Why your APIs need an edge cloud

Caching APIs to optimize performance and save costs
Overview

Users have come to expect more from both consumer and business applications, so slow and unpersonalized experiences just can’t compete. To engage users and deliver value, you need fast response times, personalized interactions, and device-specific user experiences.

Application Program Interfaces (APIs) are a critical component to any successful platform designed to achieve these goals. APIs are the building blocks used to send, receive, and modify business-critical data. But while APIs enable a more dynamic experience, performance is often a challenge.

Content delivery networks (CDNs) are typically used to enhance web and mobile performance. But legacy CDNs cannot cache frequently and unpredictably changing content like that delivered via APIs.

This is where Fastly comes in. As an edge cloud platform, Fastly allows you to process, serve, and secure data at the edge of your network. You can even cache content that changes frequently and unpredictably, which was previously considered “uncacheable.” In this brief, you will learn how Fastly can cache API responses to optimize performance, enhance personalized interactions, and dramatically reduce infrastructure costs without replacing legacy solutions.

“Being able to cache APIs with a CDN is a big benefit. If a user is logged out, they send a specific header; if a user is logged in, they send another header. With Fastly, we’re able to detect the difference in the header and cache it depending on whether or not a user is logged in, so we don’t have to go to the backend to serve the same content over and over again.”

– Alan Schaaf
FOUNDER AND CEO
IMGUR
The heat is on

To survive and thrive, you need to implement technical elements and best practices that enable delivery of optimized digital experiences — all the while controlling both capital expenditure (CapEx) and operating expenditure (OpEx).

Speed matters

Slow sites and apps can be a deal-breaker for users. One study finds that 53% of mobile site visits are abandoned if a page takes longer than three seconds to load.1 Given that speed is such an important factor, the tradeoff of fast delivery for the ability to serve constantly changing content is becoming increasingly challenging.

Consistent experience across channels

Users today expect unified and consistent experiences across web, mobile, and platform integrations. Data consistency is critical because users often consume content in multiple channels. In fact, research from PwC concludes that, “customers expect the same brand experience, quality, price and service regardless of channel, and it is that connectivity and seamless experience across channels that is the real differentiator.”2 But processes and technologies need to change in order to provide that consistent experience across all platforms.3 By delivering data via API, you can stop serving stale data and offer a truly consistent user experience.

Make it authentic

To stay competitive and relevant, your website and applications need to be able to target content to specific users based on location, language, and browsing preferences. You must also tailor content delivery to the user’s device, and personalize experiences based on actions they’ve taken within a site or app.

2 “Customers Are Calling the Shots,” January 2017.
The hits add up

Delivering personalized applications can be costly. Since legacy CDNs cannot cache API responses, every page view or app interaction results in a hit to the origin infrastructure and generates both infrastructure CapEx and OpEx costs. Furthermore, you’re double charged if you use data hosting providers — legacy CDN vendors and API gateway vendors charge to facilitate the API call and hosting providers charge to serve it.

The mighty API

In today’s distributed application platforms, many key services are provided by APIs. APIs are at the core of both web and mobile applications. They enable faster development, real-time data exchange, and more secure transactions. APIs have become so common that a recent survey of 10,000 developers, testers and executives showed that 63% of respondents work with APIs more than 10 hours per week.4

Example of an ecommerce API stack

If only you could cache your APIs, you would improve response times, ensure data consistency across channels, and future-proof your applications by making them secure and scalable. But there’s a problem... and fortunately, a solution from Fastly.

When The New York Times sends news alerts to its mobile users, 20 to 30 million push alerts are sent out in a single minute. The Times uses Fastly to mirror those alerts and take load off their central systems. The company saves around $25,000 per month by putting Fastly in front of its alerts API.

The New York Times

“2019 Postman State of the API Report,”
December 2019.
The challenge of caching APIs

Static content like images and text can be reliably cached at the edge, speeding delivery and reducing infrastructure costs. Dynamic content, such as product inventory and usage reports, change frequently and unpredictably, and traditionally have been considered too dynamic to cache by legacy CDNs since they are not designed to instantly invalidate and remove cached content.

This dynamic content typically remains at the origin server — slowing delivery and racking up infrastructure costs. Many companies employ a workaround strategy by caching content using a low Time-To-Live (TTL) setting in the CDN. TTL determines how long an object should remain in the CDN cache before it expires and therefore has to be refreshed from origin.

Unfortunately, this workaround strategy has a negative side-effect: stale, outdated content may be served if it changes at the origin before the TTL expires. A 60-second TTL is a large enough window to cause an inconsistent or confusing user experience. On the flip side, a very low TTL leads to premature expiration of content, which results in the origin infrastructure bearing more load than is necessary.

“Fastly serves our applications out faster. Because Fastly’s CDN enables us to cache APIs for the first time, our viewers get a much faster response than if requests went to our origin.”

– Sung Ho Choi
CO-FOUNDER
FUBOTV

Introducing Fastly

Fastly is an edge cloud platform that provides an extremely high level of performance and control. Our architecture allows you to serve content faster than legacy CDNs due in part to our ability to cache frequently and unpredictably changing content. We call this type of content “event-driven.” APIs can be classified as event-driven content since they contain frequently and unpredictably changing information. Fastly is trusted by technologically innovative companies including Vimeo, Pinterest, The New York Times, and GitHub.
Example of traffic flow with Fastly

A key benefit of our edge cloud platform is the ability to implement more logic at the edge of the network, allowing customers to implement functionality such as device detection, geo-fencing, request routing, and token authentication. If requests come via API call, Fastly can serve cached responses based on predefined custom logic for faster response times.

Most legacy CDNs are unable to cache APIs, but Fastly makes it easy to deliver tailored user experiences in real time, regardless of device type or network. In short, we give you an unmatched level of performance and control on a single, secure platform.

Clear outdated content in a flash

Fastly’s global content invalidation has a mean purge time of 150 milliseconds. This near real-time purge capability is fundamental to caching event-driven content such as API responses. Legacy CDNs can take hours or even days to fully clear content from cache, making them unable to cache APIs. Fastly’s Instant Purge of outdated API content delivers consistency across all channels.

Easily update content site-wide

To simplify the potentially complicated task of triggering multiple purges of related content, Fastly makes it possible to combine several API purge requests into a single action with Surrogate Keys. Administrators can tag APIs with keywords to create relational dependencies such as by product type, list price, and manufacturer.

For example, if the API call for a computer keyboard called “ABC” is tagged with categories “peripherals” and “keyboards,” then a purge request can clear the specific item, the product category, or all peripherals. Furthermore, if the marketing department initiates a
10% sale on peripherals, all cached peripherals pricing data can be immediately invalidated and refreshed with new pricing. So customers always see the most up-to-date information.

**Deliver customized content based on location**

With Fastly’s geolocation module, you can pinpoint the location of an end-user down to their longitude, latitude, continent, country, city, postal code, telephone area code, and metro code. Using this information, you can create virtual boundaries (geo-fences) to direct users to localized content and logic.

With geolocation, you can deliver localized content like coupons to a city or zip code, serve versions of your site in different languages by location, and show product availability by region — all the while keeping your API content cached at the edge.

**Serve optimized content for every type of device**

The ability to detect mobile devices and customize content in real-time allows companies to create the best possible user experience. By enabling Device Detection on Fastly, you can serve content optimized for a specific platform via API.

For instance, if a user visits your site from an iPhone, you’ll want to serve images with smaller dimensions so as to fit the smaller screen area, but with higher resolution to take advantage of the iPhone’s retina display. When serving content to a user viewing your site with a different smartphone, you can automatically send them lower resolution images that utilize less bandwidth.

**Gain real-time performance insights**

To monitor API usage and health, Fastly makes it very easy for you to get real-time, secure access to your log data. You gain visibility into how your users are engaging with your content, allowing you to identify trends and resolve any API delivery problems. Moreover, you can monitor the impact of new code deployments and, with Fastly’s versioned platform, roll back to previous stable code in case of issues.

“Fastly’s geolocation edge module allows us to determine where customers are coming from and make a logical guess based on their location as to which Boots site they’d like to access.”

— Peter Dawson
SOLUTION ARCHITECT
BOOTS UK
Secure your users’ information

Protecting your users’ identities and sensitive information, as well as the integrity and reputation of your website, requires a comprehensive approach to security. Since Fastly is a certified Level 1 Service provider, compliant with the Payment Card Industry Data Security Standard (PCI DSS), it meets the requirements for secure API delivery. While other CDNs have different levels of PCI certification for different physical networks, Fastly provides a single platform that is fully PCI compliant.

Fastly also provides Transport Layer Security (TLS), the next generation encryption protocol. As part of the standard service, you can terminate secure TLS connections at our network edge, closer to users, offloading encrypted API traffic from your web server for better performance.

To protect against API targeted distributed denial of service attacks (DDoS), the Fastly network of globally distributed POPs has capacity large enough to absorb attack traffic.

Full transparency and first class support

Unlike legacy CDN vendors who’s black box systems typically require customers to open a support ticket that can slow response times, Fastly delivers unmatched transparency and support to speed deployments and address uptime issues. A Fastly service is fully self-provisioning and quick to set up. You can have a full deployment delivering production traffic to your APIs in a matter of minutes.

We allow customers to view configurations and make changes themselves, or reach out to Fastly’s 24/7 dedicated customer support for assistance with some or all changes. Through Fastly’s developer-friendly control panel, you can easily make changes to your configurations—such as modifying the default TTL or enabling Gzip of assets for improved performance.

Fastly customers love our support. Our knowledgeable support engineers are available across a variety of channels to provide assistance in near real-time. It’s no wonder our customer satisfaction (CSAT) rating stands over 95% as of June 30, 2020.
The Fastly way

Optimized web and mobile content delivery benefits both you and your end-users. Your users get a better experience. Your faster sites can yield higher conversion and engagement rates. Caching API responses at the edge of our network enables you to attain these goals across all of your content — not just static assets. In this highly competitive world, delivering content via API at speed can be a major source of competitive advantage. Isn’t it about time you tried Fastly?

To learn more about how we can help support your modern app development and API strategy, contact Fastly at sales@fastly.com.

Fastly improves the Guardian website’s performance by caching dynamic content, such as HTML and API calls. Before Fastly, the Guardian was using a CDN for static content only, resulting in a slow experience for readers located geographically far from origin.