

EP48: Debugging Like A Pro



ALEX XU

FEB 25, 2023



Share



This week's system design refresher:

- Debugging Like A Pro (Youtube video)
- Load balancer vs. API gateway
- ChatGPT timeline
- Video content uploading on Youtube
- A beginner's guide to CDN

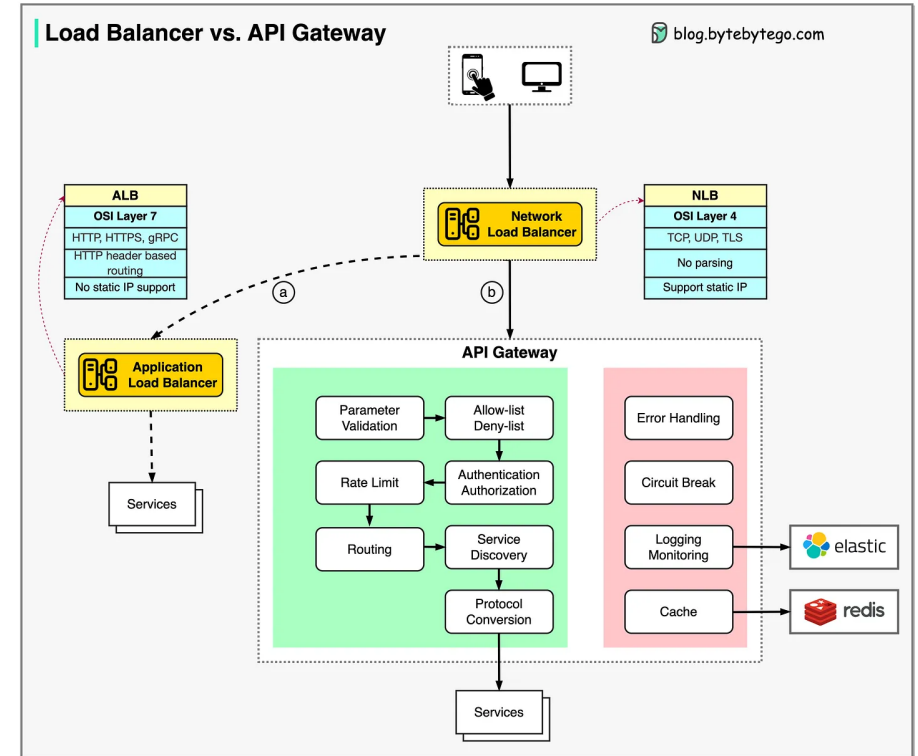
Debugging Like A Pro

Debugging Like A Pro



What are the differences between a load balancer and an API gateway?

First, let's clarify some concepts before discussing the differences.



1. NLB (Network Load Balancer) is usually deployed before the API gateway, handling traffic routing based on IP. It does not parse the HTTP requests.
2. ALB (Application Load Balancer) routes requests based on HTTP header or URL and thus can provide richer routing rules. We can choose the load balancer based on routing requirements. For simple services with a smaller scale, one load balancer is enough.
3. The API gateway performs tasks more on the application level. So it has different responsibilities from the load balancer.

The diagram below shows the detail. Often, they are used in combination to provide a scalable and secure architecture for modern web apps.

Option a: ALB is used to distribute requests among different services. Due to the fact that the services implement their own rating limitation, authentication, etc., this approach is more flexible but requires more work at the service level.

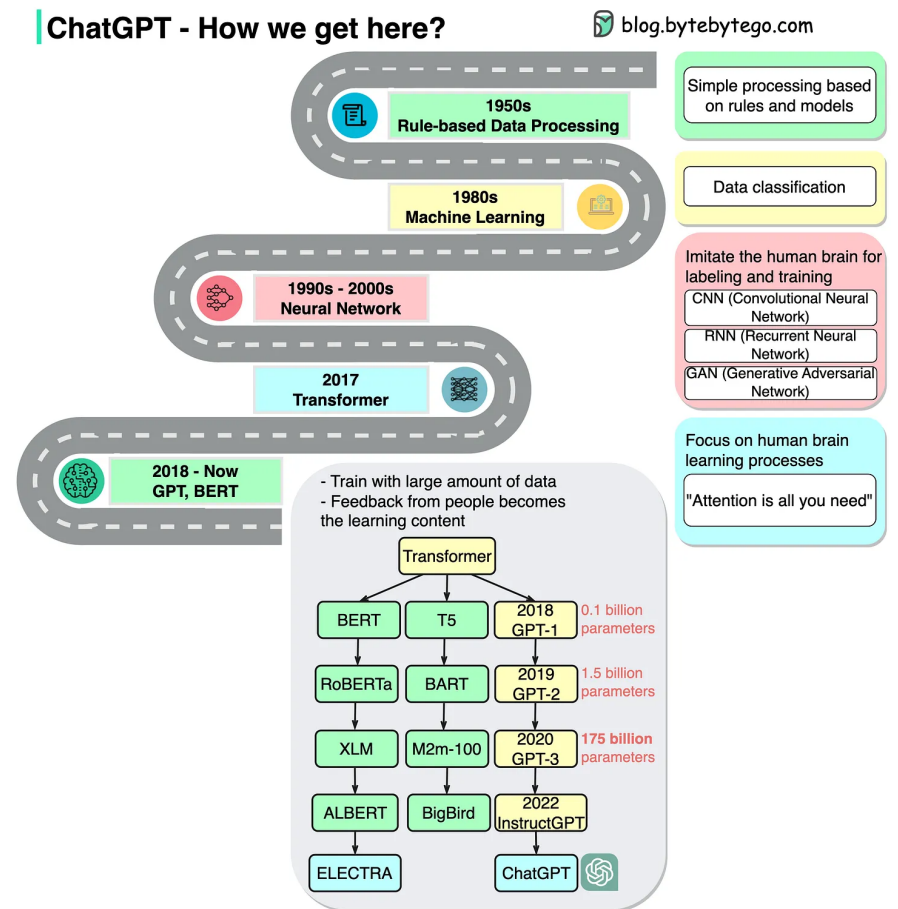
Option b: An API gateway takes care of authentication, rate limiting, caching, etc., so there is less work at the service level. However, this option is less flexible compared with the ALB approach.

Over to you: Which one should we use, a load balancer or an API gateway?

ChatGPT - timeline

A picture is worth a thousand words. ChatGPT seems to come out of nowhere. Little did we know that it was built on top of decades of research.

The diagram below shows how we get here.



- 1950s
In this stage, people still used primitive models that are based on rules.
- 1980s
Since the 1980s, machine learning started to pick up and was used for classification. The training was conducted on a small range of data.
- 1990s - 2000s
Since the 1990s, neural networks started to imitate human brains for labeling and training. There are generally 3 types:
 - CNN (Convolutional Neural Network): often used in visual-related tasks.
 - RNN (Recurrent Neural Network): useful in natural language tasks
 - GAN (Generative Adversarial Network): comprised of two networks(Generative

and Discriminative). This is a generative model that can generate novel images that look alike.

- 2017
“Attention is all you need” represents the foundation of generative AI. The transformer model greatly shortens the training time by parallelism.
- 2018 - Now
In this stage, due to the major progress of the transformer model, we see various models train on a massive amount of data. Human demonstration becomes the learning content of the model. We’ve seen many AI writers that can write articles, news, technical docs, and even code. This has great commercial value as well and sets off a global whirlwind.

Over to you: What is the next breakthrough for AI models? Can you guess?

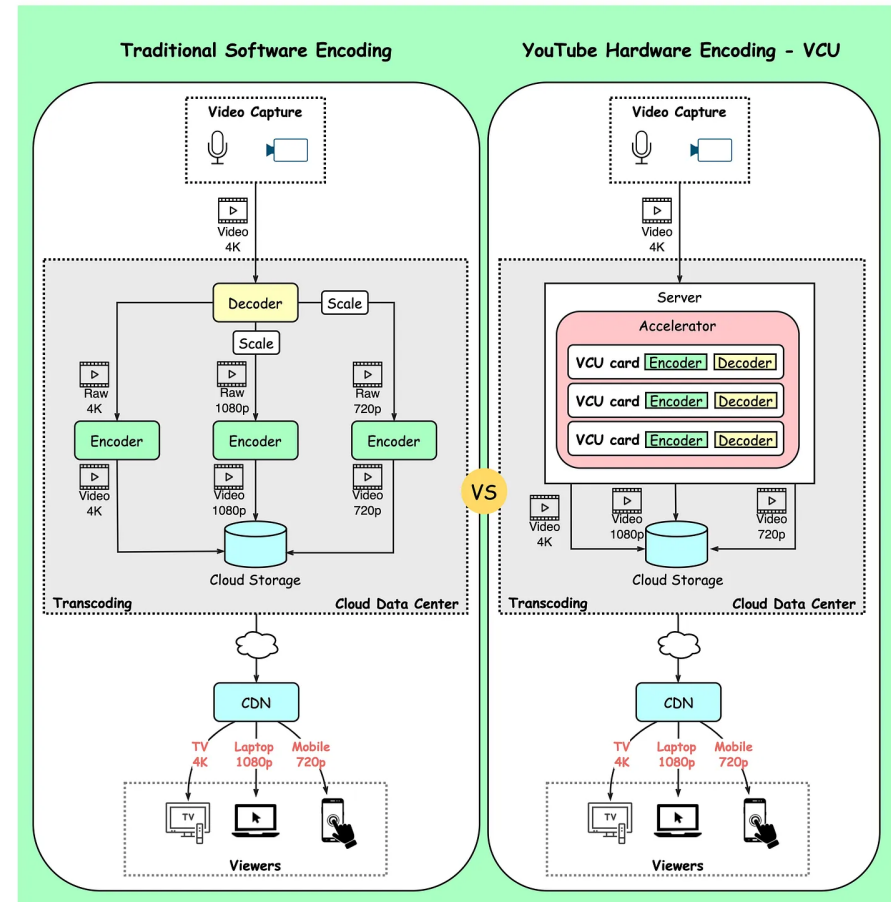
How does Youtube handle massive video content upload?

YouTube handles 500+ hours of video content uploads every minute on average. How does it manage this?

The diagram below shows YouTube’s innovative hardware encoding published in 2021.

How does YouTube Handle Massive Video Content Upload?

blog.bytebytego.com



• Traditional Software Encoding

YouTube’s mission is to transcode raw video into different compression rates to adapt to different viewing devices - mobile(720p), laptop(1080p), or high-resolution TV(4k).

Creators upload a massive amount of video content on YouTube every minute. Especially during the COVID-19 pandemic, video consumption is greatly increased as people are sheltered at home. Software-based encoding became slow and costly.

This means there was a need for a specialized processing brain tailored made for video encoding/decoding.

- **YouTube's Transcoding Brain - VCU**

Like GPU or TPU was used for graphics or machine learning calculations, YouTube developed VCU (Video transCoding Unit) for warehouse-scale video processing.

Each cluster has a number of VCU accelerated servers. Each server has multiple accelerator trays, each containing multiple VCU cards. Each card has encoders, decoders, etc. [1]

VCU cluster generates video content with different resolutions and stores it in cloud storage.

This new design brought 20-33x improvements in computing efficiency compared to the previous optimized system. [2]

Over to you: Why is a specialized chip so much faster than a software-based solution?

Reference:

[1] dl.acm.org/doi/abs/10.1145/3445814.3446723

[2] blog.youtube/inside-youtube/new-era-video-infrastructure/

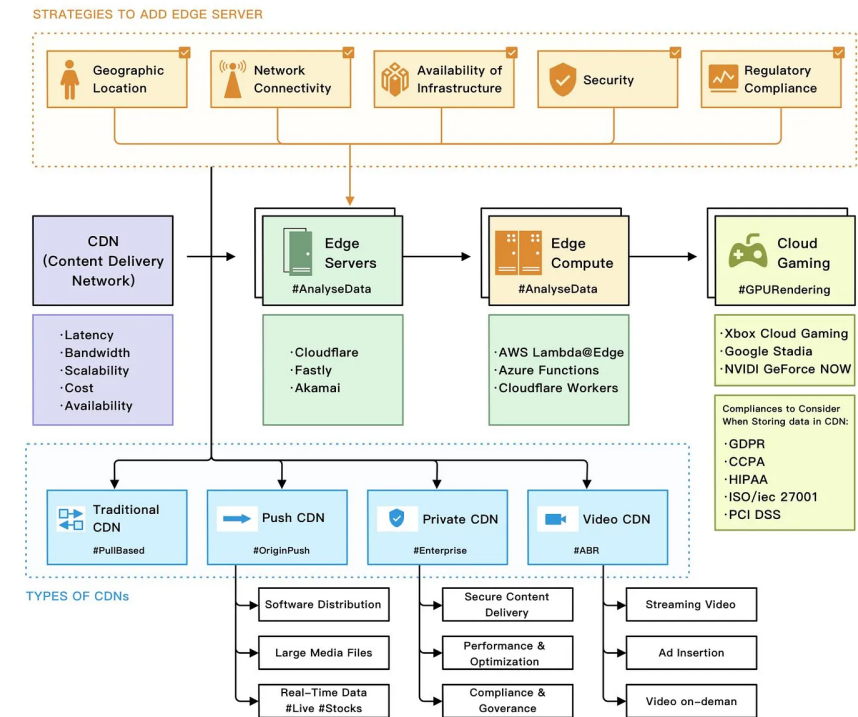
A beginner's guide to Content Delivery Network (CDN)

A guest post by Love Sharma. The link to the complete article can be found in the comment section.

This article is written by guest author [Love Sharma](#). You can read the full article [here](#).

A Beginner's Guide to CDN

ByteByteGo.com



CDNs are distributed server networks that help improve the performance, reliability, and security of content delivery on the internet.

Here is the Overall CDN Diagram explains:

Edge servers are located closer to the end user than traditional servers, which helps reduce latency and improve website performance.

Edge computing is a type of computing that processes data closer to the end user rather than in a centralized data center. This helps to reduce latency and improve the performance of applications that require real-time processing, such as video streaming or online gaming.

Cloud gaming is online gaming that uses cloud computing to provide users with high-quality, low-latency gaming experiences.


Together, these technologies are transforming how we access and consume digital content. By providing faster, more reliable, and more immersive experiences for users, they are helping to drive the growth of the digital economy and create new opportunities for businesses and consumers alike.

© 2024 ByteByteGo · [Privacy](#) · [Terms](#) · [Collection notice](#)
[Substack](#) is the home for great writing




124 Likes


2 Comments





Write a comment...


 Mert Demirok Feb 25, 2023


A specialized chip like VCU is so much faster than a software-based solution because it is designed specifically for the task of video encoding and decoding. Traditional CPUs are designed to handle a wide range of tasks, whereas specialized chips like VCUs are tailored for specific tasks, making them more efficient at performing those tasks.

 LIKE (4)

 REPLY


 SHARE





 Joe V Mar 2, 2023

Traditional CPUs like Intel/ARM processors are good for general purpose compute like rendering UI, network data processing etc. Video transcode involves taking 20 frames per sec 4k video, decode and compress them using H264 encoder to different resolutions. General purpose cpu cores nor GPUs are not good for it. You have DSP (Digital Signal Processors) cores specifically built for it. You pipe data to the DSP core and take its output and pipe it to CDN data store.

These days Intel/ARM processors package separate DSP cores, GPU cores on the same chip to help with the above activities but at a lower scale.

 LIKE (1)

 REPLY

 SHARE

