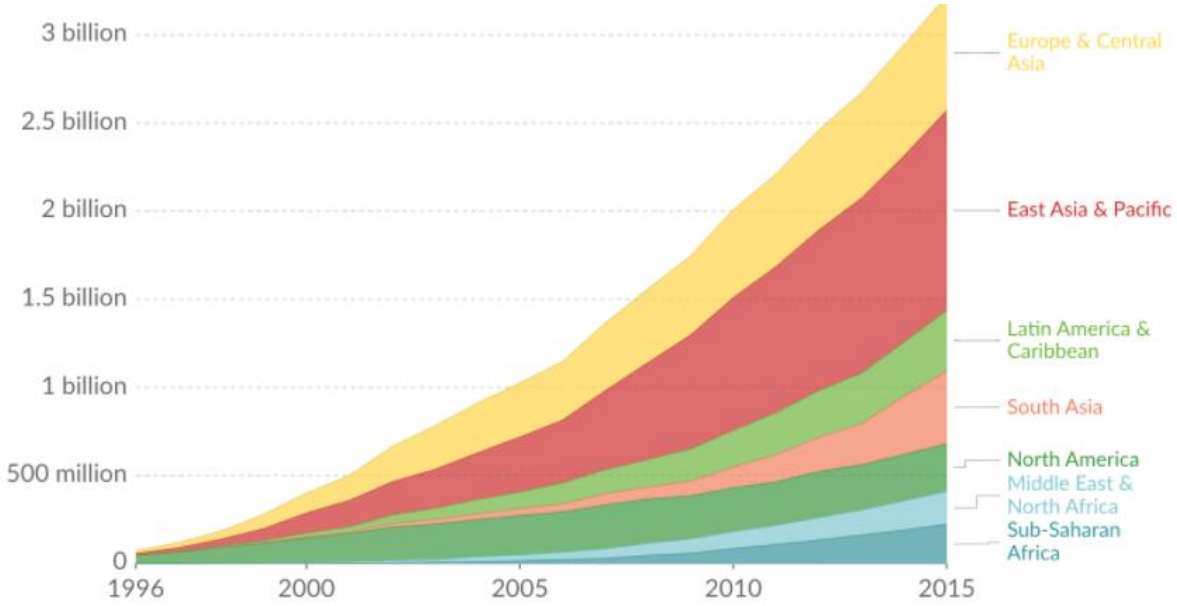
Set 13. Scalable Systems

Skill 13.01: Explain what is meant by a scalable system

Skill 13.01 Concepts

The Internet began its life as a network connecting universities and research centers. Once it became available and affordable for consumers, it shot up in popularity and is now used by an estimated 4.5 billion people.



The number of users on the Internet from 1996 to 2015, grouped by region. *Chart source:* *Our World in Data*

Fortunately, the protocols powering the Internet and the Web were designed for scalability. A **scalable** system is one that can continue functioning well even as it experiences higher usage.

Some systems aren't scalable at all and can only handle exactly the amount of usage they were designed for.
Scalable systems can handle extra usage, but their capacity varies. Some systems might only scale to handle double the current usage; other systems might scale to handle 1000x the current usage.

When we're designing systems with potentially global reach—such as the Internet itself or applications that run on top of it—we need to always consider the scalability of our approach.

Skill 13.01 Exercise 1

Skill 13.02: Identity the features of the Internet the make it scalable

Skill 13.02 Concepts

What features increase the scalability of the Internet?

- Any computing device can send data around the Internet if it follows the protocols. There is no bureaucratic process that blocks a device from joining or prevents a programmer from learning how the protocols work.
- The IPv6 addressing system can uniquely address a *trillion trillion* times the amount of devices currently connected to the Internet.
- Routing is dynamic, so new routers can join a network at any time and help to move data packets around the Internet.

The Internet was designed to be scalable, but no system is infinitely scalable.

Skill 13.02 Exercises 1

Skill 13.03: Identify the limitations of scalability

Skill 13.03 Concepts

What threatens the scalability of the Internet? Or put another way, what could go wrong if every single device in the world connected to the Internet right now and attempted to download a movie?
Here are a few ideas:

- Network connections have limited bandwidth. A huge amount of data may flow easily through the very high bandwidth connections but it could easily overwhelm low bandwidth connections, leading to delays or dropped packets.
- Routers have limited throughput (the amount of data they can forward per second). A modern consumer router has a throughput around 1 Gbps while much more expensive enterprise routers can forward up to 10 Gbps. An average movie is around 1 to 10 GB, so a worldwide download-a-thon could get bottlenecked in the routers.
- Wireless routers often have a limitation in the number of devices that can be connected to them, typically up to 250 devices. If everyone tried to use a shared WiFi network at the same time (like in a university or library), they might find themselves simply unable to join.

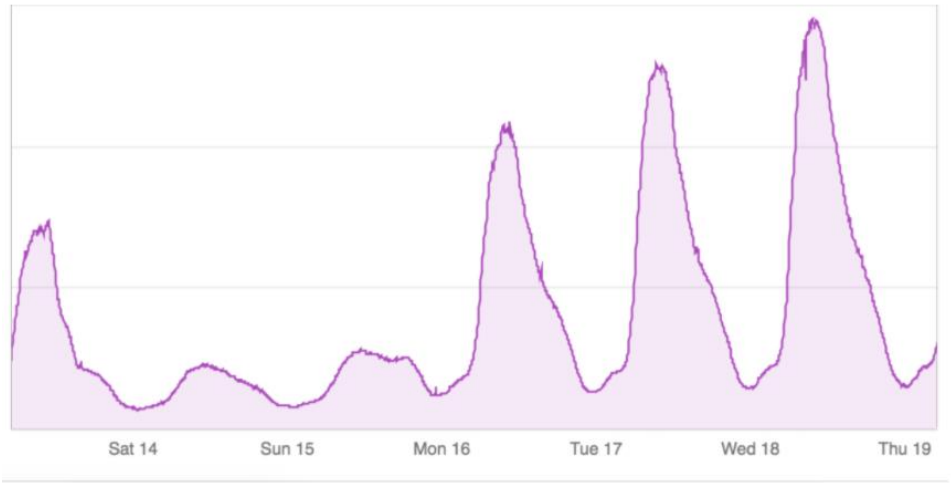
Skill 13.03 Exercises 1

Skill 13.04: Explain the importance of scalability as it applies to web applications

Skill 13.04 Concepts

A web application that runs on top of the Internet must also be scalable, whether it's an iPhone app, a website, or multiplayer game. Now that there are billions of people connected to the Internet, any application can suddenly experience a surge in users. If the application doesn't scale to meet the demand, users might experience increased latency or a complete outage. 😬

During the COVID-19 pandemic, people across the globe were asked to stay indoors to decrease infection and many of them rushed to online services for virtual versions of what they were missing in person. Due to the many students turning to Khan Academy for example, their website experienced a 250% increase in server load.

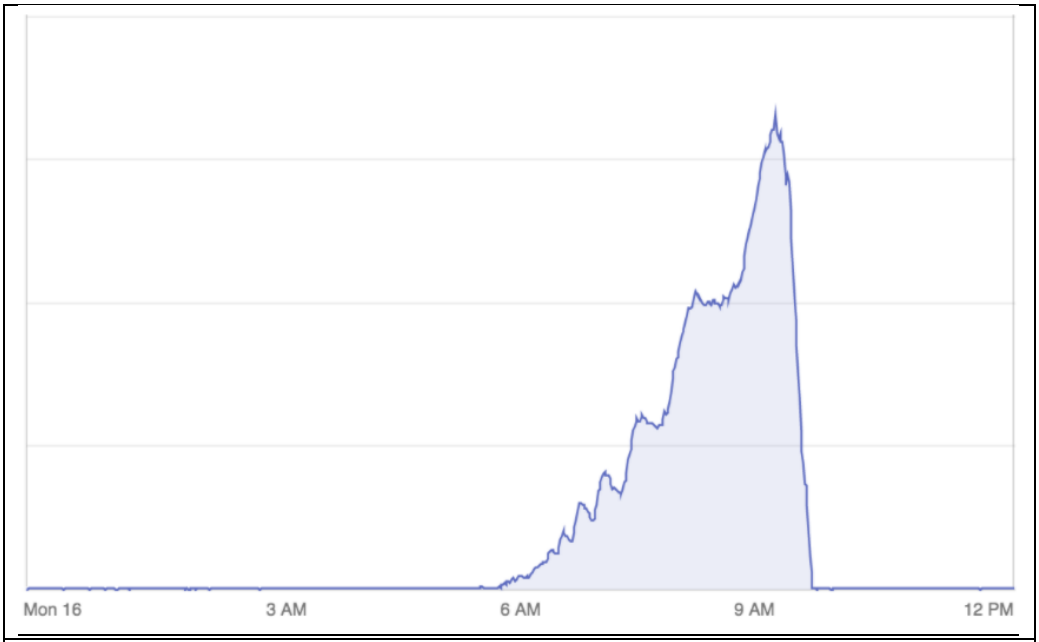


Requests to Khan Academy server from Friday, March 13 to Wednesday, March 18. Usage is always lower on Saturday and Sunday.

According to Khan Academy, that high usage took their systems by surprise, and a few of them were just barely handling the sudden onslaught of requests. For example, their system to notify students about new assignments was getting backlogged, so notifications would take a few minutes to show up (versus a few seconds).

The graph shows the backlog of tasks in the assignment notification system, with time on the x-axis and number of not-yet-completed tasks on the y-axis. The backlog typically contains close to zero tasks because the system executes the tasks as soon as they are created, but in this case, the system was assigned tasks more quickly than it could handle and accumulated a large backlog.

Fortunately, Khan Academy engineers were able to quickly bump up the capacity of those systems, and most users never noticed anything amiss.



Backlog of tasks in the assignment notification system. The engineers increased the capacity of the system at 9:30.

Skill 13.04 Exercise 1

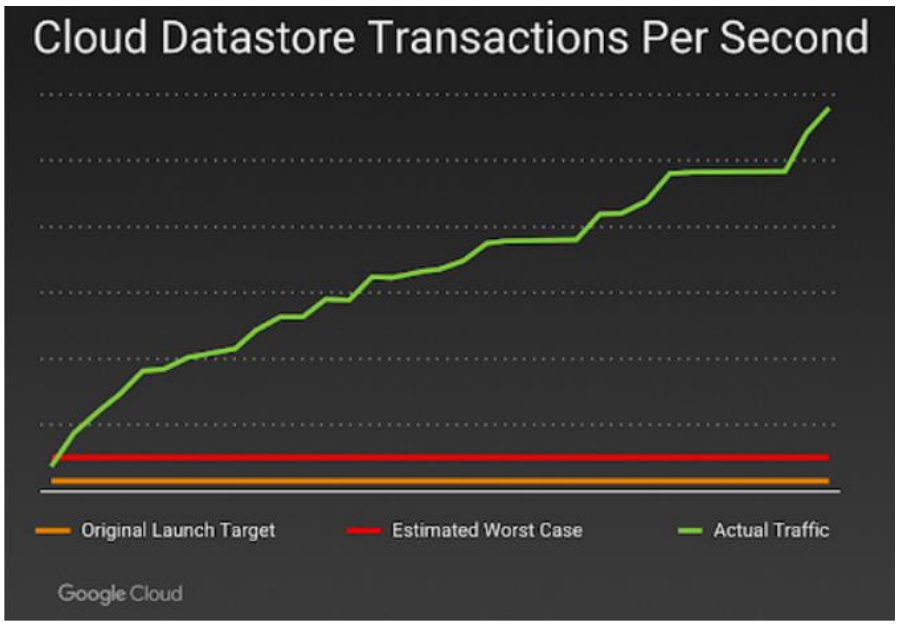
Skill 13.05: Describe load testing and how it can help engineers prepare for spikes

Skill 13.05 Concepts

Engineering teams can prepare for spikes in usage by doing **load testing**: simulating high amounts of traffic in a short period of time to see if the system buckles under the load. Load testing can uncover bottlenecks or hard-coded limits in the system.

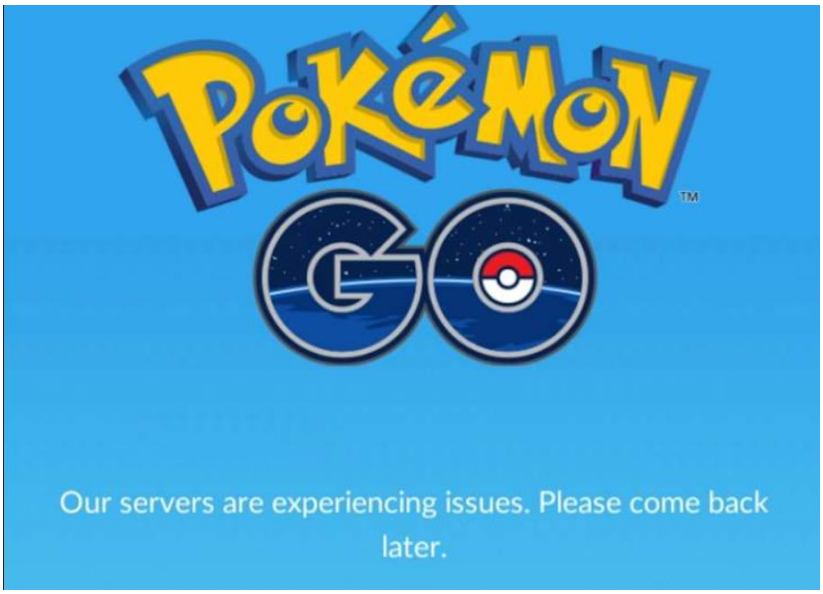
Ever played Pokémon Go? It's a mobile game that came out in the summer of 2016 and was an instant hit. The game developers did load testing before releasing the game, simulating 5 times the highest amount of traffic they expected, and the game servers handled it just fine.

They vastly underestimated the popularity of Pokémon Go, however. On launch day, their servers saw *50 times* the estimated traffic.



The transactions per second in the Pokémon Go datastore. The actual transactions vastly exceeded the estimations. *Source: Google Cloud Blog*

The game servers weren't ready for that level of extreme load, so many players were greeted with a disappointing screen:



The team scrambled to improve the scalability of the system, amidst growing demand from frustrated users plus multiple DDoS[1] attacks on their servers from cybercriminals.

After reconfiguring their server architecture to be more scalable, the team released Pokémon Go to the rest of the world. In the three years since, it's been downloaded more than a billion times from mobile app stores.

Skill 13.05 Exercise 1

[1] A distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic.