DigitalOcean

Cloud Computing for Media Streaming Businesses

Five ways to optimize performance and price

While streaming is often associated with video services, when discussing cloud applications the streaming industry also incorporates any businesses that send and receive continuous data.

This can include both on-demand and live video streaming, and other technologies such as Internet of Things (IoT) applications which are constantly sharing data between multiple sources in real-time.

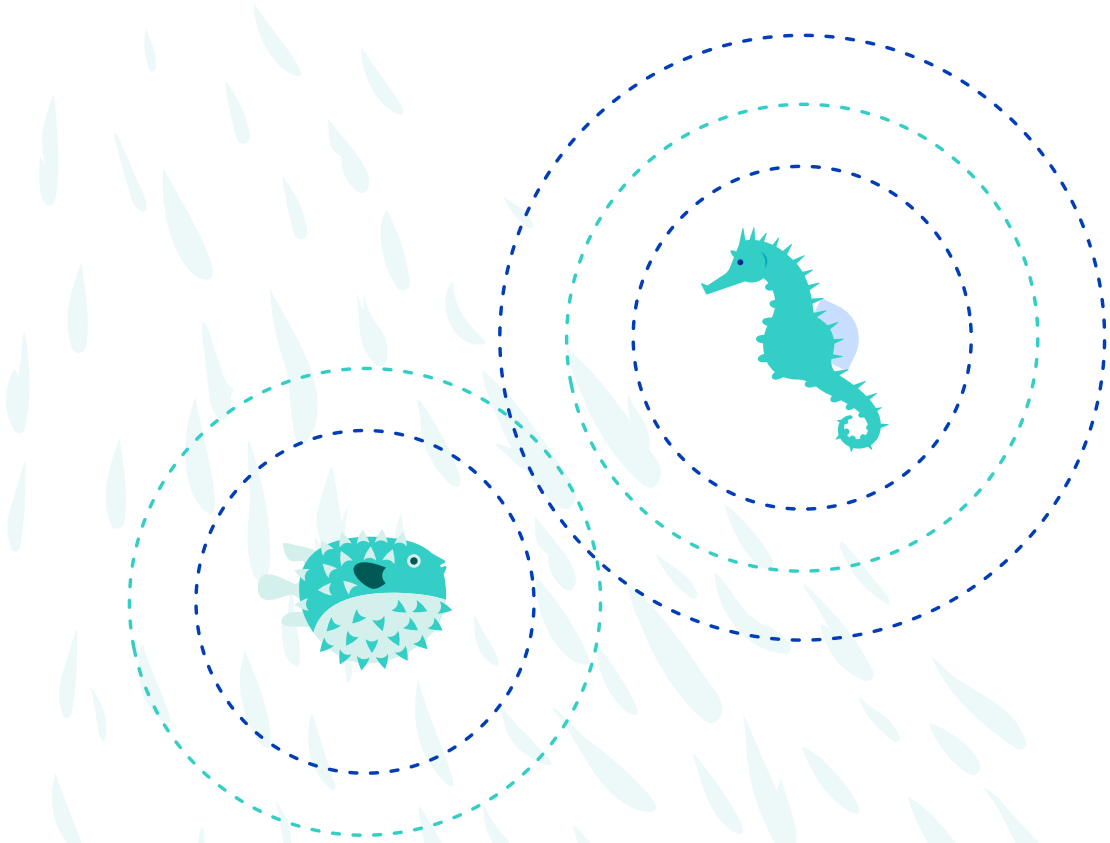
Due to the growing popularity of video and audio streaming, podcasts, online gaming, and more, the streaming industry is growing rapidly. A study by Grand View Research found that the global video streaming market alone was valued at $59.14 billion in 2021, and the market grows even more when we consider data streaming technologies. As more data streaming applications are introduced, businesses that serve both live and on-demand video and audio content need robust computing and networking capabilities to deliver an excellent customer experience. Startups entering the space and small- and medium-businesses (SMBs) working toward rapid growth need access to quality infrastructure at a reasonable price.

Streaming services need to be fast, consistent, and scalable. Streaming is unique from downloading files because data is being transferred in real-time, playing through a browser rather than through a downloaded file on a computer. Streaming also requires processing a large amount of data with minimal delay to keep applications running smoothly and customers happy.

Because of the unique needs that streaming businesses have, building a streaming business that's cost-effective, performant, and scalable enough to meet user demand is largely enabled by choosing appropriate cloud computing solutions. Maintaining low latency is essential, and storage is also an important component of application architecture. Bandwidth is another large consideration for those building streaming applications, as bandwidth costs can be substantial for network-intensive applications.

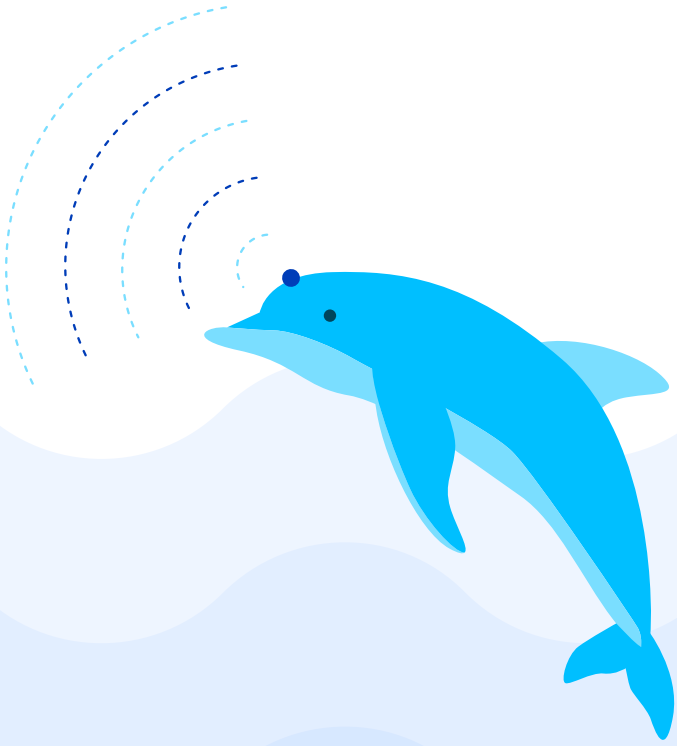
By optimizing architecture for high performance and planning for scale, builders can future-proof growing businesses and increase their chance of success. DigitalOcean is here to help, and in this guide we've outlined five ways you can optimize your performance and minimize costs when building and scaling your streaming applications.

Table of contents



Streaming terms to know

Streaming business	Any business that uses continuous data, whether it's sending or producing continuous data or receiving continuous data.
Live streaming	Live streaming refers to content that is received in real-time as it's being made. It's one connection to many users.
Video on demand (VOD)	Video on demand is a pre-recorded file delivered from a storage location a little bit at a time.
Audio streaming	Receiving audio over the internet.
Data streaming	Utilized by businesses and devices to transfer large amounts of data. Examples include IoT devices such as Amazon Alexa, Google Home, or sensor data in vehicles.



1 Choose the right virtual machine

When it comes to infrastructure management, the virtual machine (VM) is the foundation for success. Selecting the wrong virtual machine, or cutting costs in the wrong areas, can significantly impact your ability to reliably and appropriately serve customers.

Streaming businesses building their architecture through a cloud provider often have a variety of configuration options to choose from. While it may seem basic, ensuring that you have the appropriate amount of RAM, vCPUs (virtual Central Processing Units), storage, and outbound transfer can mean the difference between failure and success. As you consider what type of virtual machine is best for your workload, start with choosing whether you need shared vCPUs or dedicated vCPUs.

Dedicated vCPUs Provide resources that are fully dedicated to one business. They provide faster, more consistent performance. Dedicated vCPUs are essential for applications that are latency-sensitive and have high processing requirements.

Shared vCPUs Provide resources shared between multiple businesses. While a hypervisor can ensure that everyone can always utilize a healthy portion of its underlying hyper-threads, companies sharing vCPUs can be affected by "noisy neighbors" or another VM running a particularly CPU-intensive load.

Streaming businesses will almost always want dedicated vCPUs. Attempting to cut costs by choosing shared vCPUs can lead to increased latency, more risks of downtime, and a poor customer experience.

Many cloud providers will also allow customers to choose machines optimized for different workloads. A virtual machine explicitly optimized for your needs can help save costs on extra resources you may not utilize.

For example, DigitalOcean has both CPU-optimized Droplets (VMs) and memory-optimized Droplets. CPU-optimized Droplets provide dedicated vCPUs, but less RAM, saving costs for businesses who need compute power without a lot of RAM. Memory-optimized Droplets still provided dedicated vCPUs, but with twice the RAM, enabling them to accommodate more memory-intensive business applications. Streaming businesses on DigitalOcean can often save costs by choosing CPU-optimized Droplets, as their needs are typically less memory-intensive.

Dive deeper:

An overview of cloud computing
How to choose the right virtual Machine

2 Use a CDN and caching to deliver files more quickly

Streaming businesses serve large files or file segments to users across the globe, and one of the most effective ways to minimize stress on the origin server and deliver content faster is by leveraging a Content Delivery Network (CDN), which can cache and serve content from geographically distributed Points of Presence (PoPs). When set up correctly, a CDN can deliver on-demand video content, live-streamed video content, and streams of data more quickly and efficiently than if users were requesting content directly from the origin server. This is achieved in two primary ways, both of which should be considered when setting up your streaming architecture:

Strategically located PoPs

PoPs are distributed servers that act as the connection between an origin server and the end-user requesting content. CDNs have PoPs located around the globe and serve content to the requestor from the PoP closest to them. For example, a device located in France requesting content from an origin server in the United States may be routed to a PoP located nearby in Germany, which reduces the request time. By reducing the distance between users and data assets, businesses can deliver content more quickly and improve user experience with faster speeds. CDNs each have their own network of PoPs, and streaming businesses should consider where their users are located when creating their architecture and setting up a CDN, as some networks will have a stronger PoP network in certain parts of the world than others.

Caching

Although a user experiences a continuous stream of video content or other streamed data, streaming businesses deliver this data in multiple pieces which are then assembled and displayed in a way that creates a seamless experience. When videos pause playing or start buffering, that is because the next segment of the streamed content has not yet been fully delivered to the local device. Caching refers to creating copies of data that can then be stored and delivered to users from a cache server (which may also be a distributed PoP). By creating multiple copies of a particular asset, businesses can reduce the load on their origin server, enabling them to save on bandwidth and reduce latency so that content gets to users more quickly. Live-streamed content can also be cached in a similar way—instead of a stored video file being broken into parts and delivered, content is cached in real-time, which can still result in a faster stream delivery than if the request went to the origin server.

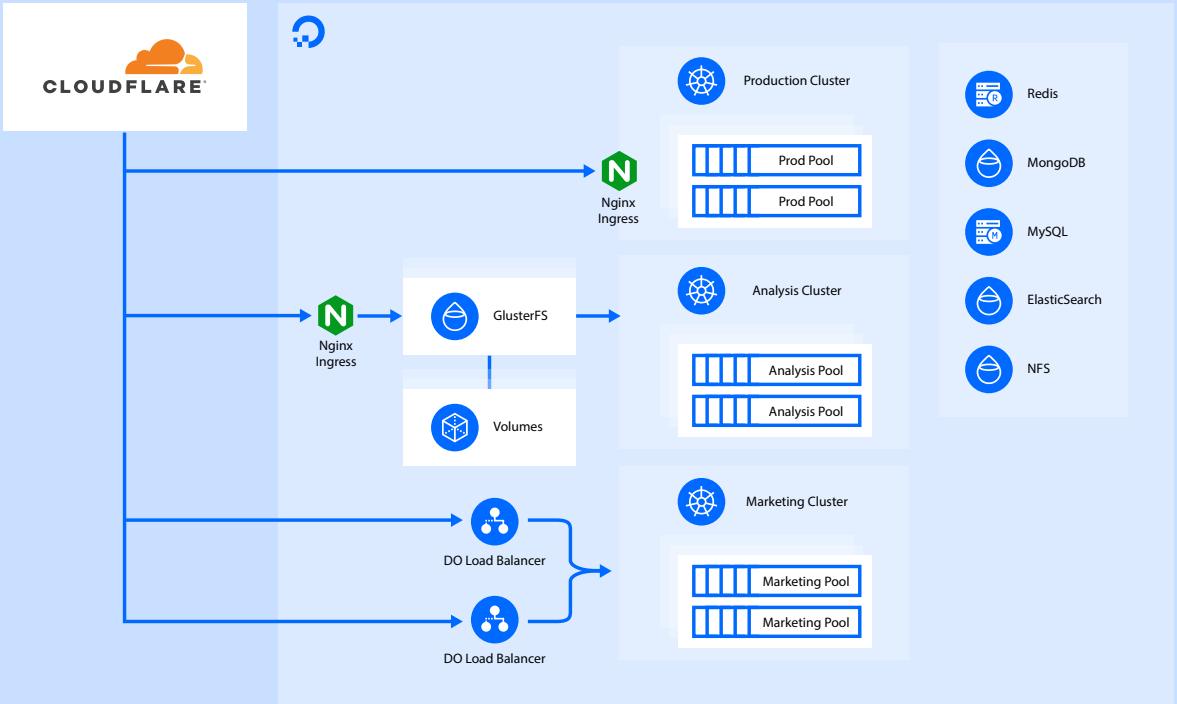
When creating the architecture for your streaming business, you will want to consider utilizing a CDN with a caching solution so that requests can be served as quickly as possible from a distributed server. DigitalOcean Spaces offers a built-in CDN with more than 25 global PoPs, and also integrates with other CDNs. Atom Learning is one example of a customer using DigtialOcean Kubernetes and Droplets, plus Cloudflare's CDN on top of their infrastructure.

Dive deeper:

Everything you need to know about a CDN
Using a CDN to speed up static content delivery
Three ways object storage with a CDN can maximize performance

How others have done it: Atom Learning

Atom Learning provides high-quality education for primary school students through a fully adaptive online learning platform. Atom Learning runs three Kubernetes clusters with multiple node pools, all of which run on DigitalOcean Kubernetes. Atom uses a microservices architecture on the backend that uses REST to communicate with each other and the frontend React app.
Atom Learning has scaled to over 350,000 accounts and continues to grow, experiencing 50% growth month to month through 2019. The team now has over 95 employees. Even with that tremendous growth, the team hasn't had to make any significant architecture changes since launching with DigitalOcean.



3 Enhance performance with adaptive bitrate streaming

Another way for streaming businesses to reduce stress on servers and save bandwidth costs is by streaming files more efficiently through adaptive bitrate streaming (ABR). Rather than sending video through a predetermined bitrate regardless of the end user's capabilities, implementing ABR allows for the adaptation of the bitrate based on the end user's device and bandwidth capabilities. This allows the server to limit its output based on the connection and send only the data that the user can receive and process.

ABR works by encoding content, a video, for example, with multiple bitrates to match various devices and download speeds. After the video is encoded, it's broken into segments. As the video plays on the end user's device, an algorithm determines the device's needs. An adjustment process is triggered at the end of each segment, evaluating if the content is playing at the appropriate bitrate. If the end user's device is struggling to download the content fast enough, it will switch over to a smaller file to finish viewing. If it's determined that the quality is poor or the device can handle a higher bitrate, that adjustment will be made as well. Typically, devices begin with the lowest bitrate option and move up as it's determined the device can handle more.

ABR allows for better quality and performance of content, providing a better customer experience while saving compute power if you're encoding content as it's sent and reducing overall bandwidth costs.

Dive deeper:

Setting up ABR
Enabling New Possibilities for Real-Time, Scalable Video Streaming

4 Maximize efficiencies to reduce bandwidth costs

Reducing bandwidth costs while maintaining performance is crucial for a growing media streaming business. Video streaming applications utilize a tremendous amount of bandwidth, and incremental efficiencies can add up to massive savings as you grow.

Foundationally, choosing an appropriate codec—or the way that you encode a video—is essential. Inefficient codecs can waste bandwidth and money. While there are several options for codec, H.264 and AV1 are typically best suited for online video streaming.

H.264

H.264 is the standard video encoding format. It's often used to record and distribute HD video and audio and is excellent for distributing video to multiple sources. Almost any device can play H.264, and it's one of the most widely used formats available.

AV1

AV1 is a more modern video encoding format. It's open source and royalty-free and allows providers to encode higher quality streaming content while delivering faster frame rates and higher resolution.

Choosing a supportive cloud provider is another way to optimize bandwidth costs. For example, data transfer into DigitalOcean and within your private networks is free. DigitalOcean offers bandwidth pooling that gives customers an outbound data transfer quota depending on the number of DigitalOcean Droplets they have. All the Droplets together form a bandwidth pool, and there are no additional charges for outbound transfer up to the bandwidth pool quota. For example, if you have two Droplets with 2TB outbound transfer available between them (1TB per Droplet), and Droplet one uses 1.5TB and Droplet two uses zero, there are no additional charges for the bandwidth. This saves significant costs as businesses grow and scale.

Dive deeper:

DigitalOcean Bandwidth Calculator

5 Automate as much as possible

A streaming service that goes down if someone isn't actively manning the keyboard to bring things back up isn't set up for success. As you set up your streaming architecture, consider how you can automate infrastructure management tasks like deploys, scaling, healing, and more. Automating infrastructure management is typically done in two ways:

Container orchestration with Kubernetes

Kubernetes improves resource utilization and shortens software development cycles. Because its self-healing and auto-scaling capabilities can ensure high reliability and great uptimes and response times, it often enhances the user experience by improving product quality and stability. Its ability to scale up and down to meet performance demands is another way to optimize bandwidth costs, ensuring you're only paying for what you're using. Media streaming workloads are great candidates for Kubernetes.

Infrastructure as Code

Streaming businesses should also consider using infrastructure as code (IaC) to help with infrastructure automation, deployment, and changes. With IaC, you can quickly create as many instances of your entire infrastructure as you need, in multiple provider regions, from your declarative code. Using a provisioning tool like Terraform, you can write code that defines your infrastructure. After Terraform builds out what the infrastructure looks like, a tool like Ansible, which is popular with media streaming services, manages the ongoing configurations such as patching, pushing software, and configuring system services.

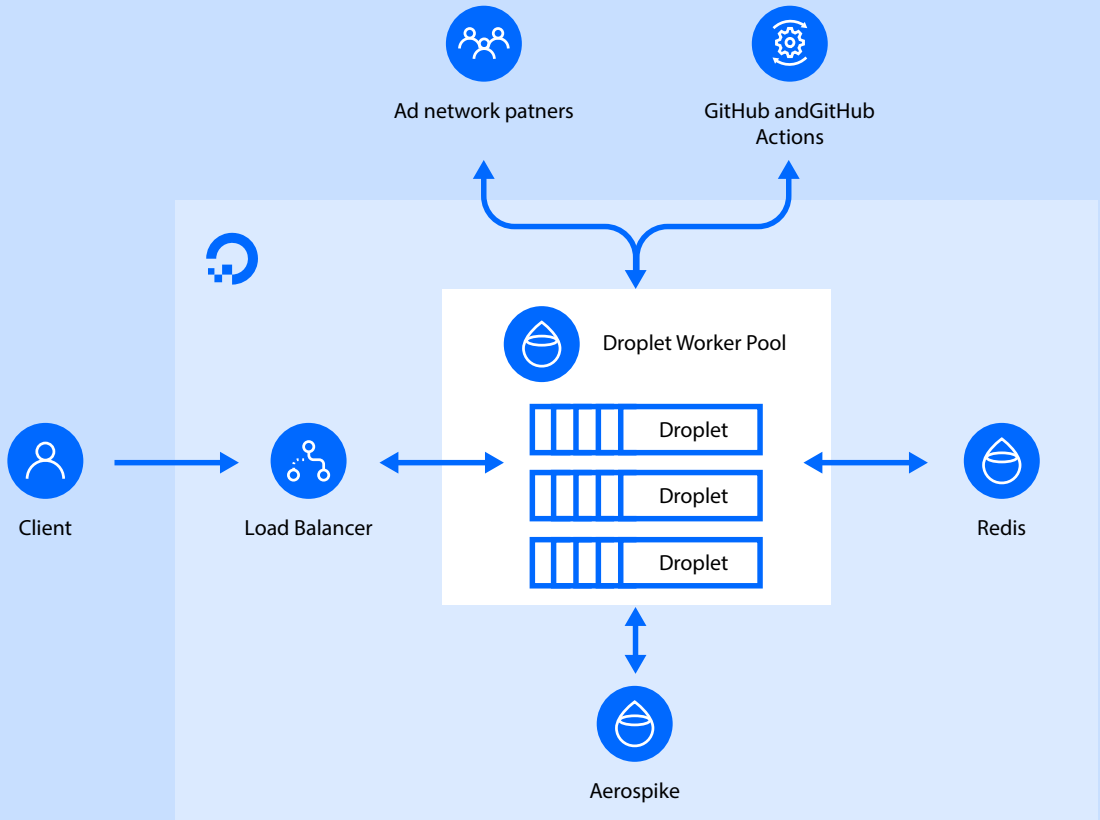
Dive deeper:

Infrastructure as code explained
How to use Ansible with Terraform
Monolithic vs microservice architecture: which is best?

How others have done it: Origin

Origin is a creative and technology company that delivers attention-grabbing, dynamic advertisements to consumers watching connected television. Origin uses 25 DigitalOcean Droplets operating behind a DigitalOcean Load Balancer. The Droplets function as stateless, algorithmic engines that exist to receive instructions and data about existing campaigns and receive bid requests from partners. Origin can receive upwards of 40,000 queries per second (QPS), and for every request that comes in, the Droplet opens up seven outbound connections to the ad partners to receive bid requests.

As Origin expands to different locations across the globe, they can easily replicate this architecture in other territories. Since it's primarily stateless, they can replicate the configuration in various data centers by simply adding a geo-load balancer. When they have to add Droplets, it can be done with the click of a button. Origin was able to create Droplets, name them, add them to the Ansible host list, and run the Ansible playbook. The DigitalOcean Load Balancer immediately added the new Droplets to the cluster. Origin houses everything in GitHub and uses GitHub Actions. Every commit creates a different version of the server, meaning they can deploy a specific version or roll back if needed.



Choosing a cloud provider

Running a streaming application is a lot of work. There's no reason to add even more complexity with your cloud provider. When researching a cloud provider for your business, consider the following:

CPU considerations

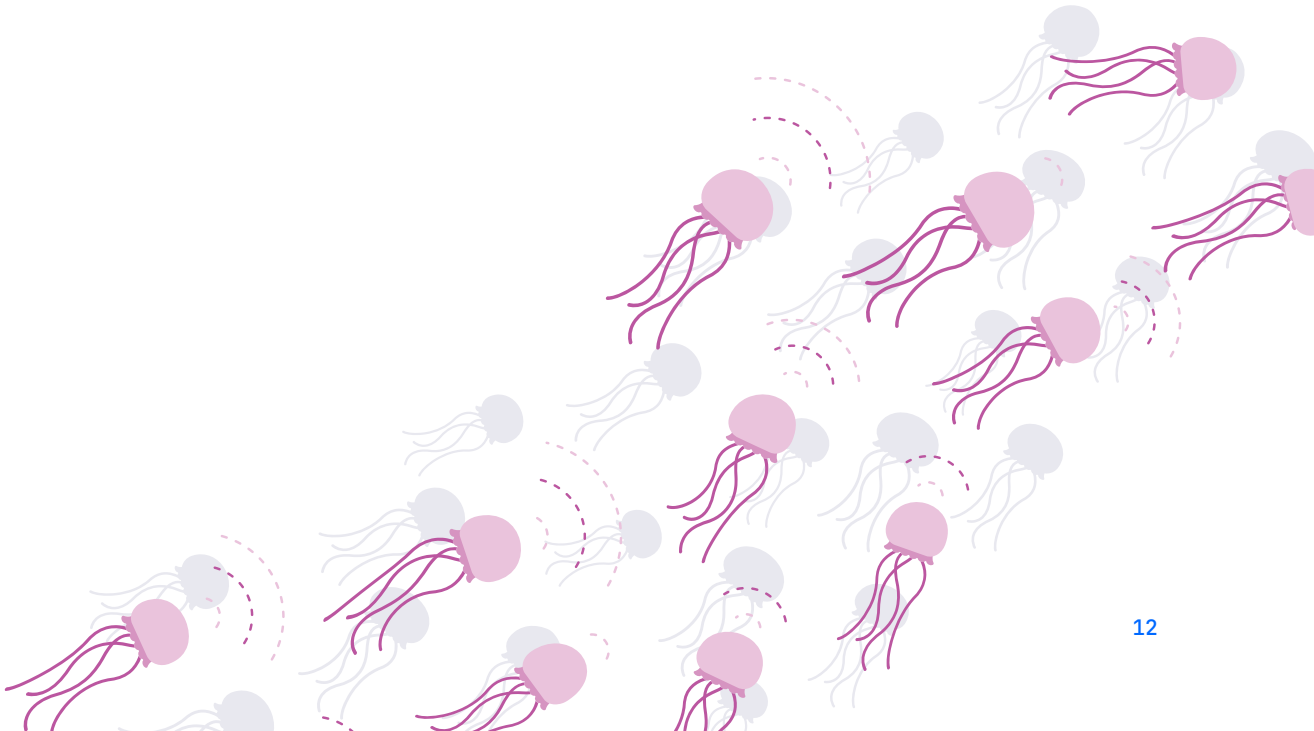
Streaming requires a significant amount of egress and ingress. Take a close look at bandwidth models, benefits, and pricing of potential providers. For example, DigitalOcean provides free ingress and bandwidth pooling. You can estimate potential bandwidth costs on DigitalOcean with our bandwidth pricing calculator.

> **"Digital Ocean's pricing model literally enables our business to exist. Specifically, bandwidth pricing."** – Gather

Uptime considerations

Streaming needs extremely reliable services, so find out about the reliability of any provider you're considering. Ask if there's lock-in with long contracts or how well they support cloud native computing, and research uptime SLAs for computing needs.

> **"The prerequisite for any cloud provider is availability. If your platform isn't reliable and functioning at peak performance, it doesn't matter if your support is great or your prices are low. I'm confident in DigitalOcean's infrastructure, service level, and scale, and I'm willing to vouch for that."** – DevOps Manager, Origin



Storage options and pricing

Object storage is vital for streaming businesses. Large cloud providers can have a complicated and layered approach to pricing their object storage solutions. The pricing is usually based on multiple factors, including the location where the object storage is created, the type of object storage, the number of requests to the object storage, and the data transfer/bandwidth costs. Due to the variable nature of the above factors, it's hard to estimate the final monthly bill. Plus, adding a CDN is an additional cost with some cloud providers. Assessing the actual cost of using the CDN service is challenging because multiple factors are involved, such as the geographic location where the data is delivered, the length of the service contract, and more.

For example, DigitalOcean Spaces is an S3 compatible storage option that provides object storage option that features a built-in CDN. Spaces is simple to use and provides predictable pricing, starting at $5 per month, including 250GiB of data storage and a built-in CDN for no extra cost. Additional storage costs only 2 cents per GiB. In addition, many libraries are built specifically for interacting with S3, like Python's Boto3, so the Spaces S3 compatibility helps from both a tooling and automation perspective as well as a cloud agnosticity perspective.

> **"One of the reasons we've been able to grow as quickly as we have is because of how scalable everything is and how easy it is to add new features on our DigitalOcean platform. We use every single one of your products now. It's been great."**
> – Tim Osborne, CTO, Atom Learning

DigitalOcean

About DigitalOcean

DigitalOcean simplifies cloud computing so developers and businesses can spend more time building software that changes the world. With its mission-critical infrastructure and fully managed offerings, DigitalOcean helps developers, startups, and small- and medium-sized businesses (SMBs) rapidly build, deploy, and scale applications to accelerate innovation and increase productivity and agility. DigitalOcean combines the power of simplicity, community, open source, and customer support so customers can spend less time managing their infrastructure and more time building innovative applications that drive business growth.

> To get started, **sign up for an account at DigitalOcean.com.** For more information or help migrating your infrastructure to DigitalOcean, **speak to a sales representative.**