NEXT-GENERATION INFRASTRUCTURE

Global-scale solutions

Bring new applications to market while modernizing existing applications to improve functionality, scale effectively, and accommodate a global audience

Table of contents

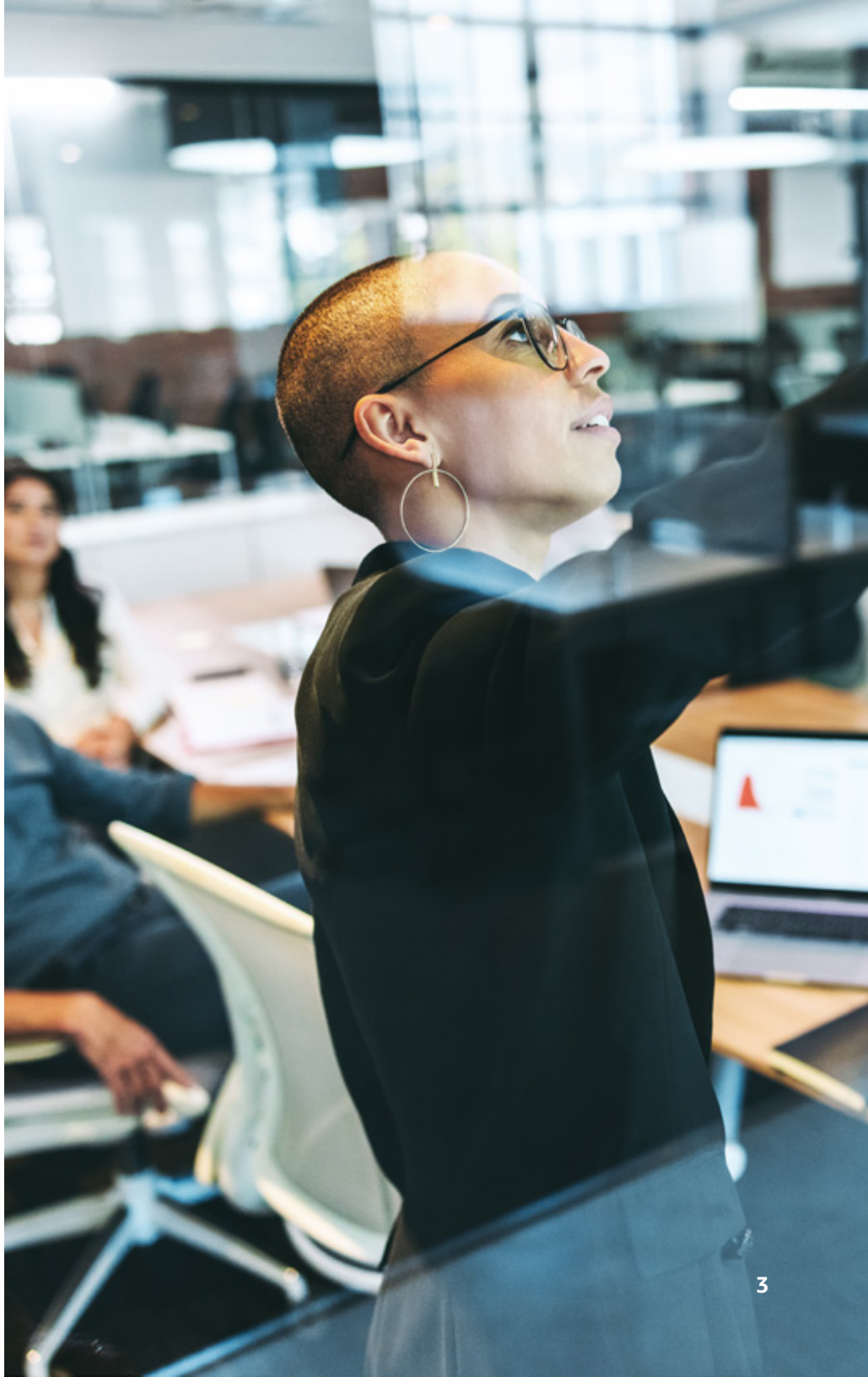
The challenge

Modern application development

Businesses are continuously working to differentiate themselves by bringing new applications to market while modernizing existing applications to improve features, scale effectively, and provide a consistent experience to users wherever they are. To keep up with user expectations, businesses need to get this work done quickly while continuing to build applications that are reliable and secure.

Today's application users have a high standard for engaging and interactive experiences and more options than ever to choose from. Exceeding these expectations is only possible when building upon a scalable, reliable, and globally available infrastructure that can support highly available applications and scale to users worldwide with improved performance.

Read this eBook to review some of the challenges AWS is seeing businesses face today and explore how you may be able to plan for and solve these challenges for your own business with infrastructure and solutions provided by Amazon Web Services (AWS).



Scale

You need to develop your applications with scale in mind, especially during growth phases, such as when you enter new markets or expect an increase in the number of users that interact with your applications. There are generally two scenarios to consider when planning for scale:

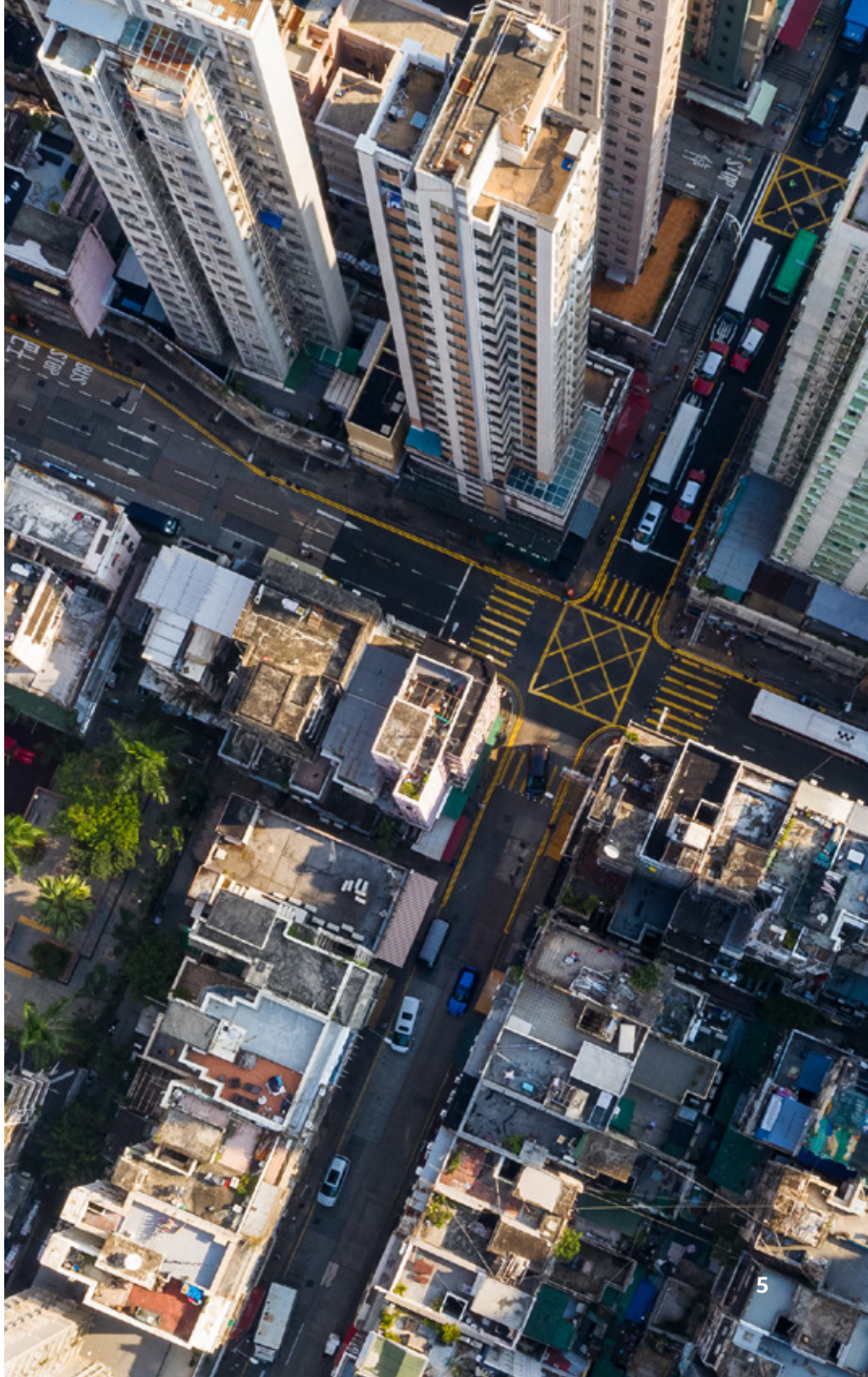
- **Short-term traffic spikes:** Modern-day applications are bursty by nature. Whether we're talking about websites, APIs, or streaming applications, the demand on your application is likely to fluctuate depending on a variety of different circumstances—often unpredictable in nature. These unanticipated spikes in demand can result in dozens or even hundreds of times the normal load on your application's compute, storage, and networking infrastructure. Hitting a scale limit in your architecture could impact your ability to grow your business. However, with the right architecture, your applications can be built to handle today's needs and the growth you'll experience tomorrow.
- **Long-term growth:** As your business grows over time, you need to be prepared to scale out in new directions, without painting yourself into a corner. While your current infrastructure may meet your current needs, it may not be able to provide the same experience to users in new markets. Similarly, as traffic grows, you may find that costs soar and performance deteriorates due to architectural limitations that inefficiently utilize your servers. Given the right selection, you can alleviate these challenges by having the right infrastructure available to run your applications where and when you need them. To achieve this, long-term growth plans need to consider not only how you can scale your existing architecture but also how you can optimize your costs, performance, reliability, and security as your demand and application types evolve.

If you don't plan out your application architecture with the right infrastructure choices to handle both short-term and longer-term scale challenges, the impact can be significant. Exceeding the capacity your architecture was designed for can affect your application, leading to slow performance, unhappy customers, and extra maintenance. Even if your response team prevents your application from going down, identifying the issue and unexpectedly provisioning new resources is costly and time-consuming.

Working with on-premises infrastructure

Traditional scaling of on-premises infrastructure continues to prove costly and time-consuming. Building new infrastructure is difficult, time-consuming, and expensive. Purchasing land, obtaining permits, and building new data centers require you to plan years ahead. These large investments require large fixed costs that limit your company's adaptability as needs shift. Even if you can staff a team of experts and plan future capacity successfully, you are still left devoting a large amount of time and resources to infrastructure rather than product quality and what makes you stand out from your competitors. In addition, the equipment you purchase could be easily obsolete before you've recovered the costs of your new data center.

While some companies still pursue this method due to sunk costs or regulatory requirements, building their own physical locations leads to a challenging set of problems. You should consider these challenges when planning out your application infrastructure.



New considerations for today's applications

Digital transformation is changing the way businesses and industries operate, delivering unparalleled efficiencies and innovation. Leveraging technology advancements, like the cloud and the Internet of Things (IOT), organizations are now using a highly diverse digital platform to drive their businesses forward.

Whether you are doing analytics or machine learning or distributing video globally, building applications today comes with a new set of challenges. Today's applications are data-intensive, they run at a larger scale with higher throughput, and need to connect to more users. Meanwhile, application users have higher expectations than ever. The security landscape is also evolving as threats become more common and complex and botnets enable DDoS, phishing, and spambots to execute at a massive scale.

If you run your applications on infrastructure (both physical and software) with a one-size-fits-all (or one-size-fits-most) approach, you will create an inefficiency that can greatly impact your costs and performance at scale. Applications with Windows-based workloads, enterprise database workloads, or even Mac workloads each have their own characteristics around CPU, memory, storage, and networking. Running on a less-than-optimal workload infrastructure is an expensive liability that opens the door for unplanned downtime, slow deployments, and poor performance.

Performance matters,s and user expectations keep getting higher. In 2006, the average online shopper expected a webpage to load in four seconds (Forrester), while by 2017, a one-second delay was reducing page conversion by 20 percent, and 53 percent of mobile visitors were abandoning a page that took more than three seconds to load (**Google**). Today, with 5G and faster devices, users expect imperceivably small load times. To meet this demand, you need to build an application with low latency in mind at every level: low-latency application servers, a performance-optimized network, and the shortest practical distance between your applications and your users. This means building a globally dispersed application architecture with optimized network paths and prioritizing large metro areas with higher demand.

Security considerations have become more complex as network and application architecture has evolved and become complicated to include on premises and hybrid cloud. This complexity also includes an increasingly distributed workforce (WFH, client machines, mobile devices) and app modernization (containers/microservices at a global scale). IT and security teams responsible for maintaining this modern and evolving architecture are looking for a holistic solution and an integrated defense strategy.

When building an integrated defense and in-depth security strategy, there are a few critical things to take into consideration. First, you need to identify the organization's physical and logical network boundaries; second, you need to define security and business policies for users, partners, and all distributed business functions inside the organization that touch data. Third, you need to consider analyzing the data as it traverses through your workflow to identify risks and put in place mitigation strategies at each access delivery point to prevent possible data theft. Finally, you need to ensure the analyzed data is then routed to the designated business function securely. To stay ahead of potential threats, you need to design processes, platforms, and IT infrastructure with **security in mind**.

To open the first **AWS re:Invent keynote** in 2012, News International talked about the value that tapping into scalable infrastructure had for their business. To this day, the core value proposition of AWS infrastructure remains true—an easy-to-use, secure, and scalable platform that provides reliable and cost-effective solutions to enable innovation. While applications have become more complex and demanding, AWS remains customer obsessed, improving availability, scalability, and cost performance while bringing new services and features to market. For these reasons, more companies are working with AWS to build cloud-native and run their existing applications than ever before.

The solution

An extensive and reliable cloud platform

AWS has over 200 fully featured services spanning 245 countries and territories. With AWS, you can run and scale your applications wherever your customers need them, and you can leverage AWS innovations wherever your business needs them.

With millions of active customers and tens of thousands of partners, AWS supports the largest cloud ecosystem in the world. Customers across virtually every industry and of every size—including startups, enterprises, and public sector organizations—are running on AWS to meet the needs of every imaginable use case.

AWS provides scalable cloud infrastructure for any requirement:

- Enable general purpose applications on AWS, such as web and application servers
- Power compute-intensive applications, such as gaming, machine learning, and high-performance computing
- Run memory-intensive applications, like databases and in-memory caches
- Deploy hybrid and edge workloads operating on Linux-, SQL-, SAP-, Mac-, and Windows-based applications

Infrastructure by AWS: Global, Available, Scalable

AWS is the most secure, extensive, and reliable cloud platform, offering over 200 fully featured services from data centers globally. AWS provides multiple options to help you build, run, and scale every type of application, including:

Regions and Availability Zones ›

Edge Network Points of Presence ›
(for content delivery)

Local Zones ›

Wavelength Zones ›
(for 5G mobile edge computing)

AWS Outposts ›
(to run AWS on premises)

Building your applications on AWS

When considering your plan to leverage AWS in building applications, it's necessary to think about where your different workloads will live, which features to implement, and what changes to make to your application's architecture. AWS provides guidance and infrastructure that is an ideal environment for applications that need to scale, providing your organization with access to highly elastic compute, storage, and networking services.

Whether you are migrating an existing application, building a new application, or considering a hybrid on-premises/cloud solution, AWS Cloud has the solution that's right for you. Doing so can improve cost-efficiency of IT investments through pay-per-use models while helping you achieve high availability of your workloads, fast and efficient storage, robust security, and the confidence your application growth can't be stifled by lack of scale or innovative tools.

Many organizations approaching the cloud are challenged with existing processes and applications, distributed teams, legacy infrastructure, and technical debt. To help better prepare teams for success in the cloud, AWS has created tools and experiences to guide our customers with best practices and expert advice.

The **AWS Well-Architected Framework** helps cloud architects deploy resilient and efficient infrastructure for all kinds of applications. Built around our six pillars of cloud success—operational excellence, security, reliability, performance efficiency, cost optimization, and sustainability—the service provides a consistent approach for evaluating architectures and implementing scalable designs. With AWS Trusted Advisor, you can gain recommendations that help your teams follow AWS best practices by identifying practical ways to optimize infrastructure, improve security and performance, reduce costs, and monitor service quotas. Pair this with Amazon CloudWatch to monitor your applications, respond to system-wide performance changes, and optimize resource utilization.

What global-scale infrastructure means to these AWS customers:

Reuters: Distributes content to billions of global users every day

Discovery Channel: Delivers 8,000+ hours of original programming per year in nearly 50 languages

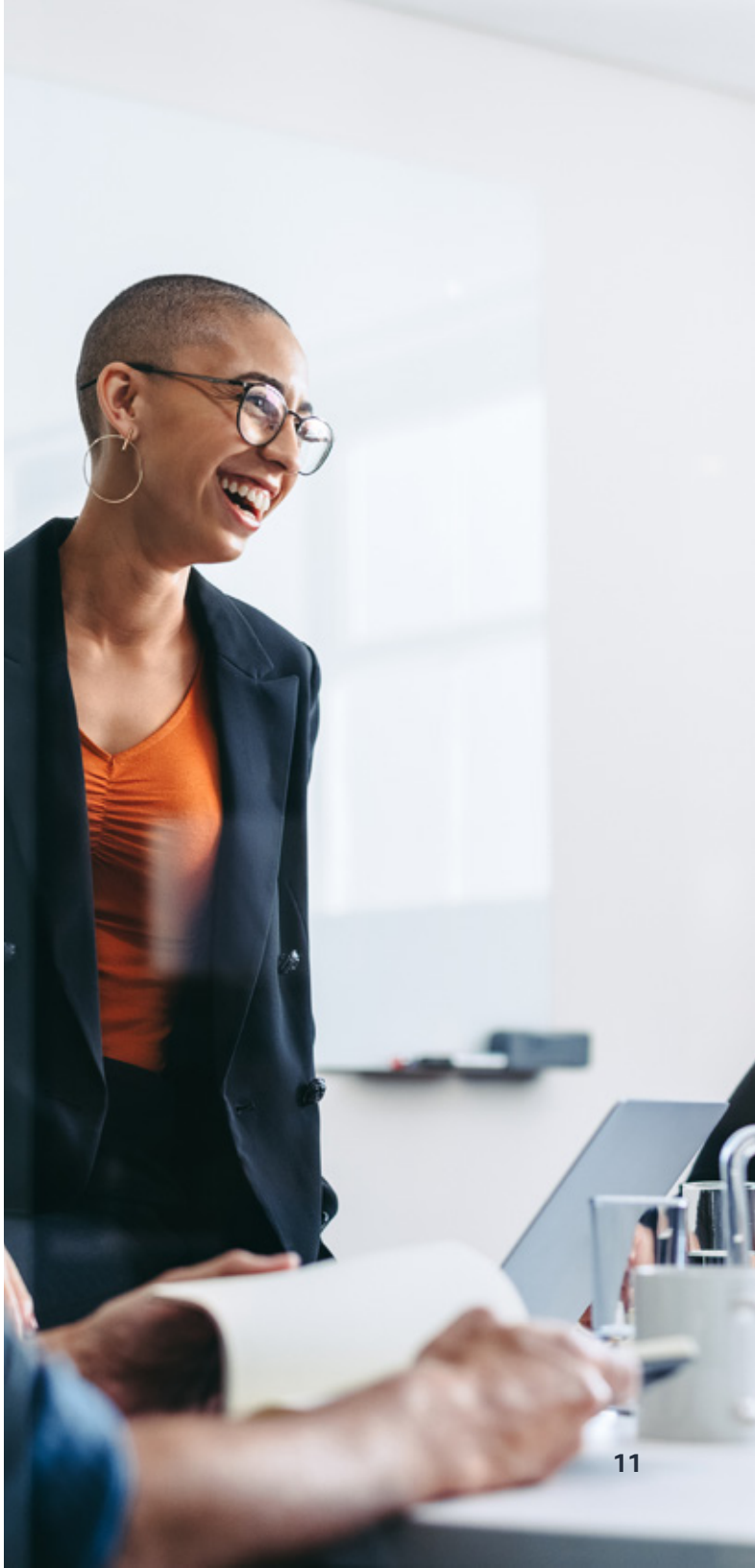
Epic Games: Supports peak usage more than 10 times that of non-peak and scales to host game events with all of its 200 million users invited

Scaling compute to run your applications

Scaling applications to serve an expanding customer base requires access to secure, reliable, and resizable compute capacity. Amazon Elastic Compute Cloud (Amazon EC2) enables you to scale compute capacity within minutes with an SLA commitment of 99.99 percent availability. Security is built in to the foundation of Amazon EC2 with the AWS Nitro System, which is the underlying platform for the latest generation of Amazon EC2 instances. Whether your applications live in the cloud, at the edge, or on premises, Amazon EC2 provides consistent compute infrastructure to build, run, and scale your applications.

The ability to rightsize your compute resources in the cloud can help you meet the needs of your applications while providing considerable flexibility to optimize your costs. Many applications, such as web servers, developer environments, and small databases, do not need consistently high levels of compute power, yet they may occasionally benefit from access to very fast CPUs, while other applications with use cases such as video encoding or HPC applications regularly require high-performance compute. AWS instances are engineered specifically for these types of use cases to eliminate the variable performance or other side effects you might typically see from oversubscription in other environments. With more than 500 Amazon EC2 instance types, AWS infrastructure allows you to scale compute resources optimally to meet precise requirements.

Once you have selected the right instances for your applications, having the tools and capabilities that enable you to scale seamlessly is critical. With AWS Auto Scaling, your teams can easily monitor applications and automatically adjust capacity to maintain steady, predictable performance at the lowest possible cost. AWS provisions compute capacity across your choice of Amazon EC2 instance types, Availability Zones, and purchase models, allowing you to optimize scale, performance, and costs. Furthermore, AWS services such as Amazon Elastic Kubernetes Service (Amazon EKS), Amazon Elastic Container Service (Amazon ECS), and AWS Batch enable you to orchestrate and batch your workloads to easily and efficiently scale your applications.



Scaling storage for virtually every application

Millions of customers use AWS storage services to run their applications to increase agility, reduce costs, and accelerate innovation. You can choose from a broad portfolio of solutions with deep functionality to store, access, protect, and analyze your data for virtually every application. Whether you want to build a data lake to run analytics and machine learning for insights, archive or back up data at the lowest cost, or develop powerful cloud-native apps that scale automatically, Amazon Simple Storage Service (Amazon S3) delivers industry-leading scalability, availability, security, and performance your applications demand. Customers of all sizes and industries use it to store and protect any amount of data from anywhere.

For your most mission-critical applications that require block storage, Amazon Elastic Block Store (Amazon EBS) delivers a wide variety of price and performance options for Oracle, SQL, and SAP HANA. It is designed to deliver millisecond latency to support your most demanding applications. It's never been a better time to migrate SAN-based applications to AWS for performance, cost savings, management simplicity, and virtually unlimited scale.

For your file-based applications on premises, AWS offers like-for-like capabilities with the Amazon FSx family, so you can migrate to the cloud seamlessly and benefit from the high-performance, low TCO, and enterprise features with fully managed services for your applications running in Windows, NetApp ONTAP, OpenZFS, and Lustre file systems. To build serverless applications using serverless file storage, you can use Amazon Elastic File System (Amazon EFS). With a "set-and-forget" elastic file system that scales automatically, you can simplify DevOps and big data analytics applications without managing storage.

How cloud infrastructure drives better outcomes:

The right compute: Leverage the industry's broadest and deepest portfolio of instances for compute infrastructure, allowing you to build with a consistent set of services and instance types globally

Built-in security: AWS is the most secure cloud provider, ensuring that your infrastructure satisfies compliance requirements for virtually every regulatory agency in the world

Optimize resource costs: Pay only for the resources you use, choose from a variety of purchase models that best suit your business needs while saving you the most, and further optimize costs through intelligent usage-based cost-optimization capabilities

Launch applications faster: Speed time to market, avoid complex capacity planning, and reduce over-provisioning with just a few clicks

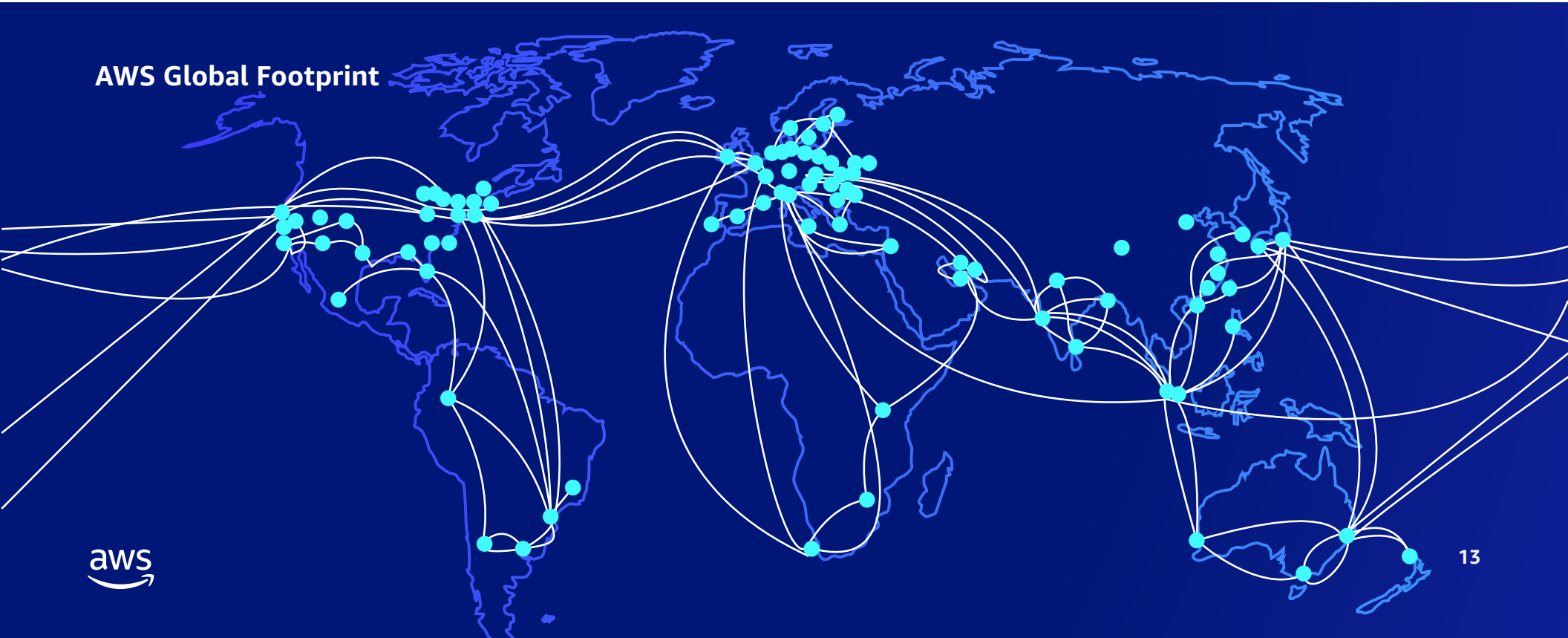
Application connectivity and networking considerations

Users expect low-latency, highly available connectivity to web applications all over the world, while your organization demands security, performance, and reliability at a reasonable cost. When building for these requirements, you shouldn't overlook the role networking plays in improving your applications and how you can take advantage of these benefits AWS provides globally through integrated networking and security solutions at the edge of the AWS Cloud.

The internet is unpredictable and non-deterministic, a labyrinth of interconnections constructed from stitching together thousands of individual network strands. For businesses seeking to make the connections work reliably at scale, this inherently non-deterministic environment—the fog between where applications are hosted and where they are consumed—creates unpredictability and a lack of visibility and control over application performance, security, cost, and availability.

In order to manage the huge volume of data transferred daily between AWS Regions, application users, and APIs, AWS has invested in building a dedicated global networking backbone. This is a critical part of what allows AWS customers to connect securely and reliably to end users at scale.

Every component of AWS cloud infrastructure is built for redundancy and reliability to maintain availability when running at a global scale—from AWS Regions and services to the physical network cables and connection points.



AWS edge networking services securely transmit your internet-facing data while delivering improved latency worldwide. By moving traffic off the internet and behind the defenses of the world's most secure cloud provider, you can limit your exposure to attack by encrypting data, removing network hops, and controlling application access. Networking services sit at AWS global edge locations, connected by dedicated 100 Gbps redundant fiber, to deliver data with ultra-low latency day in and day out. Using AWS edge networking services, you can improve performance, security, and availability, reduce cost, and better monitor network traffic.

How AWS helps you scale for always-on availability

- **Customizable and secure content delivery networking:** Customize web content delivery using Amazon CloudFront to improve latency and protect applications. Amazon CloudFront is a fast content delivery network (CDN) that accelerates the response times of Internet-facing applications and APIs to consumers all over the world. Its edge compute capabilities integrate with security services AWS Shield and AWS Web Application Firewall (AWS WAF), enabling you to build comprehensive perimeter protection to defend against DDoS and bad bot attacks.
- **Accelerate global network traffic:** Limited bandwidth on public networks and spikes in network traffic result in slowed or lost connections and dropped data. Solve this issue with AWS Global Accelerator, which moves network traffic onto congestion-free private network infrastructure provided by AWS and improves network traffic performance by up to 60 percent.
- **Optimize traffic management and routing:** End-user latency and failed connections often stem from inefficiencies in the routing of network traffic between devices and your applications. Amazon Route 53 solves the issue by translating website names into IP addresses to foster internet communication. AWS Global Accelerator further increases performance and simplifies traffic routing by providing two static customer-facing IPs. On the backend, AWS Global Accelerator allows traffic rerouting along the AWS backbone without customer-facing address changes or downtime.
- **Continually reduce costs:** Delivering data through edge locations reduces your application costs by limiting and consolidating requests. All data transferred to AWS edge locations from an AWS resource incurs no additional charge. All AWS edge networking services are "pay as you go," with no upfront costs and no minimum usage. Customers willing to make a long-term commitment can sign up for self-service discounts, such as the CloudFront Security Savings Bundle.

Bringing AWS closer to where you need it

Some applications cannot easily migrate to the cloud. Perhaps you need to work with large local datasets, share data with on-premises applications with single-digit millisecond latency, or meet data residency requirements. In these use cases, constantly transmitting data through the cloud is too slow, too cumbersome, or simply not permitted.

Traditionally, building your hybrid and edge infrastructure and solutions has been complex and challenging, with the focus on trying to integrate legacy systems and architectures with cloud services. All this to maintain different infrastructure platforms, different application programming models, and separate tools between your on-premises, edge, and cloud environments. To resolve this complexity, you need an extension of the cloud to where your on-premises and edge applications reside. An extension that provides you the same cloud infrastructure and services to be deployed wherever you need them, for a truly consistent AWS experience.

AWS has an extensive cloud service portfolio to address some of your most common use cases:

- **Low latency** – For use cases such as manufacturing automation, media and entertainment content creation, real-time gaming, financial trading platforms, electronic design automation, and machine learning inference at the edge.
- **Local data processing** – Some of your datasets may need to be processed locally because they can't be easily migrated to the cloud. Build on AWS for a consistent hybrid cloud architecture to process data on premises and easily move data to the cloud for long-term archiving.
- **Data residency** – Security/tax regulations, data sovereignty, and geopolitical dynamics may require you to store data in a country, state, or municipality. This is especially common in financial services, healthcare, oil and gas, and other highly regulated industries.

What bringing the cloud closer to where it's needed looks like in action:

- DISH automated its network to bring 5G applications to market faster
  Watch the video
- Riot Games is providing players with the best possible gaming experience
  Watch the video
- Aurrigo is achieving the bandwidth, latency, and security necessary to connect its autonomous shuttles in Cambridge
  Watch the video

Where to learn more

Customers choose Amazon Web Services because they recognize AWS can help them innovate faster and scale their business better than any other provider. As modern compute, storage, hybrid, and edge applications evolve, businesses take advantage of AWS infrastructure and services to tackle larger and more complex challenges at the best price performance for elastic usage and global scale.

Businesses adopting AWS infrastructure have access to the most secure, extensive, and reliable cloud platform available in the market, with services across 245 countries and territories. AWS provides customers of all sizes the ability to scale application resources seamlessly and optimally to meet customer demand with a wide selection of available solutions across compute, storage, networking, hybrid, and edge infrastructure.

For businesses that need to deploy their applications across the globe—or are looking to build and deploy specific applications closer to their end users with ultra-low latency—AWS provides enterprise-grade cloud infrastructure where and when they need it.

Learn about AWS Free Tier ›

Learn more about how AWS makes it simpler and more secure than ever to scale your applications for the needs of today and tomorrow.