
Software Scalability

CS480 Software Engineering

Yu Sun, Ph.D.

<http://yusun.io>

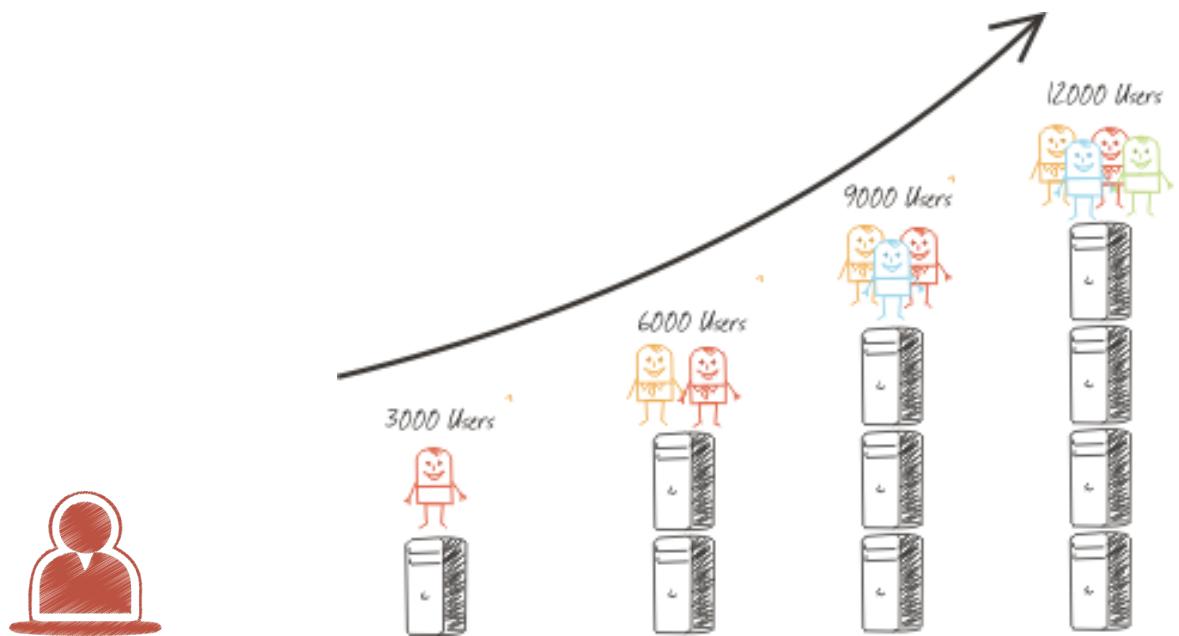
yusun@cpp.edu



CAL POLY POMONA

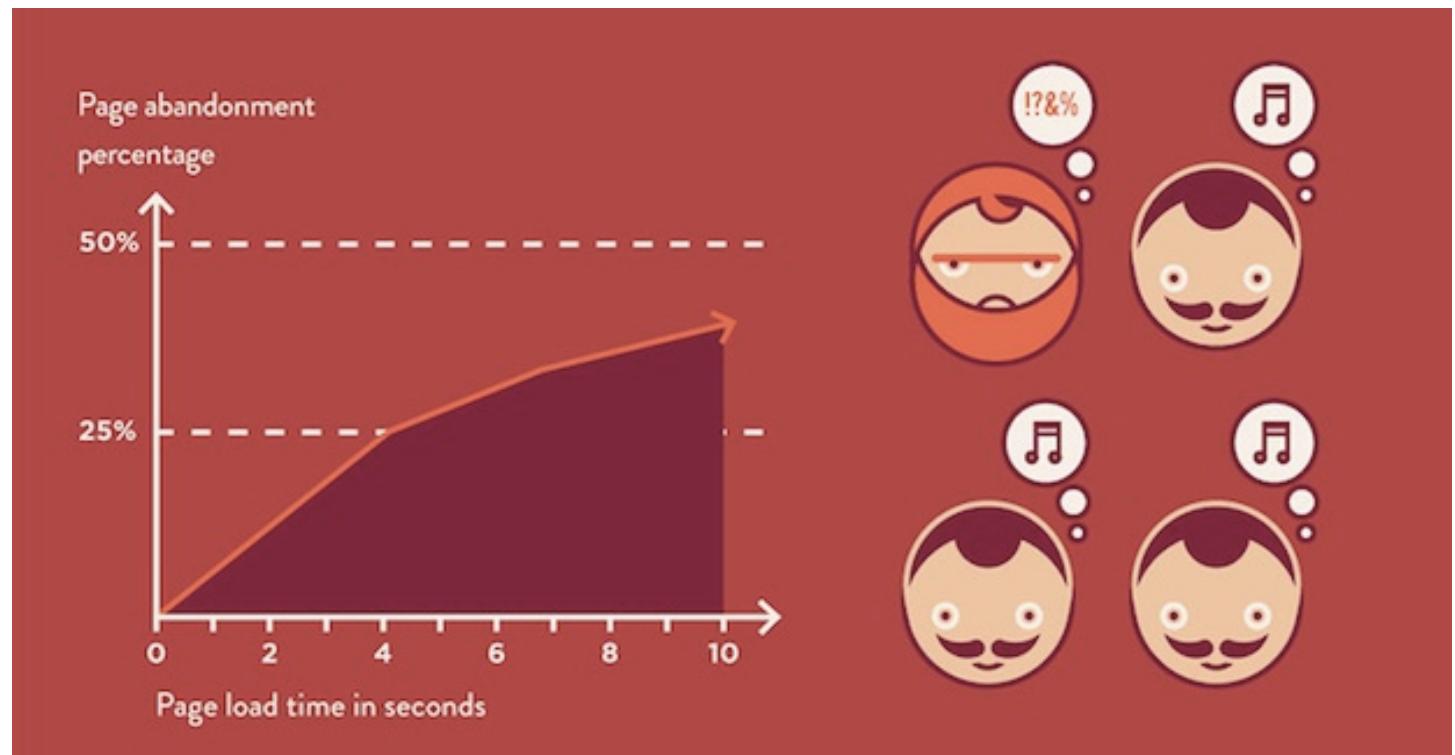
Software Scalability

- ◆ *Scalability* is the ability of a system to handle a growing amount of work in a capable manner or its ability to be enlarged to accommodate that growth



Amazon.com

- ◆ 426 items were sold per second during Christmas
- ◆ A page load slowdown of just one second could cost it \$1.6 billion in sales each year



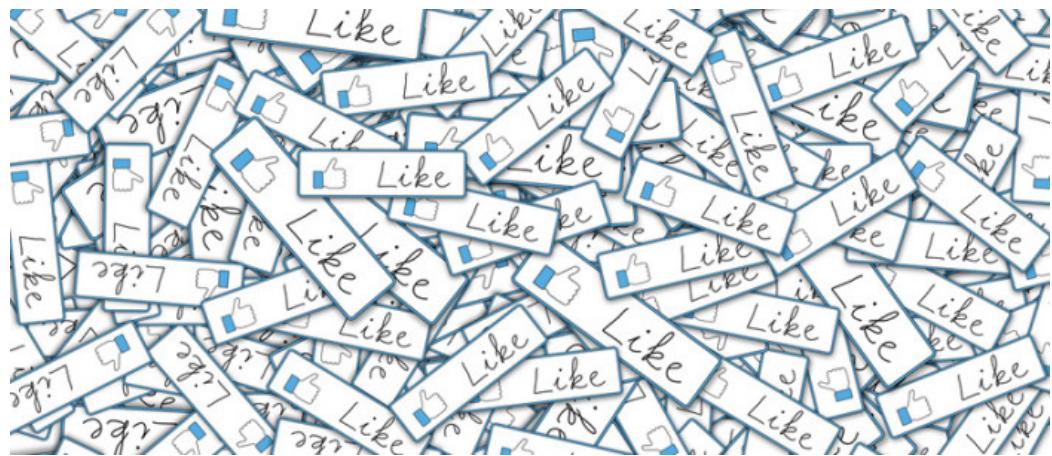
Google

- ◆ 3.5 billion searches / day
- ◆ 1.2 trillion searches / year
- ◆ “by slowing its search results by just 4/10 of a second they could lose 8 million searches per day”



Facebook

- ◆ People spend over 700 billion minutes per month on Facebook
- ◆ In 20 minutes 10.2 million comments are posted
- ◆ 750 million photos were uploaded to Facebook over New Year's weekend

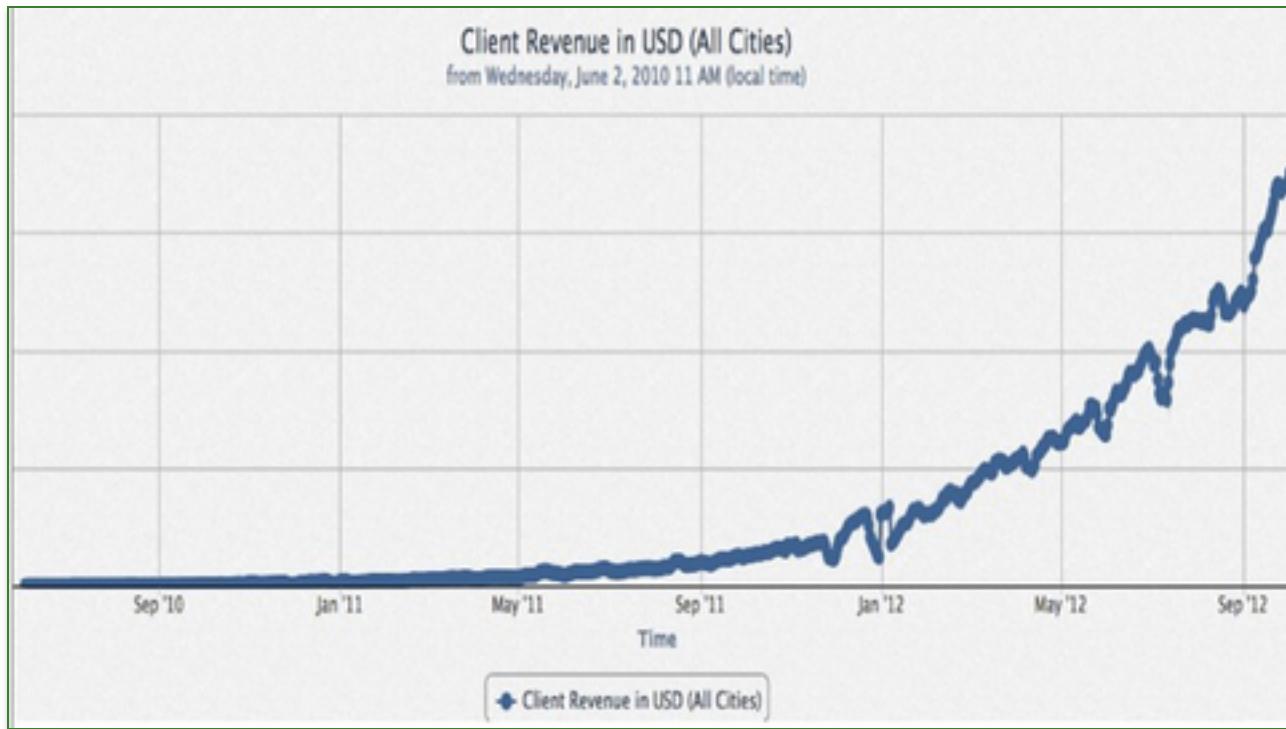


Amazon S3

- ◆ 1.3 trillion objects stored
- ◆ 1.1 million requests / second



Uber Growth



BroncoDirect



Cal Poly Pomona

February 10 at 11:47am ·

Broncos, if you're trying to register for classes right now, you've probably noticed that BroncoDirect is having some serious issues right now.

We apologize for this. Really, we do. We know that you don't need this extra aggravation while you're trying to get your classes for next quarter.

We are working as quickly as we can to fix the situation. We will keep you updated as much as possible.

Again, our sincere apologies.



Like · Comment · Share · 106 29 1

Ian I pray Lord for this matter to be resolved...Amen.

· February 11 at 11:56am

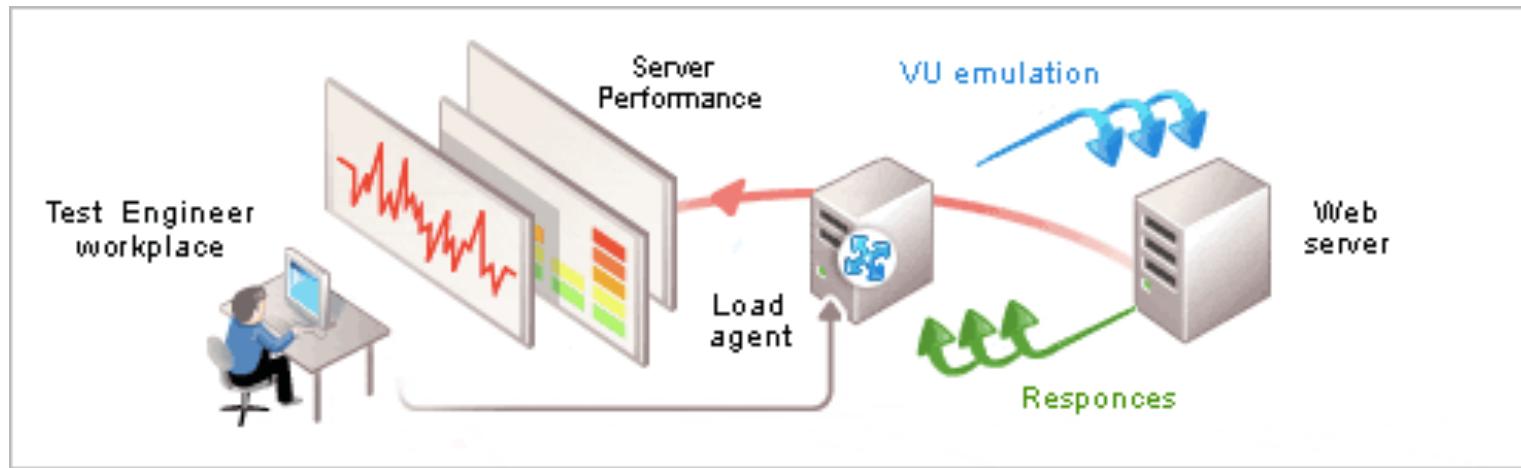
LOL Broncodirect... the worst thing about Cal Poly Pomona.

February 10 at 1:17pm

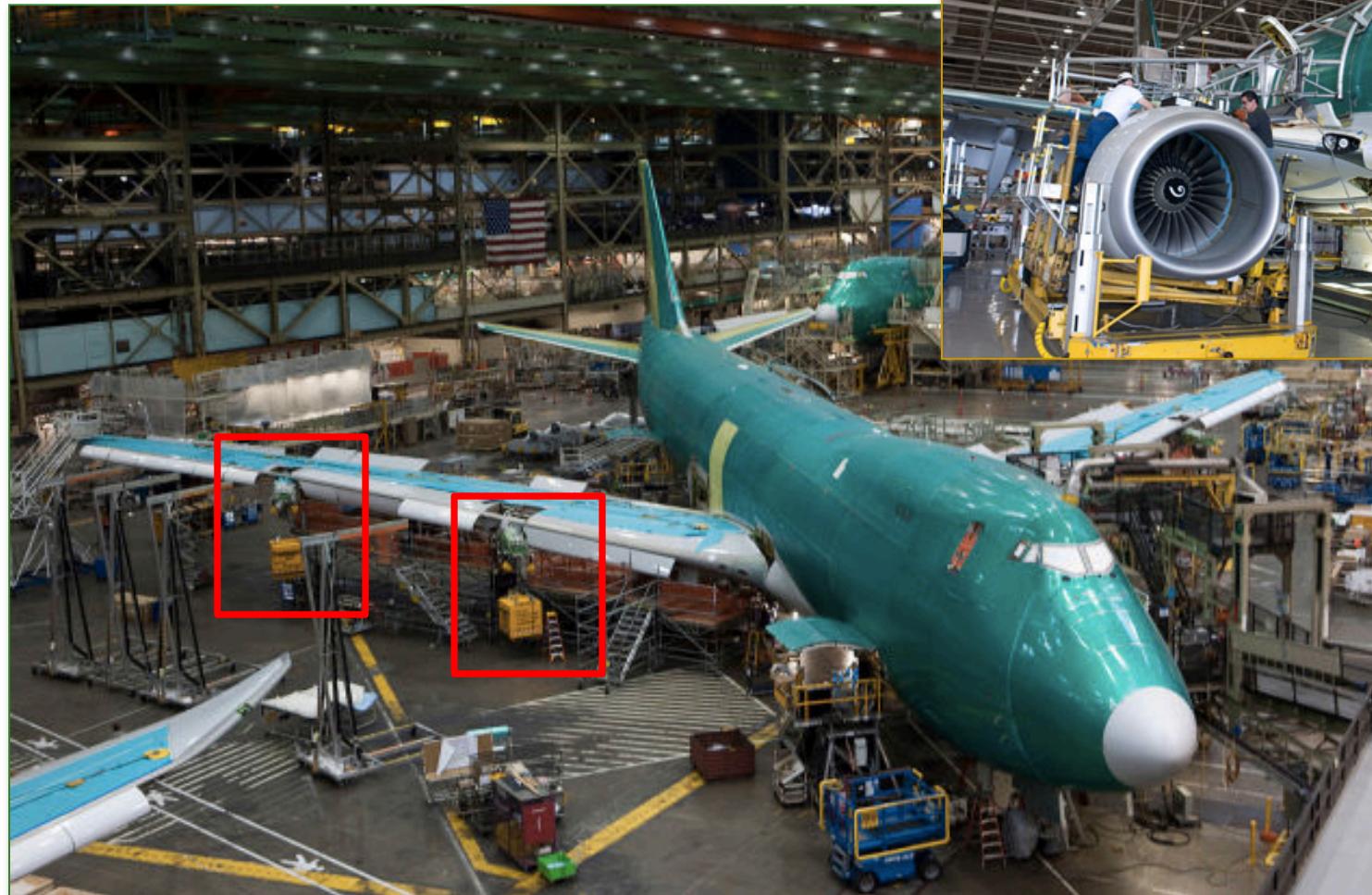
Gabriel Horowitz Lol one of the most stressful parts about going to school at CPP

Michelle Cassidy Every college has these problems during registration, I don't blame cpp, hope it's resolved soon though!

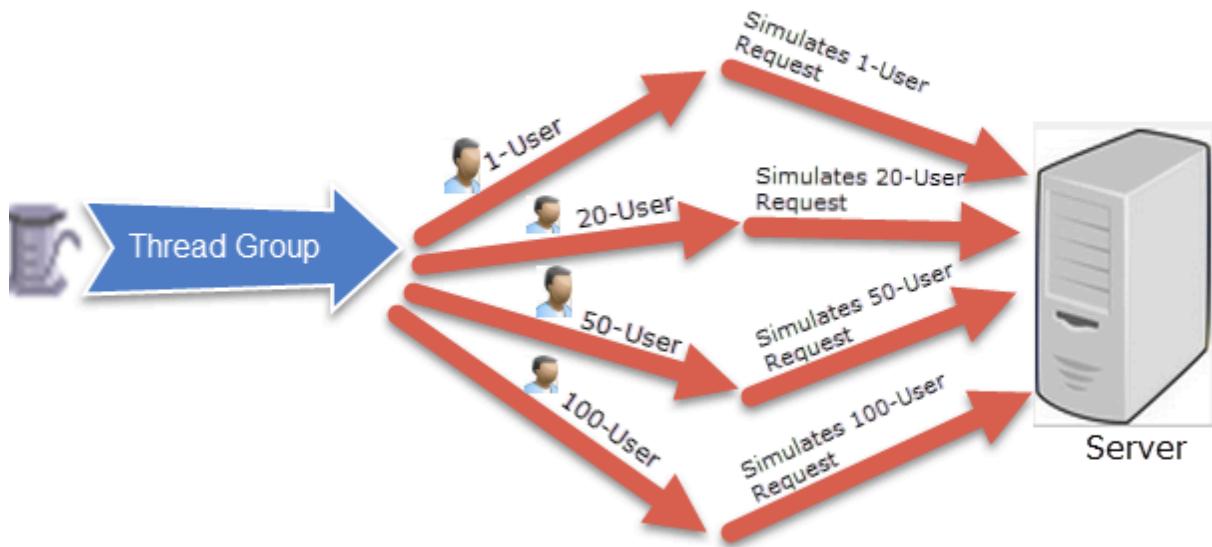
Scalability Verification – Load Test



Scalability Verification – Load Test



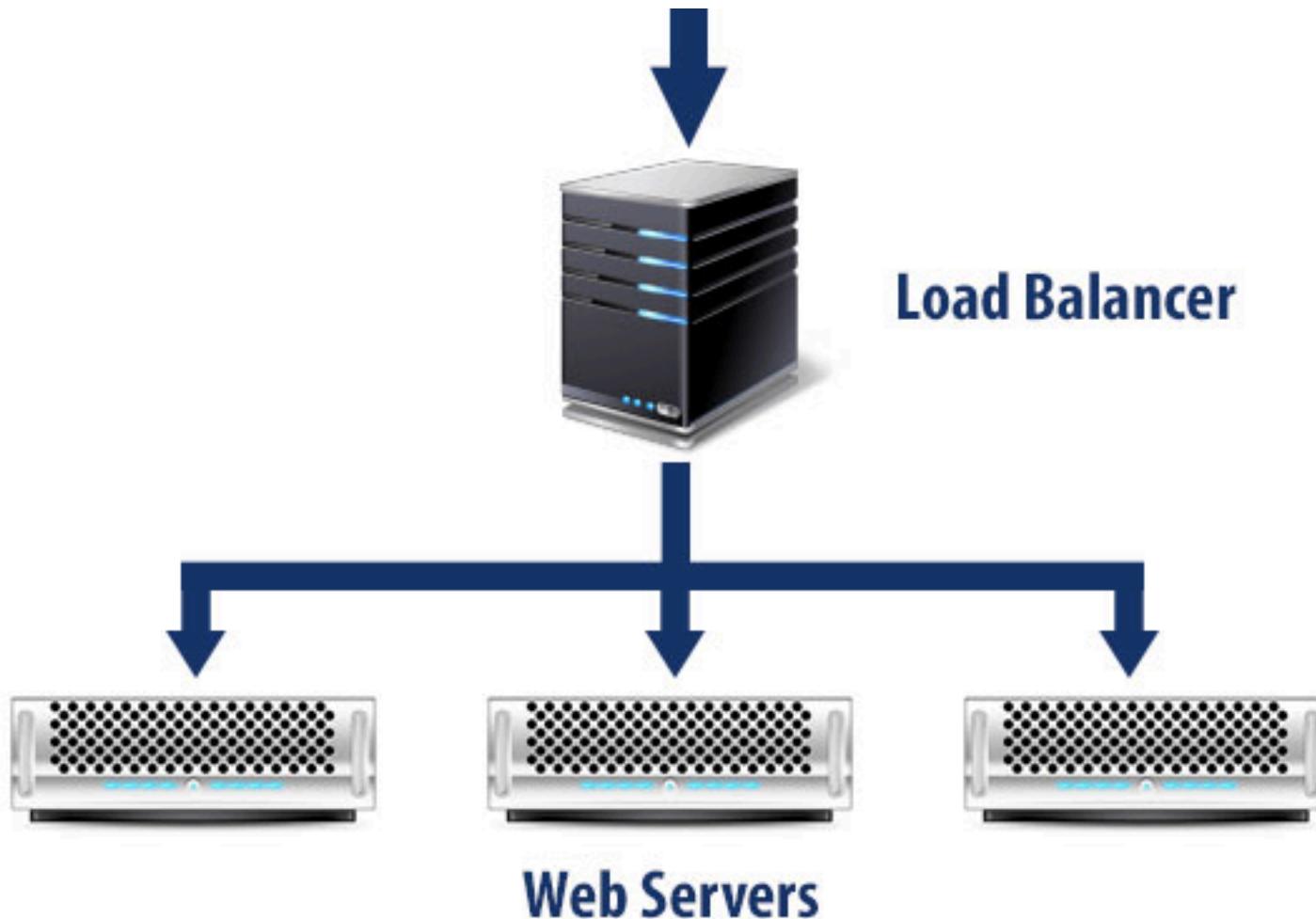
JMeter Demo



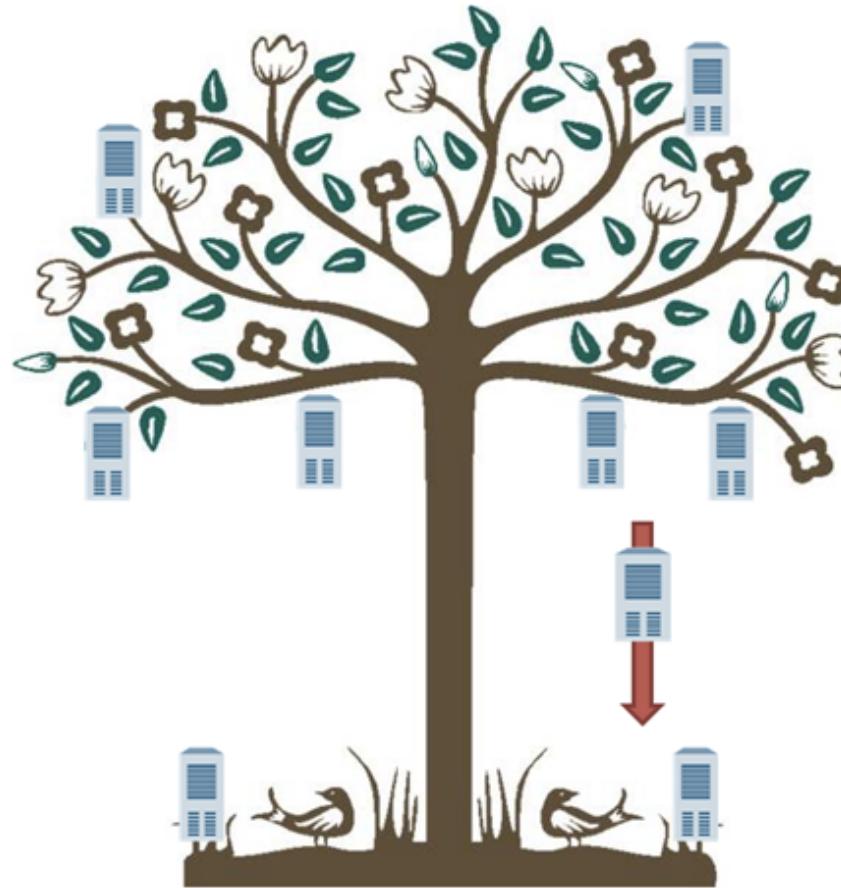
How to Improve Scalability?



How to Improve Scalability?



How to Make Scalability Easy?



Cloud Computing

- ◆ Cloud computing shifts computing from local dedicated resources to distributed, virtual, elastic, multi-tenant resources
 - ◆ On-demand access to computing, storage, and software services
 - ◆ Based on a utility cost model



Cloud Computing & Amazon

- ◆ The popularization of the term can be traced to 2006 when Amazon.com introduced the Elastic Compute Cloud (EC2)



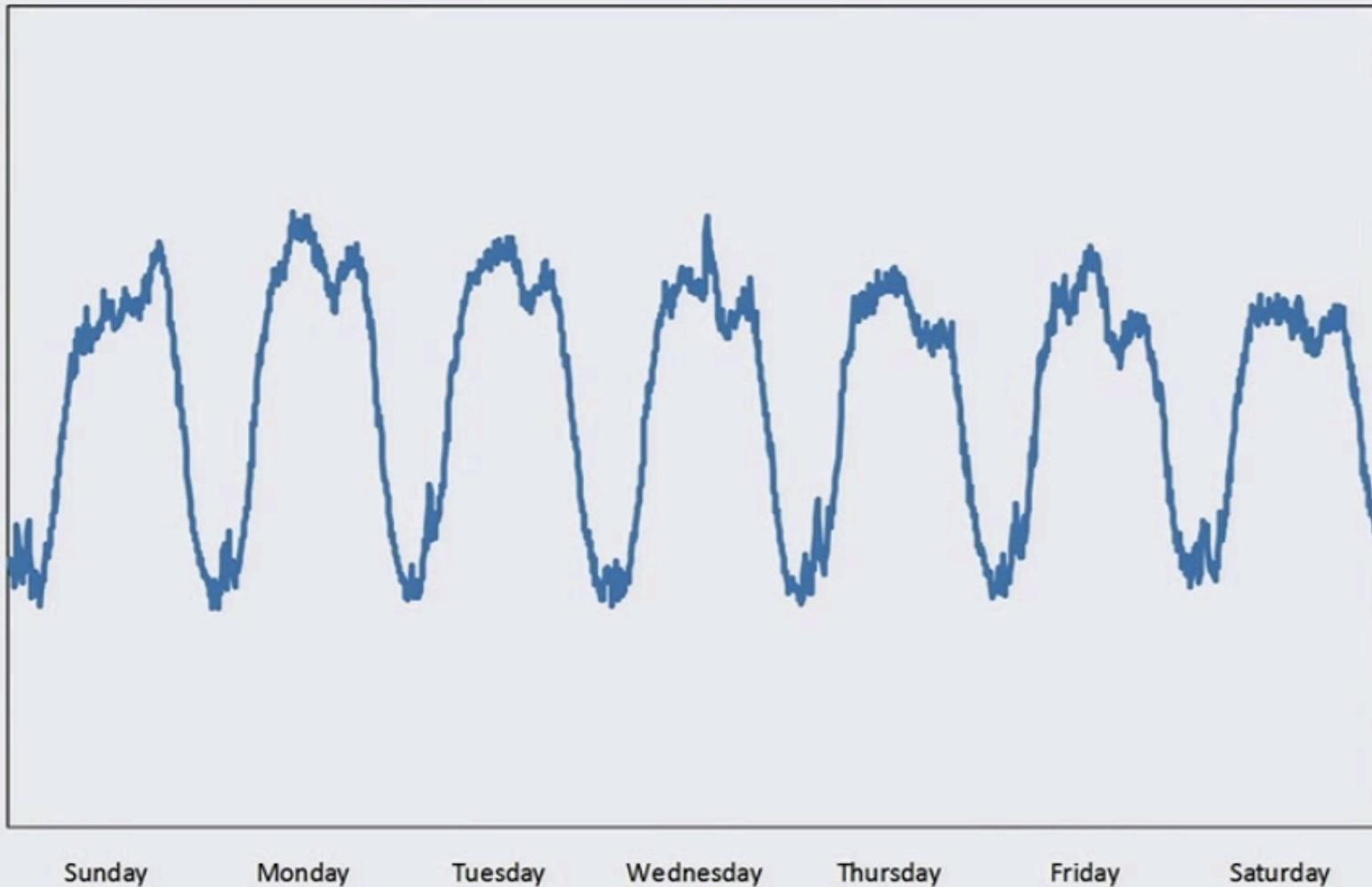
Case Study: Amazon Retail Website

The screenshot shows the Kindle Store section of the Amazon website. At the top, there's a banner for "Best Books of 2011 So Far". Below it, a large image of a Kindle device with the text "Kindle The #1 Bestseller on Amazon". To the left, there's a sidebar for "Kindle Store" with links for buying Kindle devices and reading apps. The main content area shows Kindle devices for sale: Kindle (\$139), Kindle 3G with Special Offers (\$139), and Kindle 3G (\$189). There are also sections for "Kindle Daily Post" (our editors' blog) and "Great Deals on Kindle Accessories". A "More Items to Consider" section lists books like "THE INNOVATOR'S DNA", "BASIC ECONOMICS", "The Innovator's Dilemma", and "venture deals". A "Bestsellers" section shows the top 100 paid Kindle books, with "Hidden in Plain View" by Darryl Billups at the top.



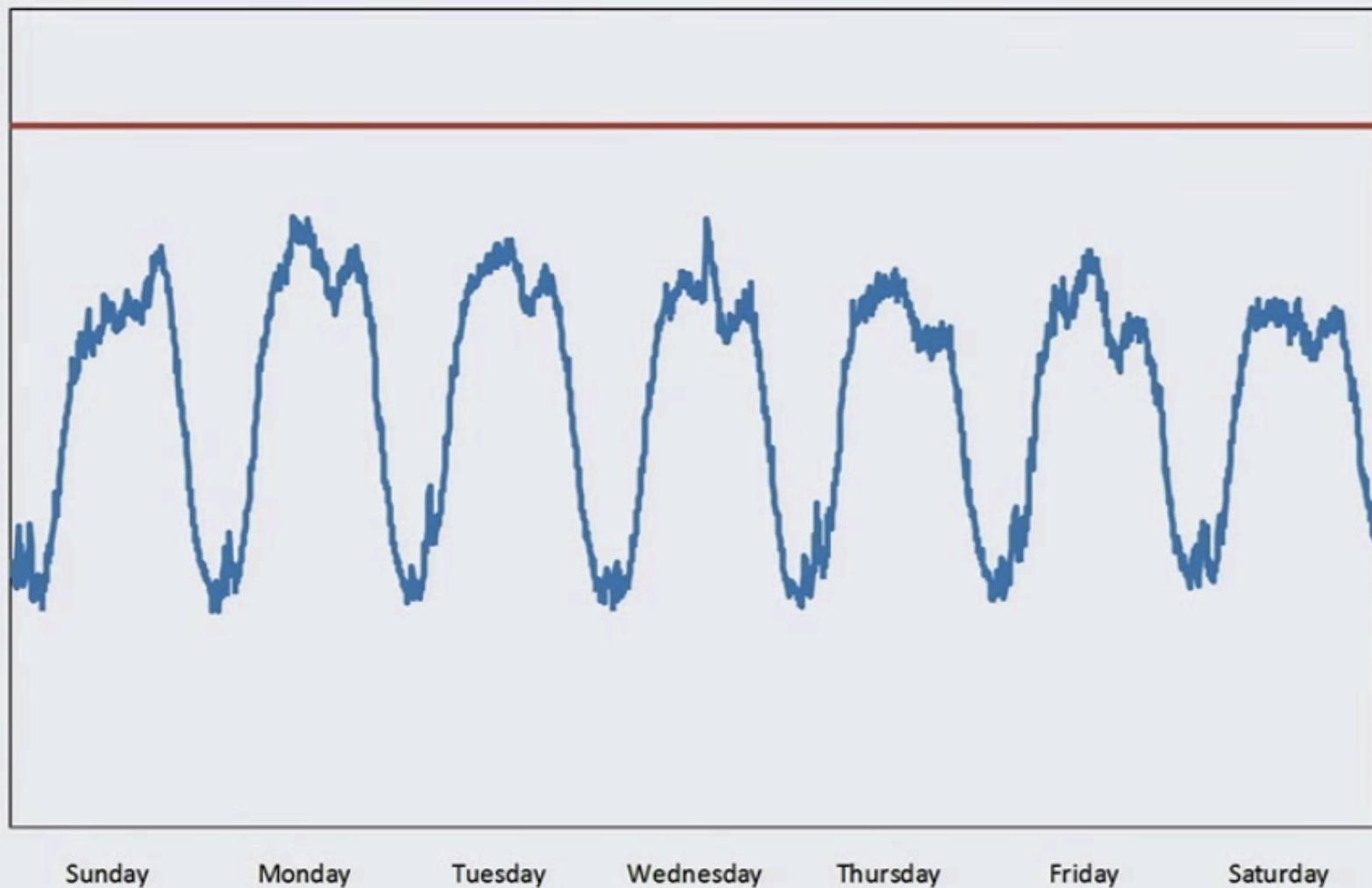
Case Study: Amazon Retail Website

Typical Weekly Traffic to amazon.com



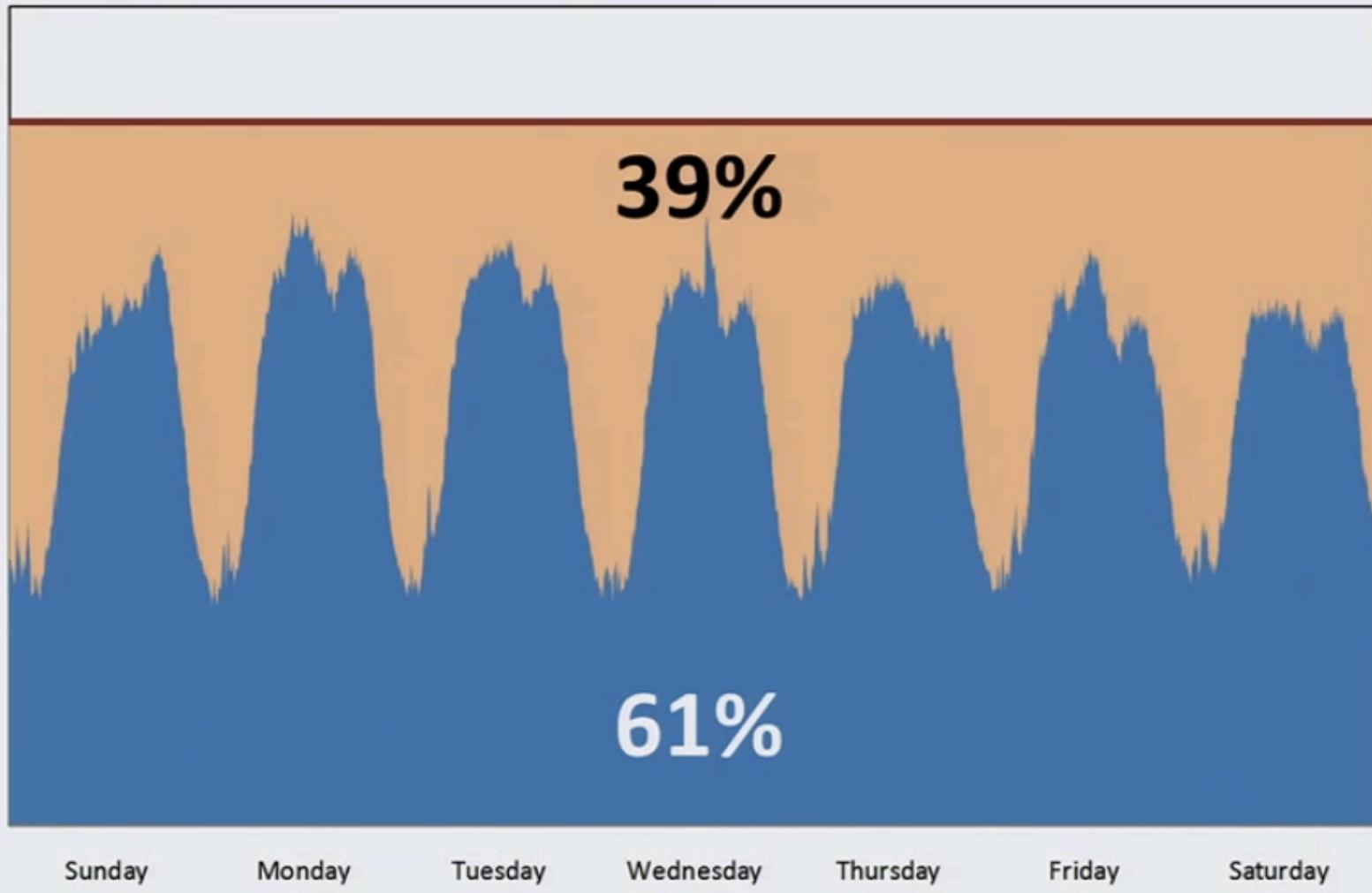
Case Study: Amazon Retail Website

Typical Weekly Traffic to amazon.com



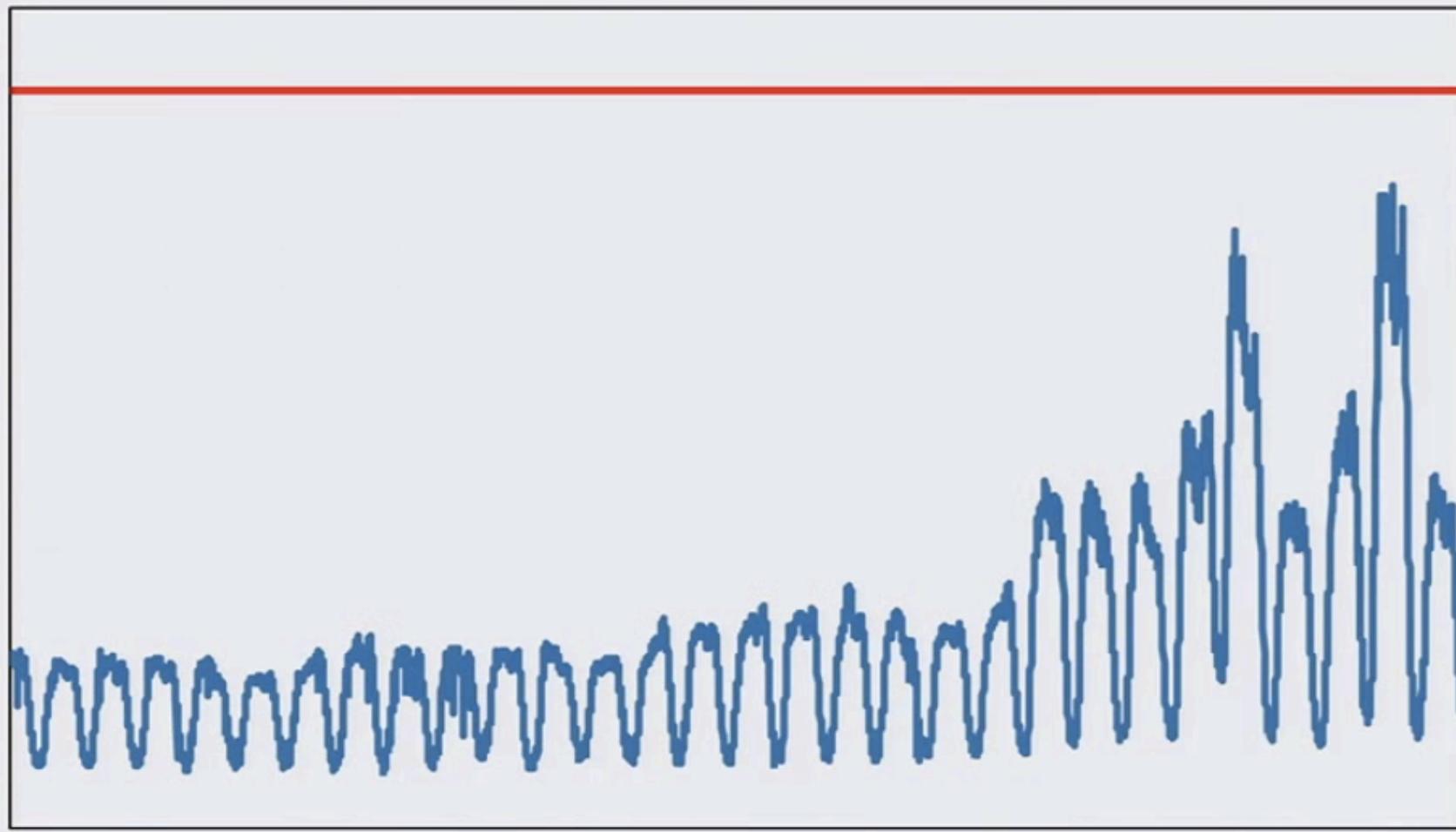
Case Study: Amazon Retail Website

Typical Weekly Traffic to amazon.com



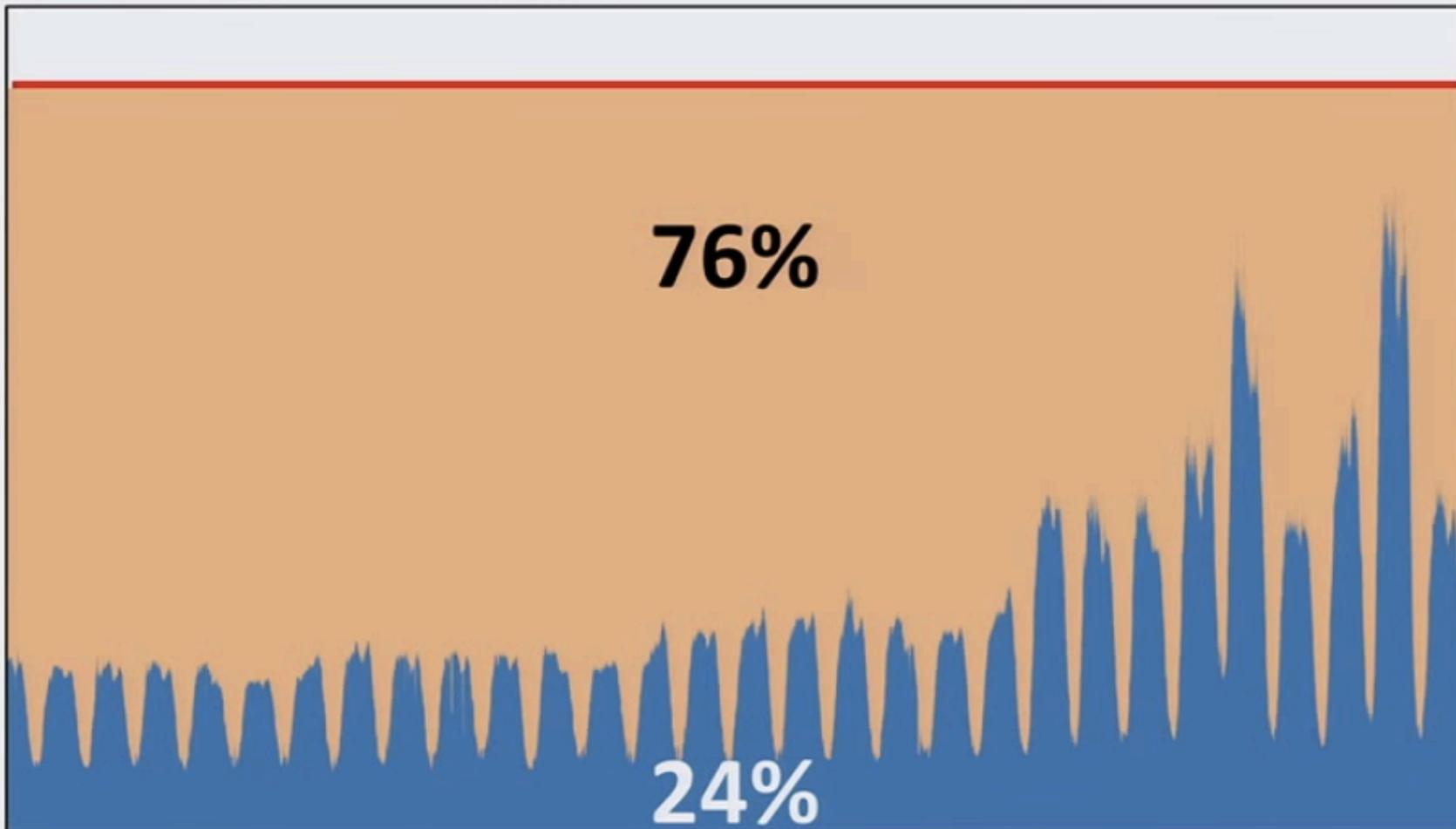
Case Study: Amazon Retail Website

November Traffic for amazon.com



Case Study: Amazon Retail Website

November Traffic for amazon.com



76%

24%

Motivation of Cloud Computing (I)

November Traffic for amazon.com



Capacity Planning = Spending Money

Capacity Optimization = Saving Money



Motivation of Cloud Computing (2)

How can we better utilize the available computing resource?

ember Traffic for amazon.com

76%

24%



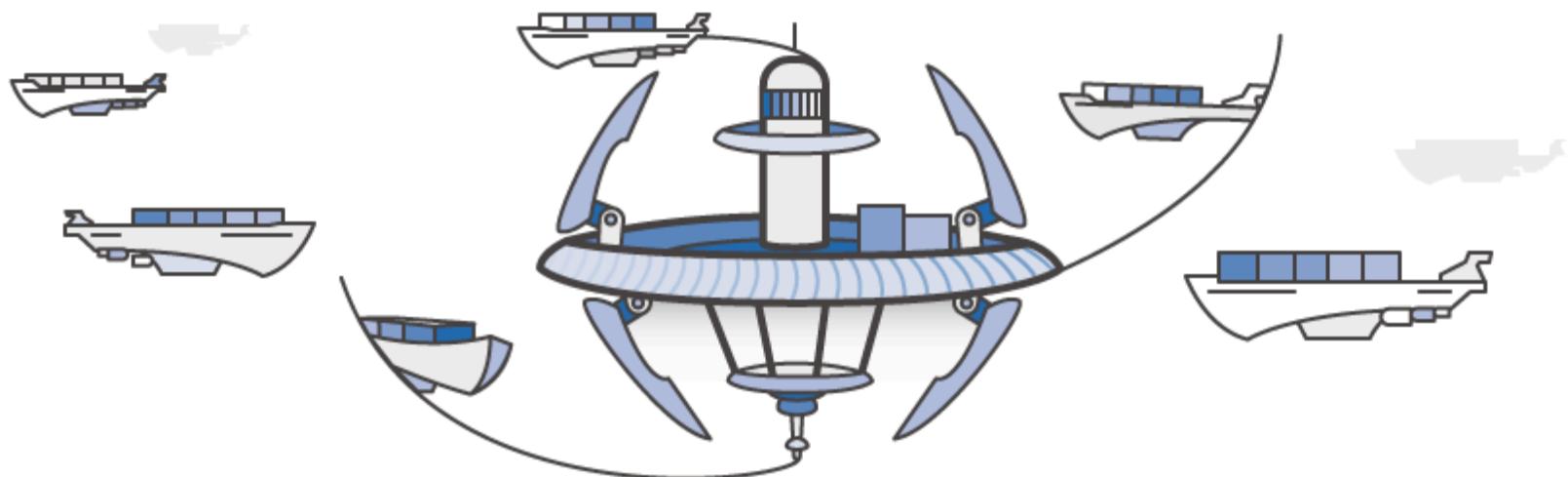
Cloud Computing

- ◆ Cloud computing shifts computing from local dedicated resources to distributed, virtual, elastic, multi-tenant resources
 - ◆ On-demand access to computing, storage, and software services
 - ◆ Based on a utility cost model

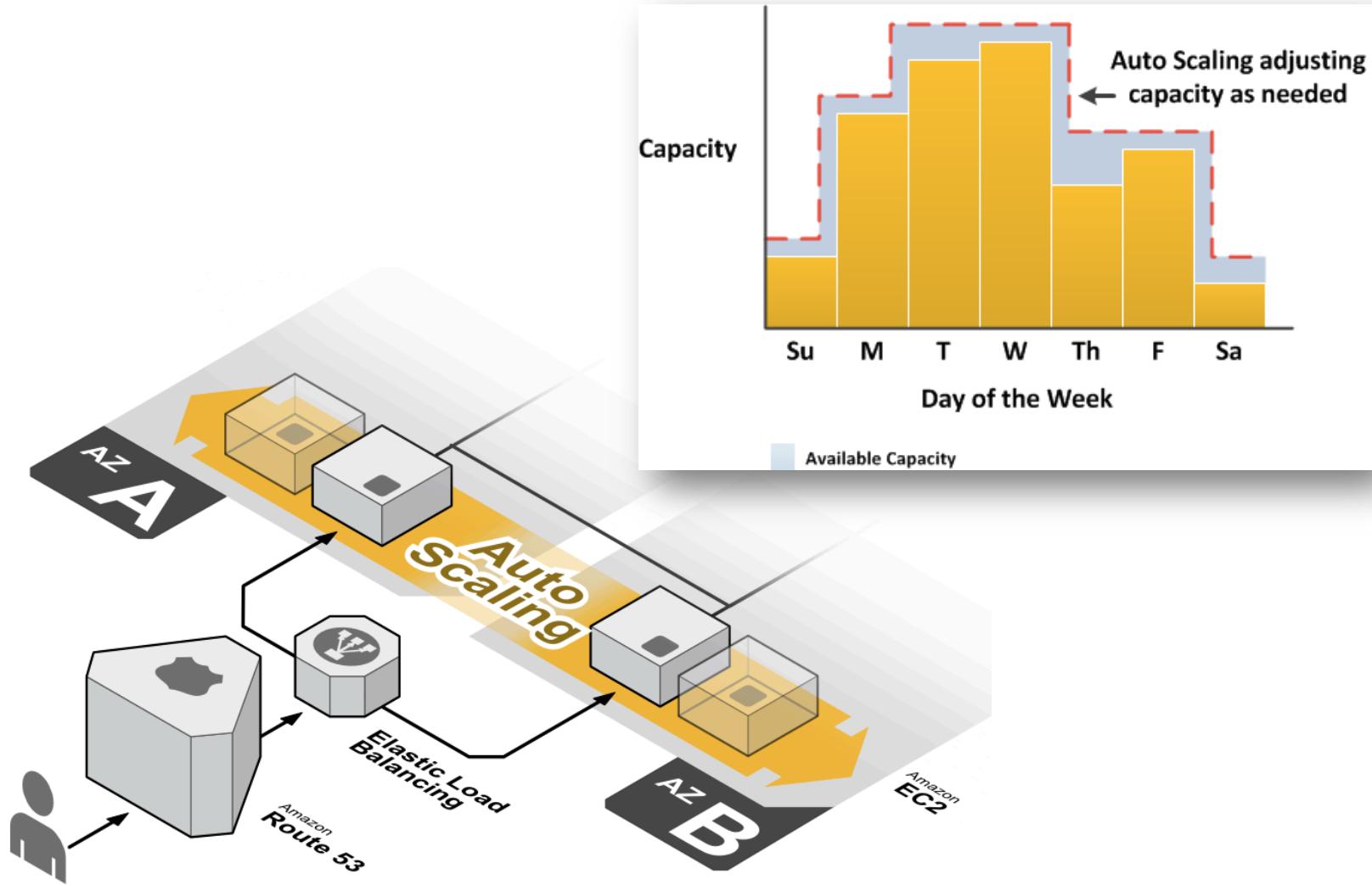


Amazon EC2 On-Demand Pricing

United States	Europe	
Standard On-Demand Instances	Linux/UNIX Usage	Windows Usage
Small (Default)	\$0.10 per hour	\$0.125 per hour
Large	\$0.40 per hour	\$0.50 per hour
Extra Large	\$0.80 per hour	\$1.00 per hour
High CPU On-Demand Instances	Linux/UNIX Usage	Windows Usage
Medium	\$0.20 per hour	\$0.30 per hour
Extra Large	\$0.80 per hour	\$1.20 per hour



Auto Scaling with Cloud Computing



Auto-Scaling Demo

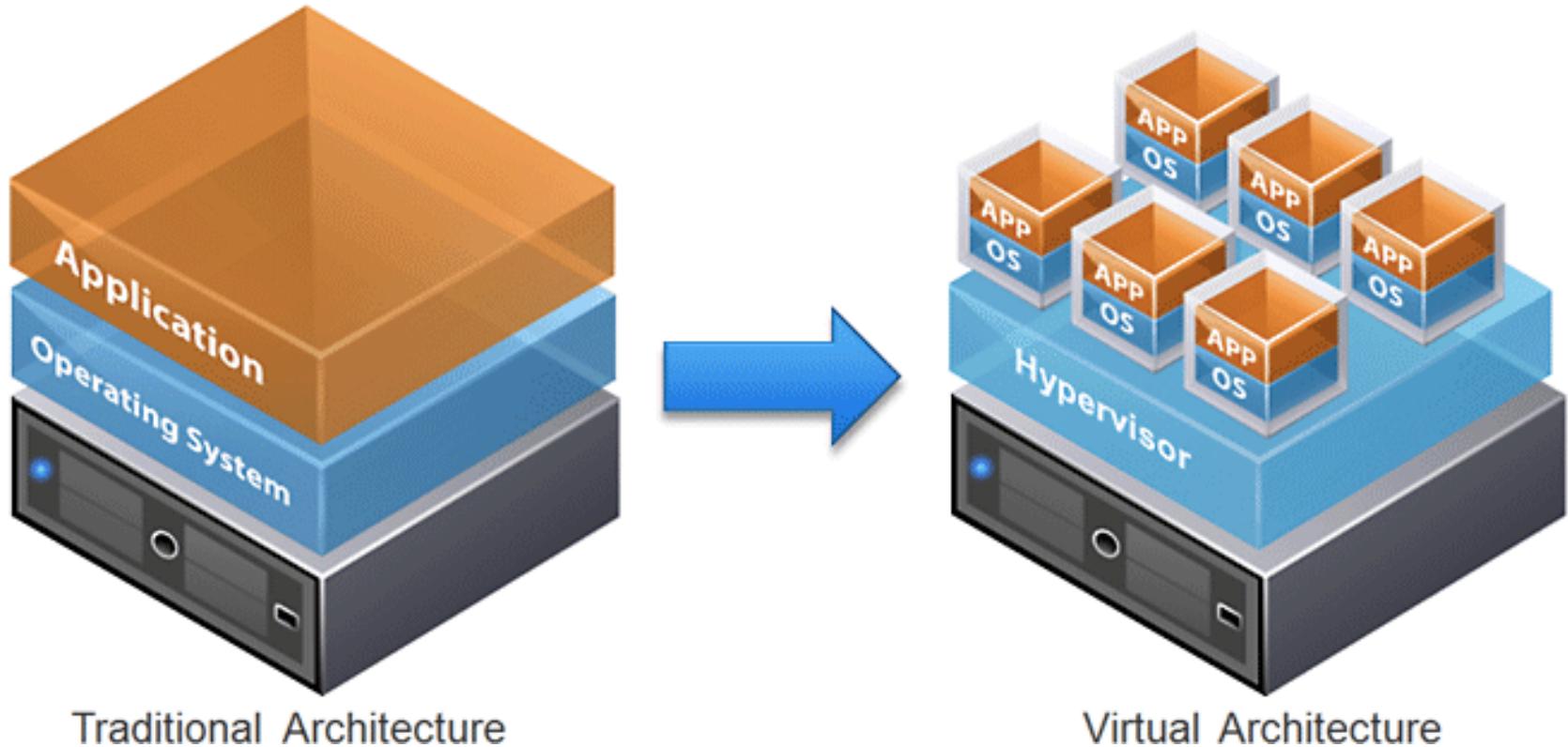


Amazon.com is Fully Served by EC2

- ◆ Reduced spending on server capacity
- ◆ Fleet scales dynamically in increments as small as a single host
- ◆ Traffic spikes can be handled with ease
- ◆ Cultural change



Virtualization

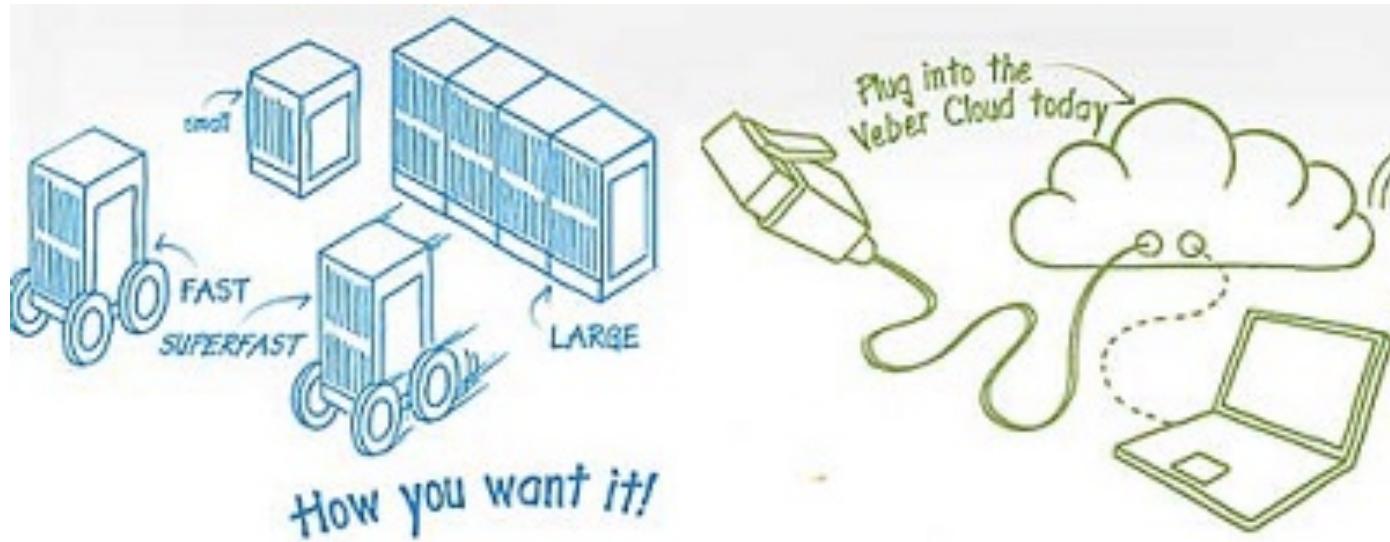


Flexible Options

Google vs. AWS On-Demand Pricing

Google Instance Type	CPU Cores	RAM	AWS Instance Type	CPU Cores	RAM	Google New On-Demand (per hour)	AWS On-Demand (per hour)	New Google Price vs. AWS
n1-standard-1	1	3.75	m3.medium	1	3.75	\$ 0.070	\$ 0.113	-38.05%
n1-standard-2	2	7.5	m3.large	2	7.5	\$ 0.140	\$ 0.225	-37.78%
n1-standard-4	4	15	m3.xlarge	4	15	\$ 0.280	\$ 0.450	-37.78%
n1-standard-8	8	30	m3.2xlarge	8	30	\$ 0.560	\$ 0.900	-37.78%
n1-higmem-2	2	13	m2.xlarge	2	17.1	\$ 0.164	\$ 0.410	-60.00%
n1-higmem-4	4	26	m2.2xlarge	4	34.2	\$ 0.328	\$ 0.820	-60.00%
n1-higmem-8	8	52	m2.4xlarge	8	68.4	\$ 0.656	\$ 1.640	-60.00%
n1-highcpu-2	2	1.8	c3.large	2	3.75	\$ 0.088	\$ 0.150	-41.33%
n1-highcpu-4	4	3.6	c3.xlarge	4	7.5	\$ 0.176	\$ 0.300	-41.33%
n1-highcpu-8	8	7.2	c3.2xlarge	8	15	\$ 0.352	\$ 0.600	-41.33%
n1-highcpu-16	16	14.4	c3.4xlarge	16	30	\$ 0.704	\$ 1.200	-41.33%

Rapid Resource Allocation



Dedicated Vs Cloud

High Availability

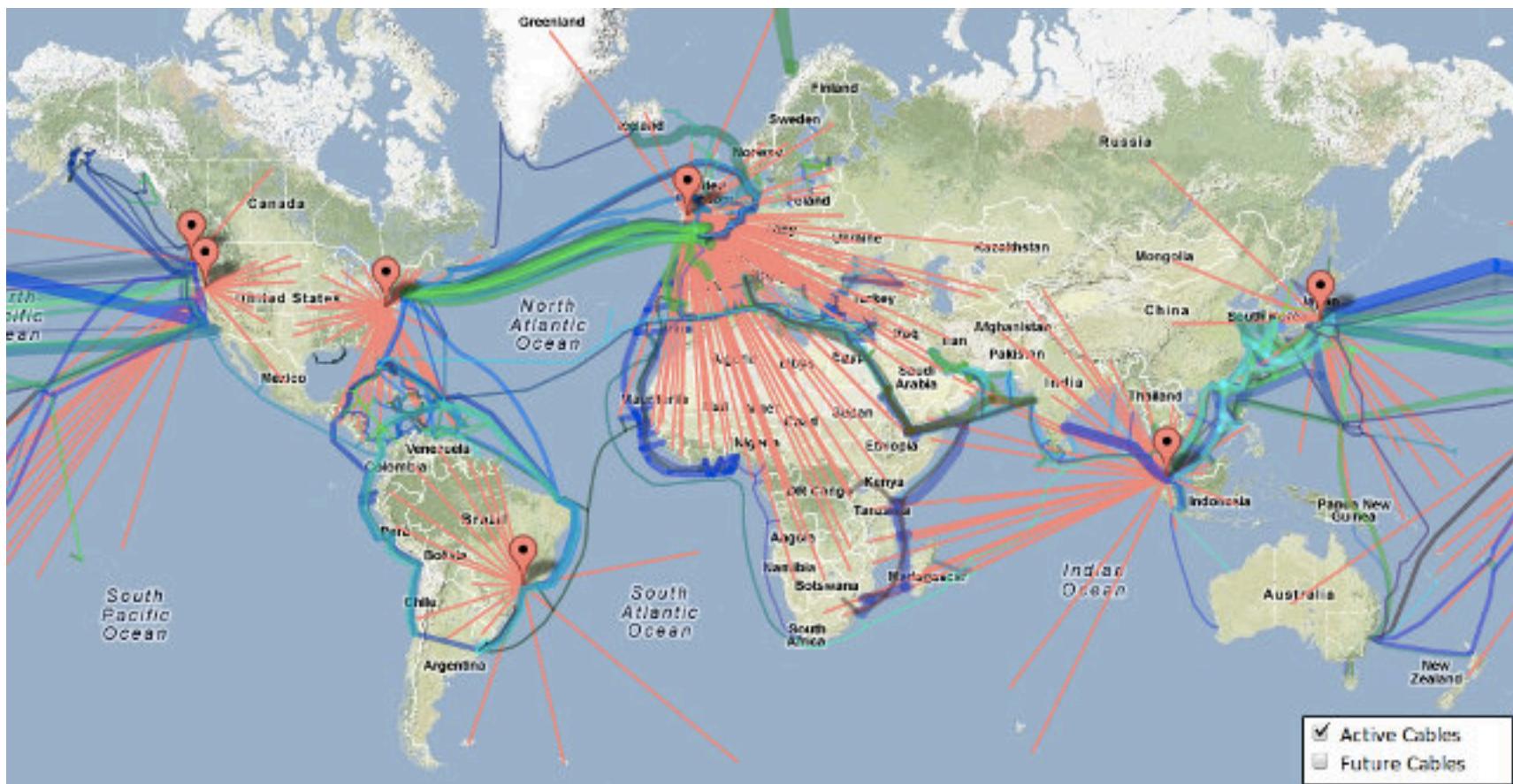


**TRADITIONAL
HOSTING**

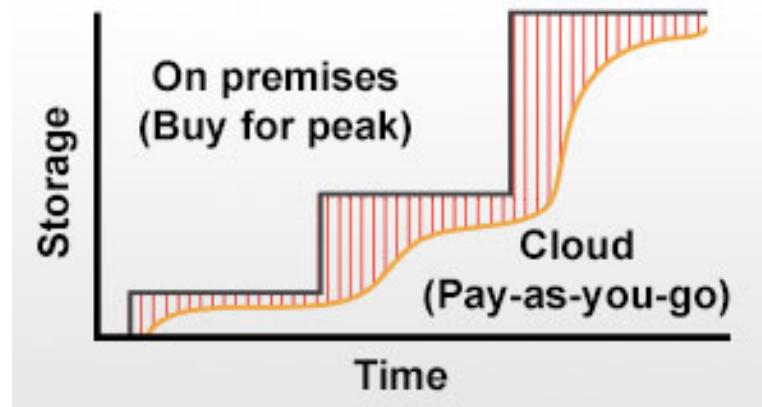
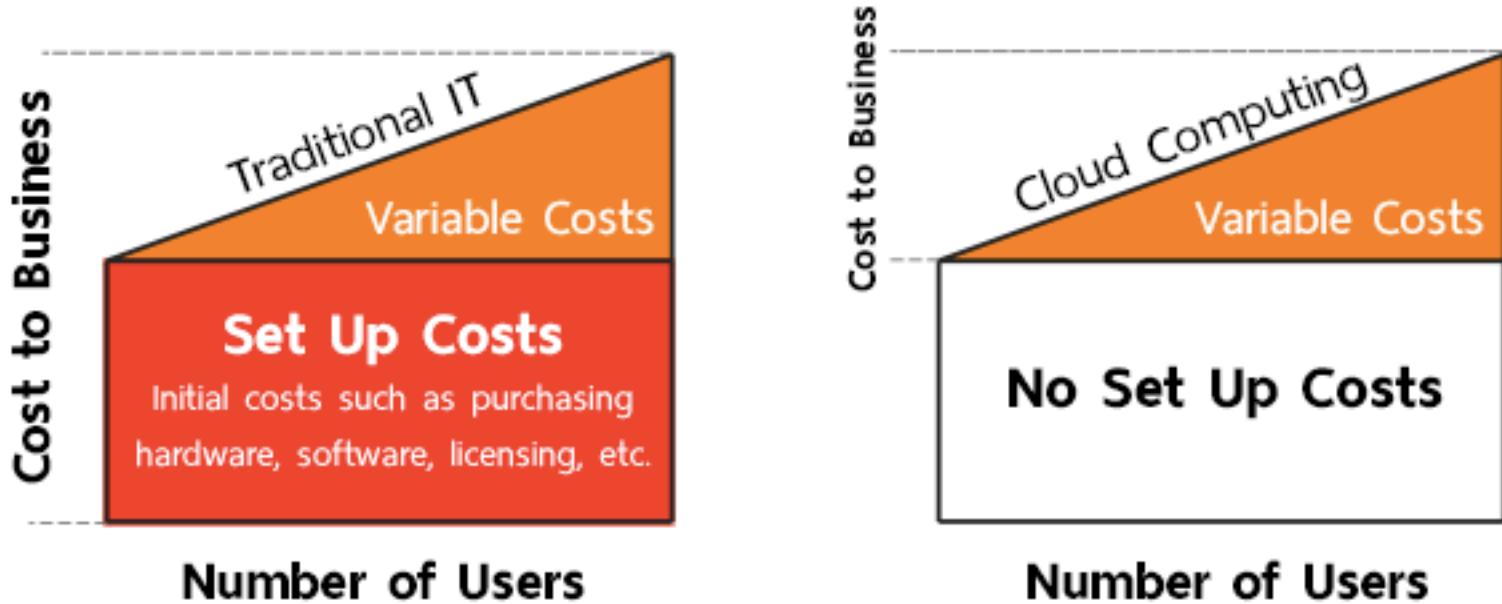


**CLOUD
HOSTING**

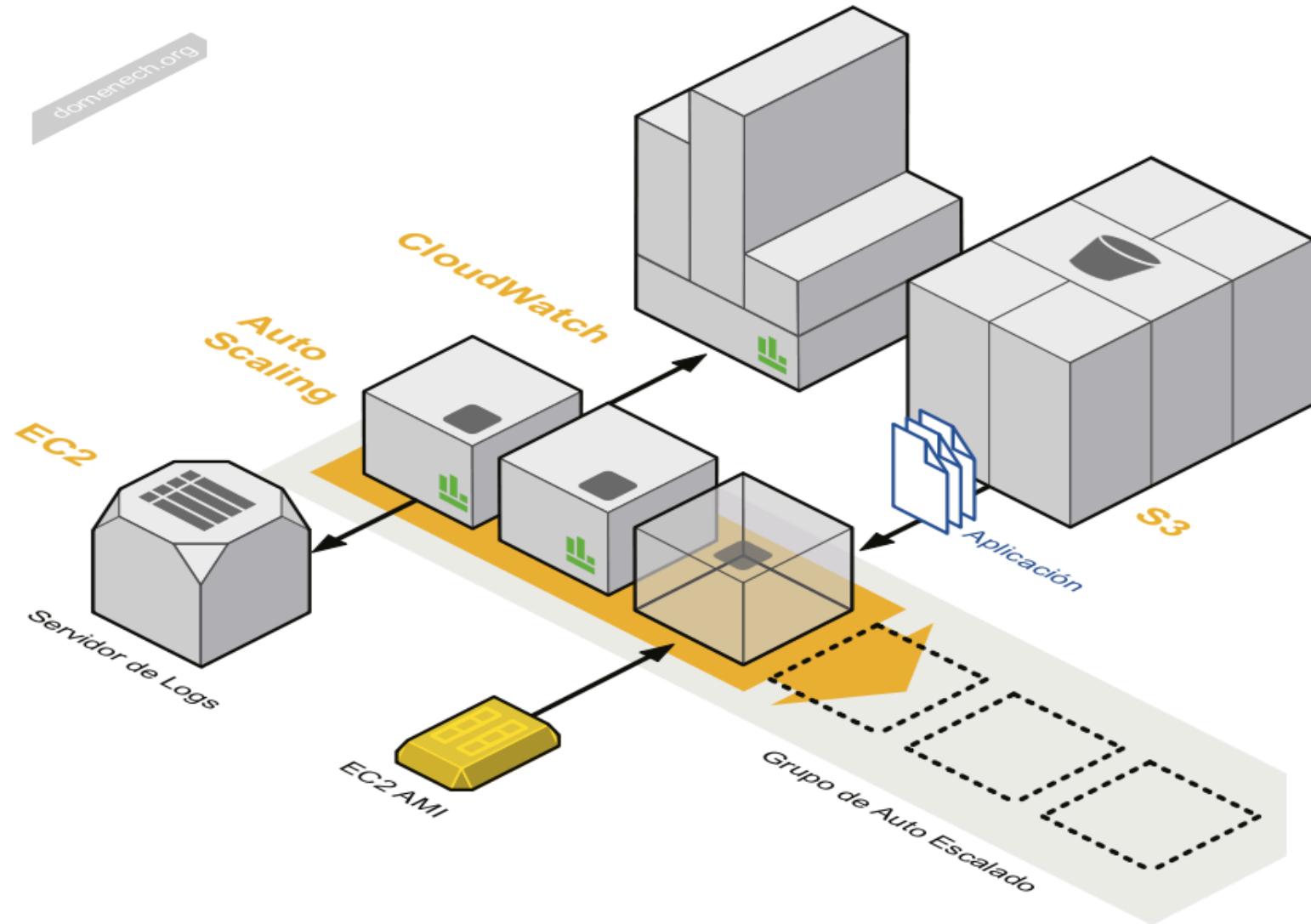
High Availability: AWS Data Centers



Cost-Effective



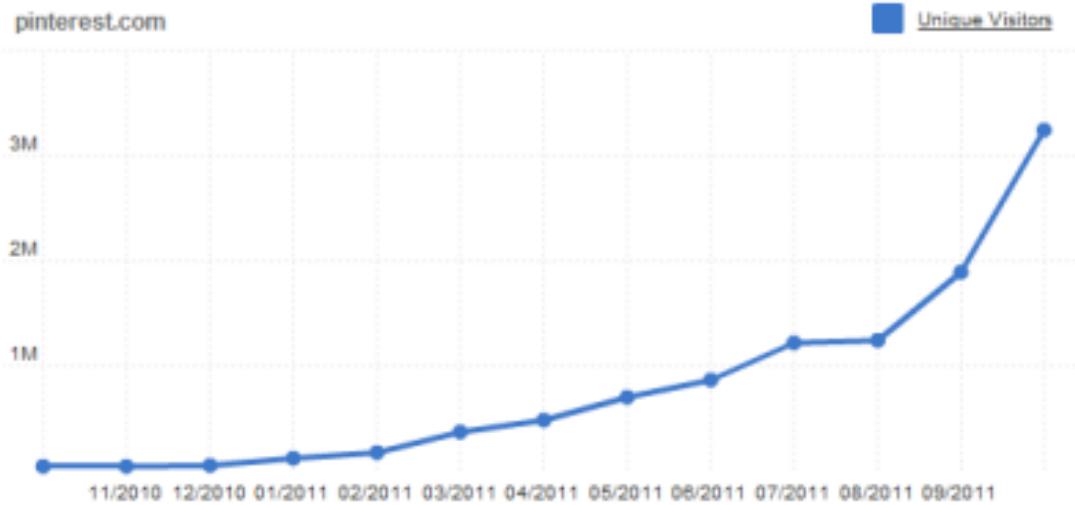
Scalability: Auto-Scaling



Focus on the Applications



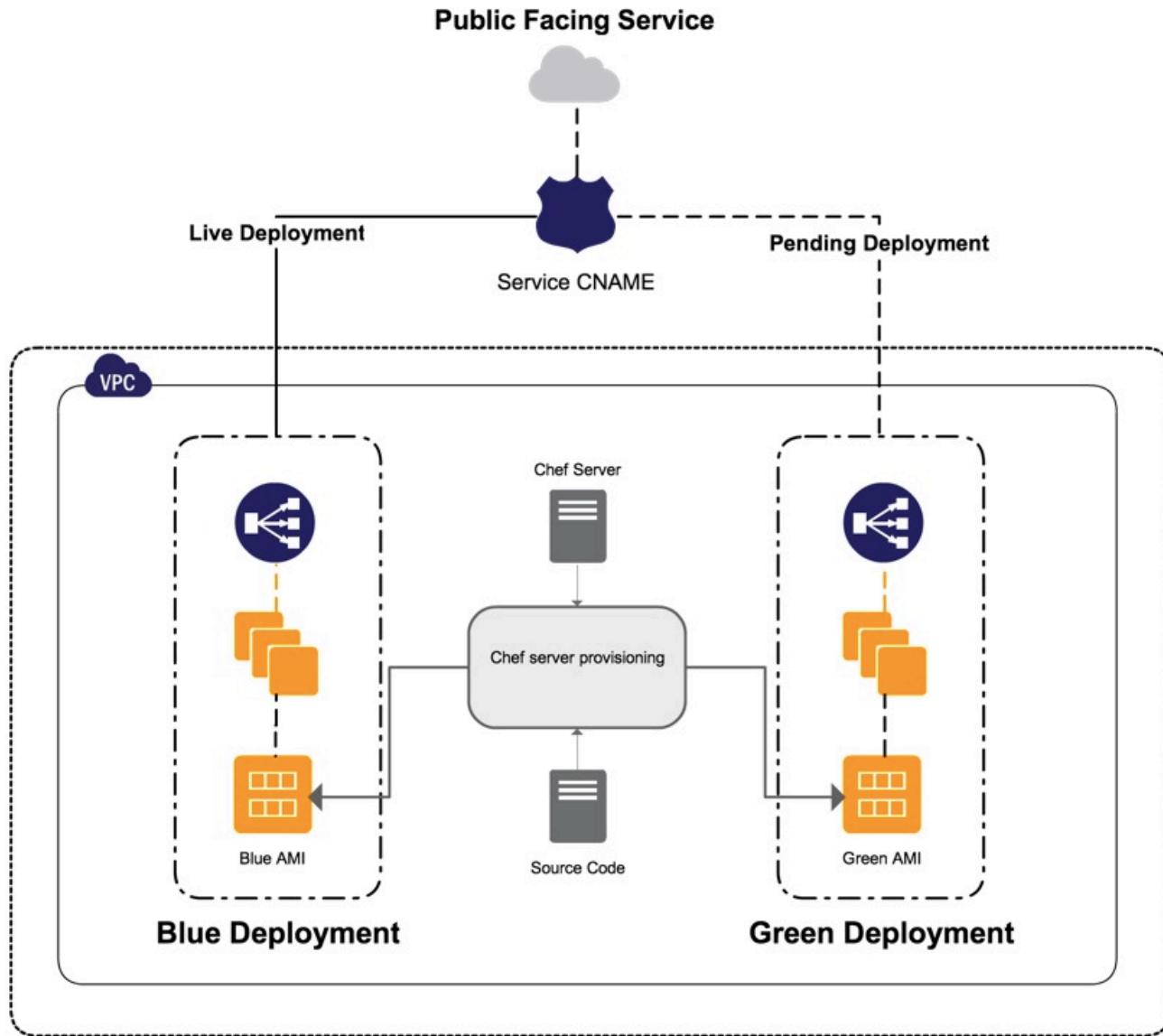
Startups Made Easy



Etsy

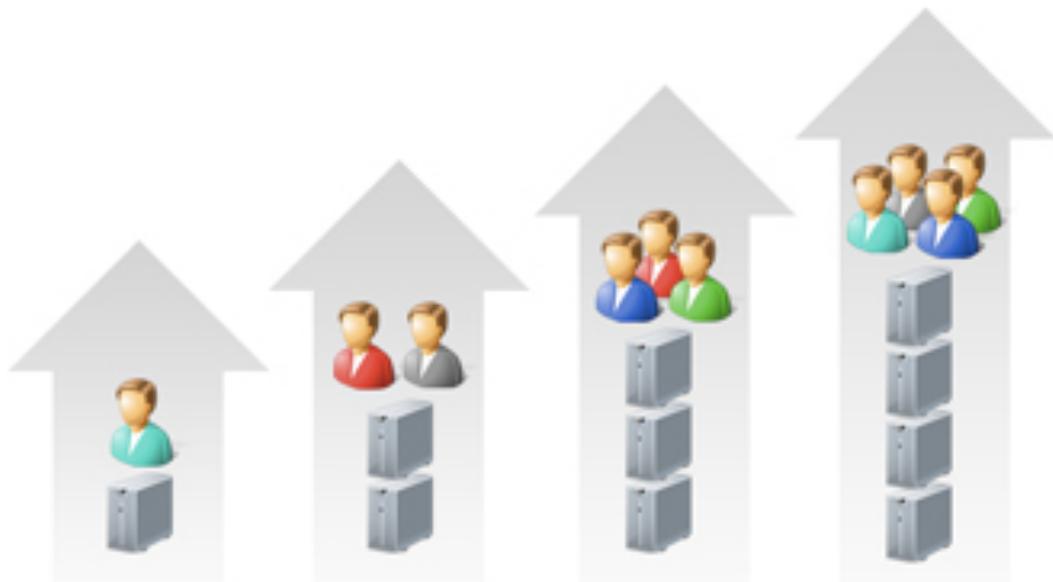


Deployment with Ease

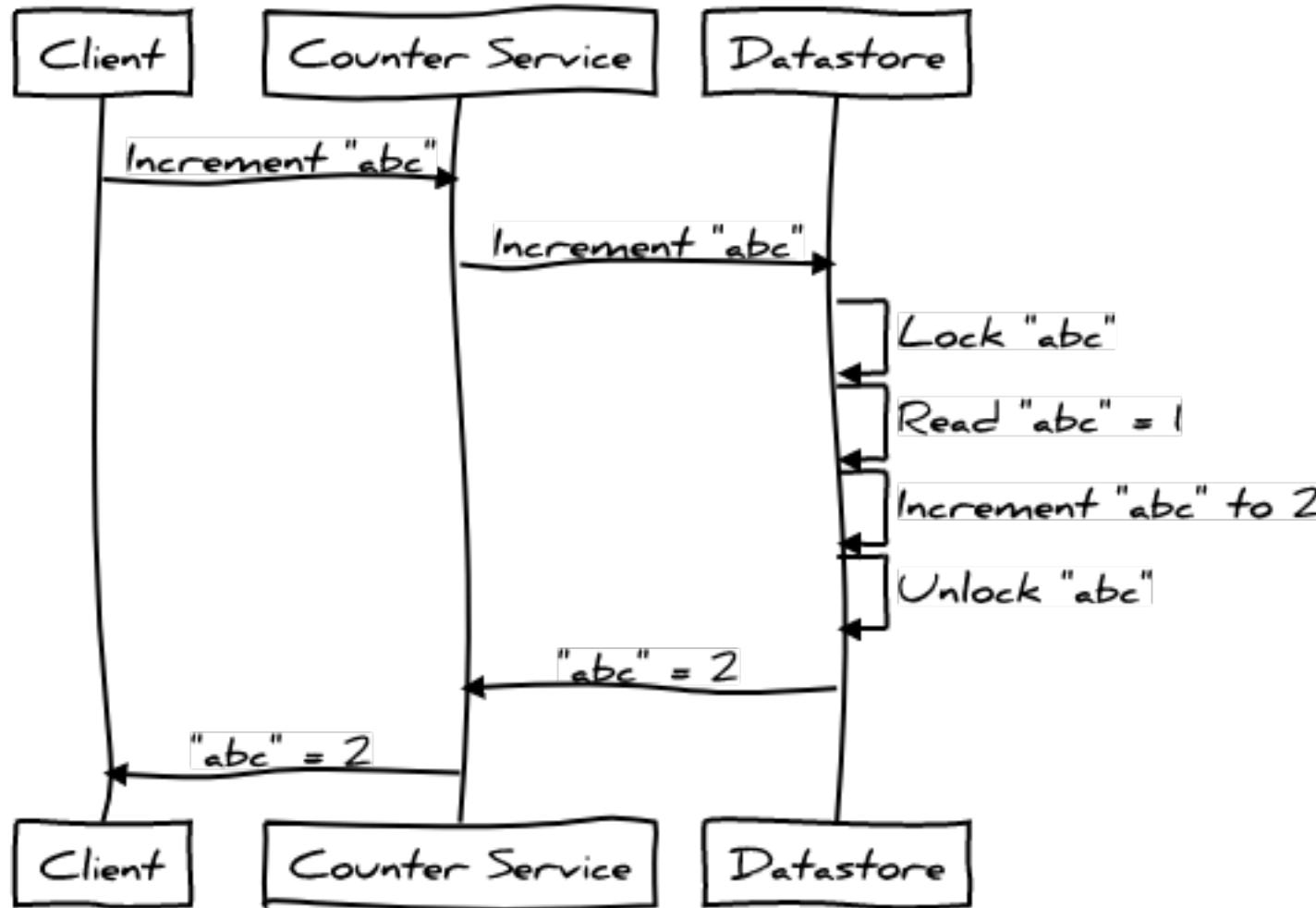


Always Keep Scalability in Mind

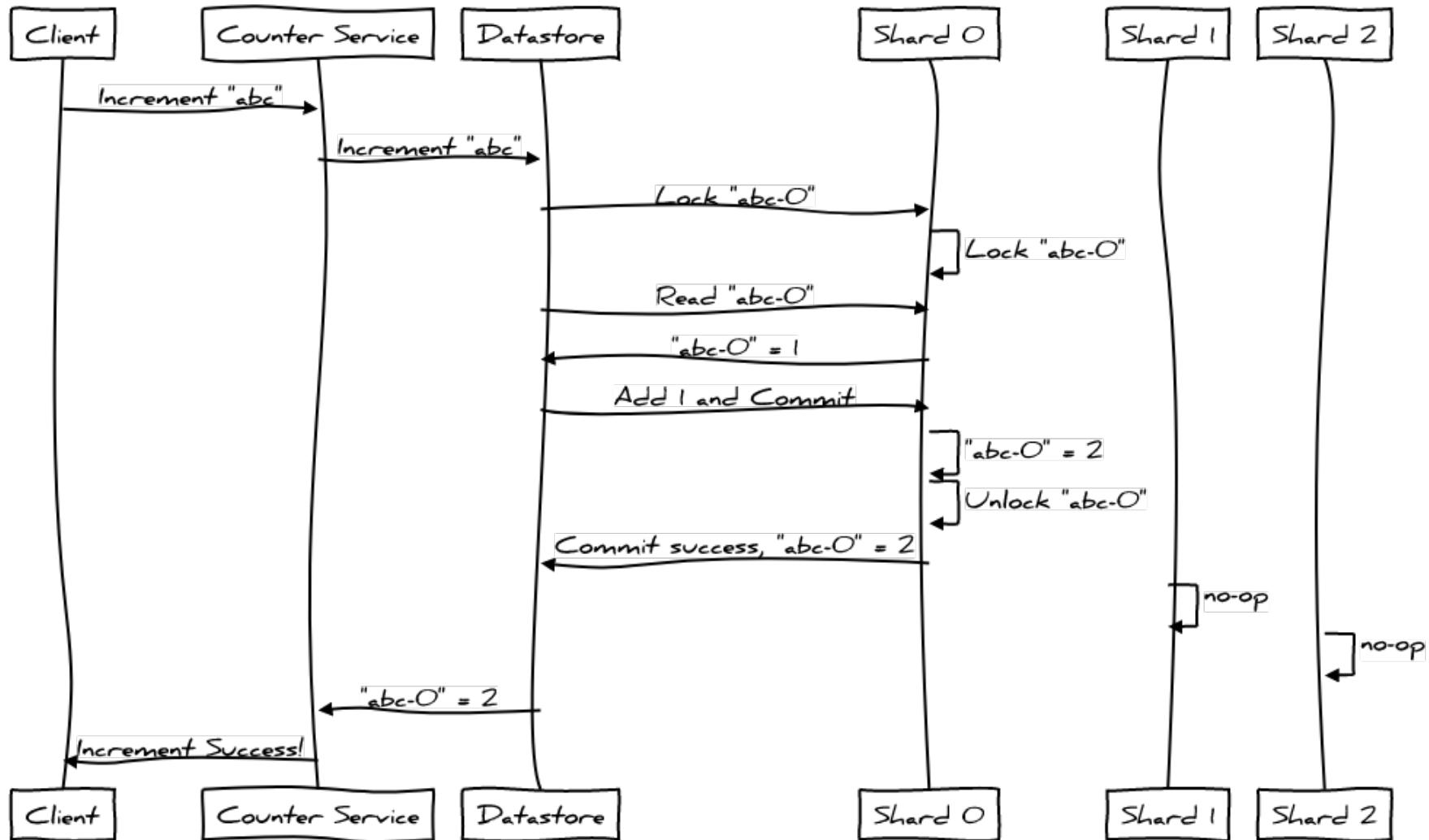
- ◆ Minimizing work
- ◆ Paging through large datasets
- ◆ Avoiding datastore contention
- ◆ Sharding counters
- ◆ Effective memcache



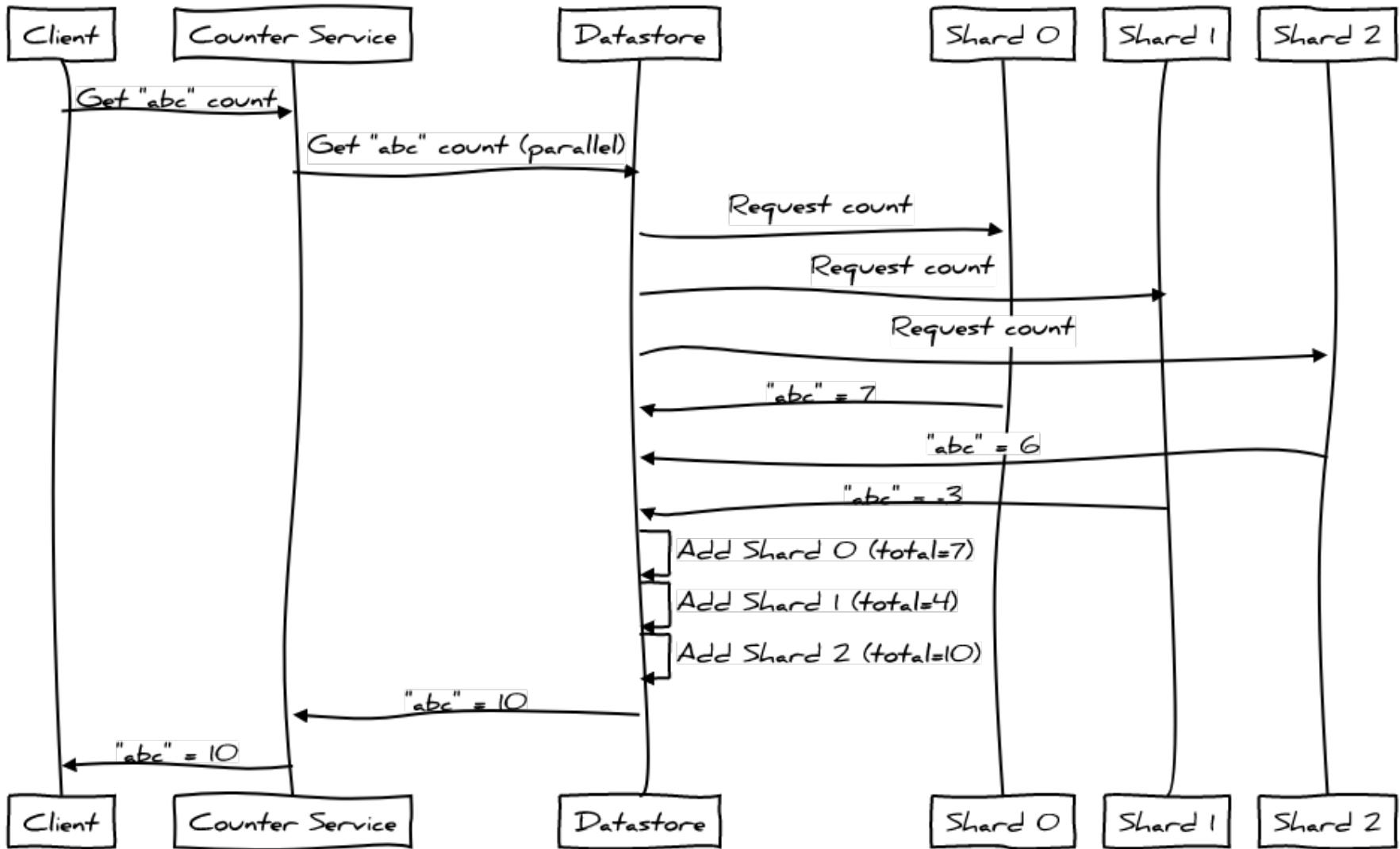
Sharding Counters



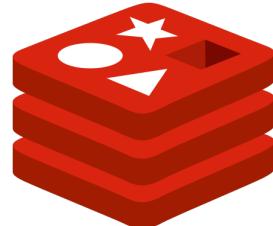
Sharding Counters



Sharding Counters



Effective Memcache



redis