16. They also support DDR5 memory, which provides 50 percent more memory bandwidth compared to DDR4.

AWS Graviton-powered managed services

AWS Graviton-based instances are also available in more than 25 popular managed AWS services. These services deliver the price performance benefits of AWS Graviton processors while providing a fully managed experience. AWS Managed Services using AWS Graviton include serverless solutions, such as AWS Lambda and AWS Fargate, and AWS Graviton–based databases, such as Amazon Aurora, Amazon Relational Database Service (Amazon RDS), and Amazon ElastiCache.

METHOD #5

Select the compute purchase models that best fit your budget

AWS offers you a choice of flexible, cost-effective purchase models to meet your infrastructure needs while keeping you within your budget:

On-Demand Instances

On-Demand Instances let you pay for compute capacity by the hour or second, depending on which instances you run. No long-term commitments or upfront payments are needed. On-Demand Instances are ideal for applications with short-term, spiky, or unpredictable workloads that cannot be interrupted and applications being developed or tested on Amazon EC2 for the first time.

Savings Plans

Savings Plans consist of flexible pricing models that can help you reduce your bill by up to 72 percent compared to On-demand prices in exchange for a one- or three-year hourly spend commitment. AWS offers three types of plans: Compute Savings Plans, EC2 Instance Savings Plans, and Amazon SageMaker Savings Plans.

Compute Savings Plans apply to usage across Amazon EC2, AWS Lambda, and AWS Fargate. The EC2 Instance Savings Plans apply to EC2 usage, and the Amazon SageMaker Savings Plans apply to Amazon SageMaker usage. Savings Plans automatically and simultaneously apply to eligible AWS usage and enable you to innovate faster by leveraging the newest instances, families, generations, and regions while staying on the same plan. Since the launch of Savings Plans in 2019, customers have saved more than $15 billion.

Amazon EC2 Spot Instances

Amazon EC2 Spot Instances let you take advantage of unused EC2 capacity in the AWS Cloud. Spot Instances are available at up to a 90 percent discount compared to On-Demand prices. You can use Spot Instances for various stateless, fault-tolerant, or flexible applications, such as big data, containerized workloads, continuous integration and continuous delivery (CI/CD), web servers, HPC, and test and development workloads. Moreover, you can easily combine Spot Instances with other purchase models, giving you the flexibility to grow and change over time while still getting the lowest cost available on AWS. Since 2015, Spot Instances have helped our customers save more than $8 billion.

$15B+

Since 2019, customers have saved $15 billion with Savings Plans

$8B+

Since 2015, customers have saved $8 billion with Amazon EC2 Spot Instances

METHOD #6

Optimize your resource capacity to fit demand

AWS Compute Optimizer and AWS Auto Scaling allow you to provision with precision. These two services help lower your costs and respond to changes in demand, and they're free to use.

AWS Compute Optimizer

Over-provisioning resources can lead to unnecessary infrastructure costs, while under-provisioning can lead to poor application performance. Using ML to analyze historical utilization metrics, **AWS Compute Optimizer** recommends the optimal AWS resources for your workloads, further reducing your infrastructure costs. With just a few clicks, AWS Compute Optimizer automatically generates recommendations based on current utilization data, eliminating the need to invest time and resources to set up rule-based thresholds. Since its launch in December 2019, AWS Compute Optimizer has generated recommendations for more than 80 percent of Amazon EC2 usage and provided more than 10 billion recommendations, resulting in reduced costs and improved performance for a variety of workloads.

AWS Auto Scaling

AWS Auto Scaling monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. You can easily set up application scaling for multiple resources across multiple services with a single intuitive interface and maintain optimal application performance and availability even when workloads are periodic, unpredictable, or continuously changing. When demand spikes, AWS Auto Scaling automatically increases the capacity of constrained resources, so you can maintain a high quality of service. When demand drops, it removes any excess resource capacity to help keep you from overspending. There's no need to add AWS Auto Scaling as a separate tool—it is already built into AWS solutions.

10B+

AWS Compute Optimizer has provided more than 10 billion recommendations since launch, resulting in reduced costs and improved performance for a variety of workloads.

METHOD #7

Optimize your workload costs with AWS Storage

Minimize your TCO with AWS storage services that eliminate on-premises capital equipment investment, management complexity, and infrastructure maintenance. With AWS Storage, you get the right mix of price and performance for your workloads, and you pay only for the storage that you use. And AWS gives you the broadest array of storage services of any cloud provider. This includes more ways to optimize storage costs through a choice of storage classes and intelligently tier data to lower cost storage and data reduction capabilities, such as compression and data deduplication.

For block-based workloads, **Amazon Elastic Block Store** (Amazon EBS) provides easy-to-use, high-performance block storage at any scale. You select the storage that best fits your workload, service level, and budget. EBS scales fast for your most demanding, high-performance workloads, including SAP, Oracle, and Microsoft products. **Amazon EBS Snapshots** provide a solution to maintain compliance and further reduce snapshot storage costs by up to 75 percent by using **Amazon Data Lifecycle Manager** to automatically move seldom-used snapshots to the **Amazon EBS Snapshots Archive**.

Amazon Simple Storage Service (Amazon S3) is the lowest-cost object storage in the cloud. No matter the size of your organization, Amazon S3 lets you store and protect any amount of data for virtually any use case, such as data lakes, cloud-native applications, and mobile apps. **Amazon S3 storage** classes provide the lowest-cost storage for specific access patterns, and **Amazon S3 Intelligent-Tiering** has saved customers more than $750 million in storage costs compared to Amazon S3 Standard since the launch of S3 Intelligent-Tiering in 2018. With cost-effective storage classes and easy-to-use management features, you can optimize costs, organize data, and configure fine-tuned access controls to meet specific business, organizational, and compliance requirements.

AWS offers the industry's widest portfolio of fully managed file storage. This means that we handle all of the infrastructure–provisioning, patching, and backups allowing you to choose the right file system technology to meet your workload requirements. **Amazon Elastic File System** (Amazon EFS) is a serverless elastic file system that scales automatically as files are added, removed, and burst to higher throughput levels when necessary. You can also reduce costs by up to 92 percent by automatically tiering infrequently accessed files. The Amazon FSx file system family includes **Amazon FSx for NetApp ONTAP**, **Amazon FSx for OpenZFS**, **Amazon FSx for Windows File Server**, and **Amazon FSx for Lustre**. Amazon FSx enables you to optimize your price and performance to support a broad spectrum of use cases, from small user shares to the most demanding compute-intensive workloads. Amazon FSx file systems support a rich set of storage efficiency features, including data deduplication, compression, and usage quotas.

Cost savings with AWS Storage

75%

Amazon EBS Snapshots Archive reduces snapshots storage costs by up to 75 percent

$750M

Amazon S3 Intelligent-Tiering has saved customers more than $750 million in storage costs compared to Amazon S3 Standard since the launch of S3 Intelligent-Tiering in 2018.

Amazon S3

Amazon S3 Glacier storage classes are purpose-built for data archiving, providing you with the highest performance, most retrieval flexibility, and the lowest-cost archive storage in the cloud

92%

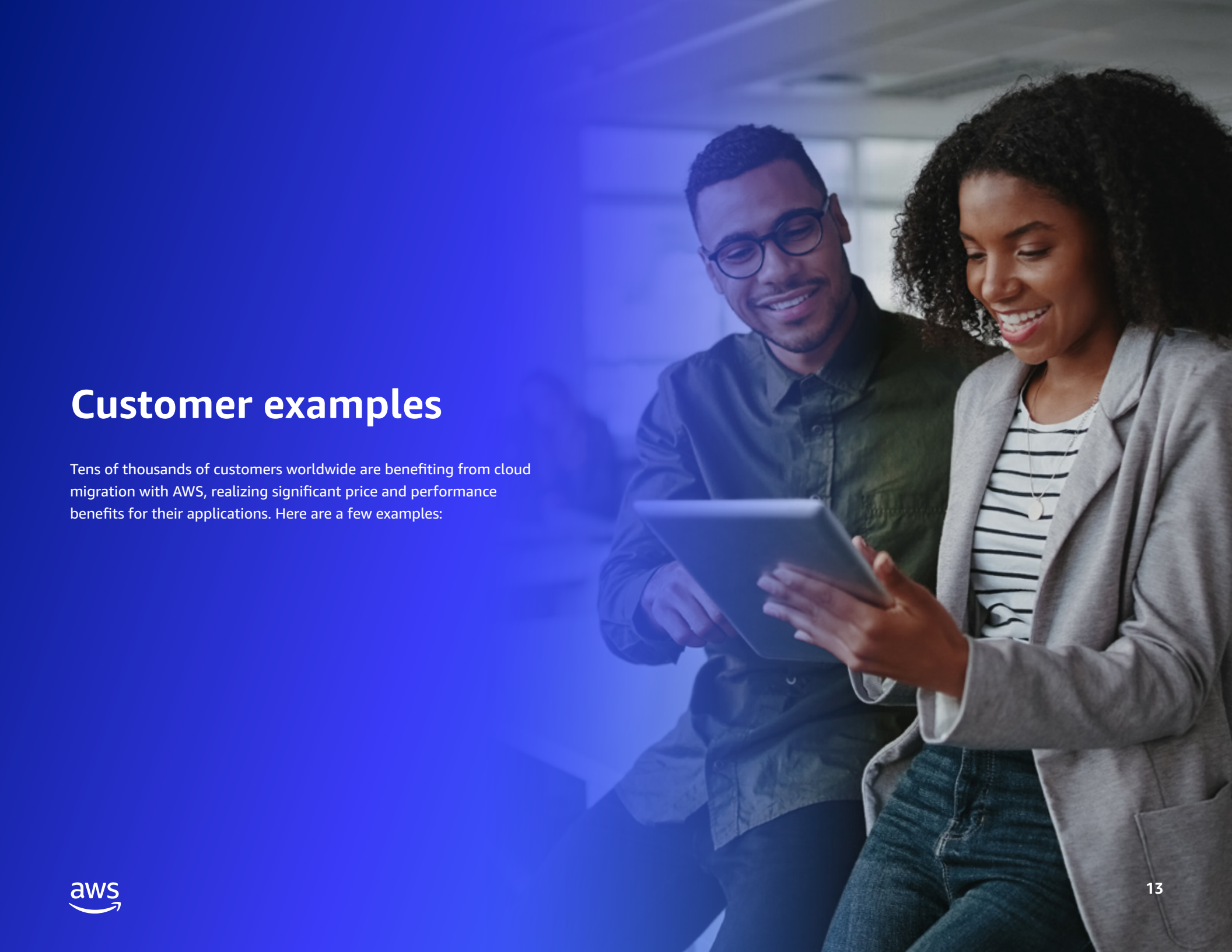
Amazon EFS Intelligent-Tiering reduces storage costs by 92 percent

50%+

Amazon FSx family data reduction technology reduces storage costs by 50–65 percent

Customer examples

Tens of thousands of customers worldwide are benefiting from cloud migration with AWS, realizing significant price and performance benefits for their applications. Here are a few examples:



Salesforce slashes processing times by 90 percent and saves $1 million monthly with AWS

Using AWS for a mix of instance-provisioning models from Amazon EC2, the Salesforce team was able to build a scalable elastic compute infrastructure. With its remodeled infrastructure, it takes the company less time to process twice as much data while lowering compute costs by more than 60 percent, saving the company more than $1 million a month.

“We use the capacity of the cloud and the wide range of Amazon EC2 instance types to do things we couldn’t do on premises. Amazon EMR Managed Scaling plays a big part in our ability to use the elastic capability of the cloud. And we significantly reduce costs just by using Spot Instances in an innovative way.”

Eric Legault, Principal Engineer, Salesforce

Read the story ›

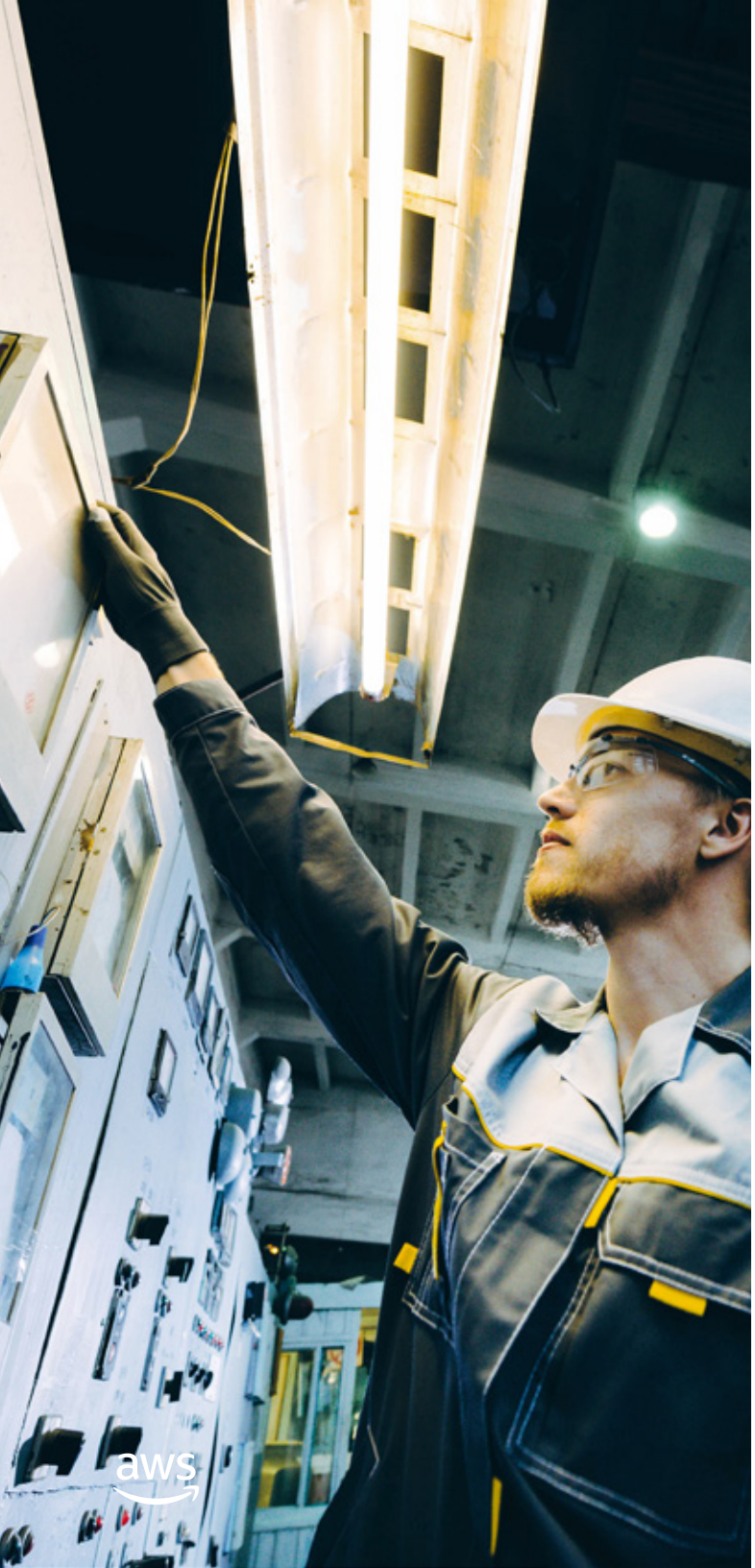
FORMULA 1: Making races more exciting while lowering costs by 30 percent with AWS

FORMULA 1 (F1) used a combination of Amazon EC2 instances to reduce its computational fluid dynamics (CFD) simulation time by 80 percent and lower the cost of running workloads by 30 percent. As a result, F1 can better support its strategic priorities of increasing competitiveness and unpredictability on the track and producing a world-class spectacle for fans.

Read the story ›

“We can now run huge models with more than half a billion cells. This is only possible because of AWS.”

Pat Symonds, Chief Technology Officer, FORMULA 1



Siemens: Reducing infrastructure costs by 85 percent

Large volumes of sensor data can result in “alert overload”—for a Siemens power plant, about 5,000 control-system alerts per day. Reducing and prioritizing them can tie up two full-time employees for six months. With a serverless platform on AWS that includes Amazon S3 and Amazon Simple Queue Service (Amazon SQS), Siemens was able to decrease alerts by 90 percent while substantially reducing infrastructure costs.

“In the 18 months since we’ve been serverless, we haven’t had one minute of unplanned downtime.”

Stefan Lichtenberger, Program Manager, Siemens Gas and Power

Read the story ›

CONCLUSION

Start maximizing your savings now

With innovations in silicon and serverless technologies and a variety of flexible pricing options, AWS is committed to helping you get the most from your cloud infrastructure spend. We provide you the capability to optimize your costs while building modern, scalable applications to meet your needs. Our breadth of services and pricing models demonstrate our commitment to giving you all the compute performance and capacity you require at the lowest cost—now and as your business evolves.

Start improving your application performance today and reach your highest savings potential with AWS.

Take advantage of AWS Free Tier and get free, hands-on experience with AWS services ›