WHITE PAPER

Roku Slashes Compute and Egress Costs

with Akamai Cloud Wrapper

Table of contents

Content consumption is evolving

Roku is a leading streaming platform that transformed how people watch television. Founded in 2002, Roku continues to be at the forefront of innovation in the streaming industry, creating powerful and easy-to-use devices that allow users to access a vast range of streaming services, including Netflix, Hulu, and Amazon Prime Video. Roku's devices come in various shapes and sizes, from popular streaming sticks to the high-end Roku Ultra, which offers advanced features such as 4K and HDR streaming. In addition to Roku's devices, Roku's robust operating system, the Roku OS, is integrated into many smart TVs and provides an intuitive user interface that makes searching and streaming content a breeze. With over 70 million active accounts and a commitment to constantly improving its products, Roku remains a significant player in the streaming space.

In 2017, Roku launched The Roku Channel, which rapidly rose to become a top free ad-supported streaming service. Roku has built its own video platform for streaming video content and was focused on maintaining an excellent user experience while delivering media at scale.

Approaches to optimize users' streaming experience

One approach is to use dynamic packaging for media preparation. Dynamic packaging is a widely used technique many video streaming platforms employ to optimize online content delivery. It allows for more efficient video delivery by breaking the video into smaller, more manageable chunks. These chunks enable faster and smoother playback, regardless of the user's internet speed or device type. Dynamic packaging also enables adaptive bitrate streaming. Adaptive bitrates allow user devices to automatically select the best video quality possible without inducing buffering or playback interruptions.

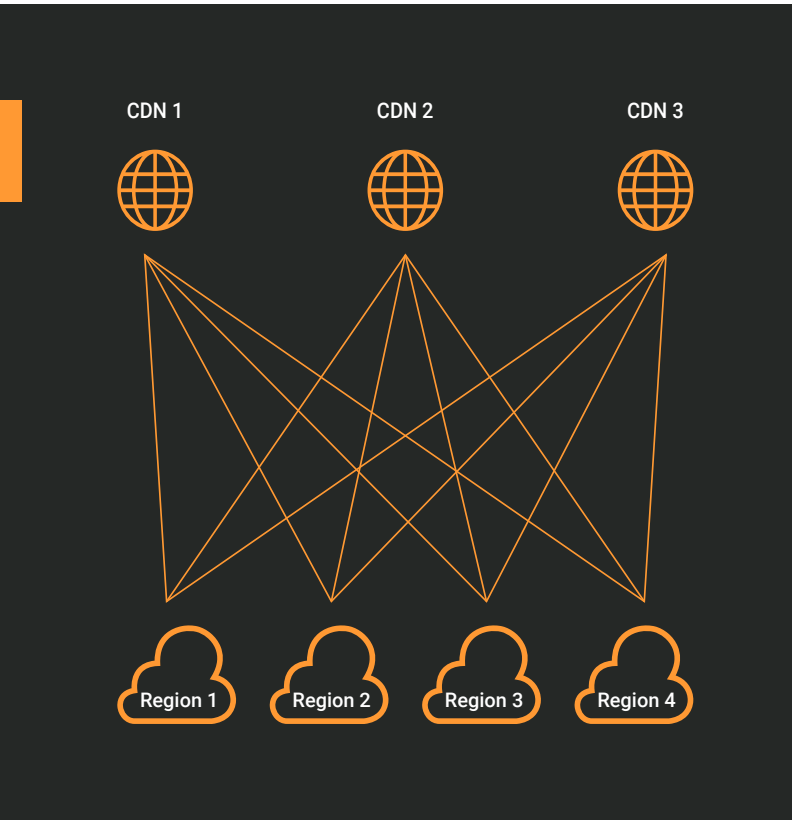
However, these benefits come at a cost. Segmenting the video into smaller chunks requires additional computing resources, increasing cloud computing costs. The amount of resources depends on how many bitrates are needed. Fewer bitrates save on computing and storage, but may create trade-offs between quality and experience for some users. More bitrates improve the overall experience, but each new version requires more CPU to generate and more object storage to maintain.

All things being equal, more bitrates also reduce the caching effectiveness of content distribution networks. CDNs operate on the principle that a single object can be retrieved from the cloud once and then delivered to millions of users. That reuse reduces the number of requests to cloud endpoints, reducing egress bytes transferred and cost. It also allows infrastructure to scale more efficiently by allowing the CDN to absorb most traffic without the need to scale computing or storage. Increasing the number of objects, like adding more bitrates, results in more file versions, each needing to be cached. Unless addressed, this could decrease the cache hit ratio, increase egress bytes, and demand more computing resources.

Challenges in production environments

Multicloud and multi-CDN architectures exacerbate these challenges. Using multiple CDNs to deliver multiple bitrates decreases the cache-hit ratio further, because each must now individually retrieve content before it can be delivered. Egress costs are multiplied by the number of CDNs in use.

Integration is also a challenge. Clouds and CDNs vary in interfaces and capabilities — one reason to use multiple providers. The differences in features provide flexibility, but create challenges with integration and operations.



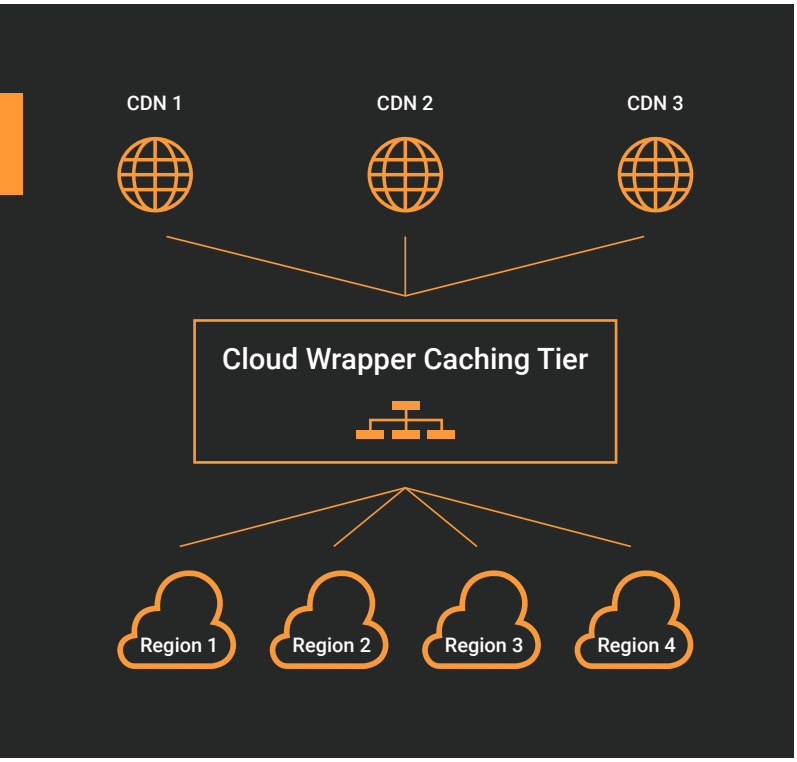
With millions of subscribers streaming exabytes of content daily, these challenges are significant for a company like Roku. Controlling costs, while ensuring the optimal user experience, is critical to continue innovating. To mitigate cost and complexity, Roku turned to Akamai to enhance its caching infrastructure.

Roku uses various cloud regions and CDNs to deliver service to a global audience. Global load balancing directs end-user traffic to the most appropriate CDN based on multiple conditions. The CDN then serves the requested video without consuming cloud resources or egress. But what happens if that's the first user that requested that video? In that case, it won't be in the cache. As a result, the CDN will have to retrieve the video from the appropriate region.

First, however, compute resources are needed to prepare — or package — the video segments. This process is dynamic because the packaging is only performed when a user requests a stream. As a result, the video not being in the CDN's cache results in 1) additional CPU to prepare the stream and 2) additional cloud egress charges. These cache misses not only drive up costs, they also impact user experience due to the latency associated with packaging execution and transfer. Since caches aren't shared across providers, adding CDNs only exacerbates the situation by increasing the number of potential cache misses. The challenge is ensuring CDN caches are populated while minimizing cloud computing and the required egress.

Add simplicity and redundancy

The solution is an additional multicloud and multi-CDN caching tier — Akamai Cloud Wrapper. Cloud Wrapper provides a dedicated caching footprint for each customer to maintain cache even in the face of noisy neighbors. It efficiently caches and replicates objects across distinct, logically grouped regions to maintain high availability. Cloud Wrapper adapts to traffic characteristics and replicates objects more as popularity increases.



Cloud Wrapper is deployed inline between CDNs and cloud regions. It provides a single, resilient aggregation point for cacheable content, regardless of the source cloud or destination CDN. That additional tier automatically caches objects any CDN delivers, effectively enabling cache sharing across CDNs without requiring expensive cloud computing or networking resources. The extra caching layer also reduces requests to cloud infrastructure from individual CDN regions, decreasing cloud costs in single-and multi-CDN architectures.

In addition to a centralized caching layer, Cloud Wrapper optimizes regions used for caching objects. Consistent hashing ensures Cloud Wrapper always fetches a particular video file from the same regions. It's also intelligent enough to consider headers and query strings in caching decisions — both critical to maximizing offload and reducing the consumption of cloud resources.

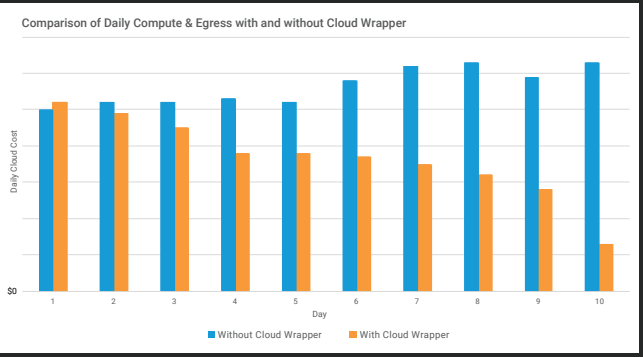
Of course, resilience can't be sacrificed in the name of resource savings. Architectures must maintain user experience during the inevitable component failures. That requires redundancy at multiple levels. Cloud Wrapper's additional caching tier provides that redundancy. It continues to serve content during origin failures to maintain service. The service itself is also highly redundant. It replicates content across multiple independent regions. This replicated content is accessible to any CDN, eliminating single failure points.

Reduce compute and egress costs

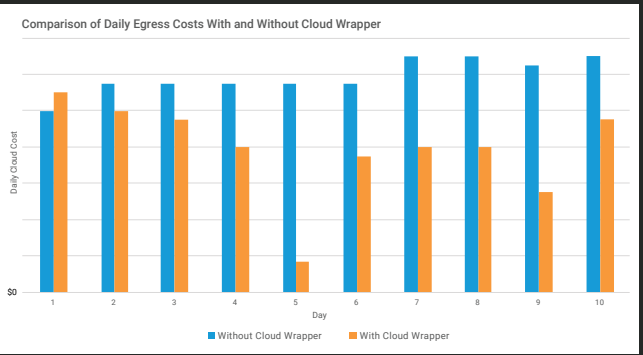
To realize a solution's value, you must implement it first. Cloud Wrapper's native support for multiple clouds and CDNs ensures it isn't a barrier. Once in place, Roku immediately saw results in the amount of computing and egress consumed by their dynamic packaging services. The trend in the usage of both is dramatic. The charts below compare 10-day periods before and after Cloud Wrapper implementation, and highlight the reduction in both.

Content is still king. A negative user experience when interacting with that content erodes the goodwill of even the most dedicated fan. Technologies like dynamic packaging exist to provide consumers and content with the experience they deserve. These technologies require additional computing and networking resources. That drives up costs and complicates effective, low-latency content distribution. Roku addressed these challenges by rethinking its approach to content caching — by addressing the complexity but not sacrificing flexibility. A small change in architecture — adding a cloud-and CDN-agnostic caching tier — allowed them to reduce costs without impacting their users, and set them up for their next 50 million viewers.

A 35% reduction in dynamic packaging compute costs



A 34% reduction in cloud egress costs



To learn why the world's top brands trust Akamai, visit akamai.com