Blog | Automation | July 8, 2021

Horizontal vs. Vertical Scaling in the Cloud

Among the many reasons to make the move to the cloud, scalability is one of the most compelling. What is scalability in cloud computing? Scalability is the ability to easily add or subtract compute or storage resources. Horizontal and vertical scaling in cloud computing makes it easier for enterprises to provision the right number and size of resources without the overhead of running a data center.

Cloud vs. Data Center

The switch to cloud has improved the computing power for organizations that used to run on-premises servers. The leading cloud providers — Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform — offer flexibility for organizations that may need to add or reduce resources at a moment's notice.

With on-prem data centers, however, the process of scaling resources was and remains incredibly costly, slow, and difficult to manage. In a data center, scaling up means buying new server hardware and disk arrays. This ends up being a costly and time-consuming process for many organizations. Even after leaders approve new equipment purchases, it can take months after the order is placed before the equipment arrives. That doesn't even include the time and resources IT personnel spend installing, maintaining, updating, and repairing the servers. In short, data centers are costly and time-consuming for businesses to run today.

Why Pay Attention to Scalability?

Even a thriving business might encounter times when there is more or less demand.

In a data center world, reducing capacity was almost never practical, so companies were left provisioning enough resources to cover their expected peak demand. In other words, an eCommerce site would need enough computing resources to handle Black Friday levels of traffic every single day. Utilization rates were very low, especially because most companies would provision resources based on expected peak demand, plus some.

The alternative is to provision just enough resources for daily use and not for peak traffic. Yet the consequences of not having enough compute or storage resources are dire. First come performance issues, then users start getting error messages and getting locked out of the application. In a business setting, that equals lost revenue. Conversely, resources are not free. Over-provisioning can lead to ballooning IT costs.

Horizontal and Vertical Scaling Strategies

The cloud has dramatically simplified these scaling problems by making it easier to scale up or down and out or in. Primarily, there are two ways to scale in the cloud: horizontally or vertically.

When you scale horizontally, you are scaling out or in, which refers to the number of provisioned resources. When you scale vertically, it's often called scaling up or down, which refers to the power and capacity of an individual resource.

What are the differences between horizontal and vertical scaling in the cloud?

Horizontal scaling refers to provisioning additional servers to meet your needs, often splitting workloads between servers to limit the number of requests any individual server is getting. Horizontal scaling in cloud computing means adding additional instances instead of moving to a larger instance size.

Vertical scaling refers to adding more or faster CPUs, memory, or I/O resources to an existing server, or replacing one server with a more powerful server. In a data center, administrators traditionally achieved vertical scaling by purchasing a new, different instance sizes, so vertical scaling in cloud computing is possible for everything from EC2 instances to RDS databases.

Horizontal vs. Vertical Scaling Pros and Cons

Pros and cons of horizontal scaling:

Pros: Horizontal scaling is much easier to accomplish without downtime. Horizontal scaling is also easier than vertical scaling to manage automatically. Limiting the number of requests any instance gets at one time is good for performance, no matter how large the instance. Provisioning additional instances also means having greater redundancy in the rare event of an outage.

Cons: Depending on the number of instances you need, your costs may be higher. Additionally, without a load balancer in place, your machines run the risk of being over-utilized, which could lead to an outage. However, with public cloud platforms, you can pay attention to discounts for Reserved Instances (RIs) if you're able to predict when you require more compute power. Following cloud cost management best practices can help you efficiently scale in or out.

Pros and cons of vertical scaling:

Pros: In the cloud, vertical scaling means changing the sizes of cloud resources, rather than purchasing more, to match them to the workload. This process is known as right sizing. For example, right sizing in AWS can refer to the CPU, memory, storage, and networking capacity of instances and storage classes. Right sizing is one of the most effective ways to control cloud costs. When done correctly, right sizing can help lower costs of vertically scaled resources.

There's also downtime to consider. Even in a cloud environment, scaling vertically usually requires making an application unavailable for some amount of time. Therefore, environments or applications that can't have downtime would typically benefit more from horizontal scalability by provisioning additional resources instead of increasing capacity for existing resources.

Which Is Better: Horizontal or Vertical Scaling?

The decision to scale horizontally or vertically in the cloud depends upon the requirements of your data. Remember that scaling continues to be a challenge, even in cloud environments. All parts of your application need to scale, from the compute resources to database and storage resources. Neglecting any pieces of the scaling puzzle can lead to unplanned downtime or worse. The best solution might be a combination of vertical scaling in order to find the ideal capacity of each instance and then horizontal scaling to handle spikes in demand, while ensuring uptime.

Types of Cloud Scalability: Manual vs. Scheduled vs. Automatic Scaling

What also matters is *how* you scale. Three basic ways to scale in a cloud environment include manual scaling, scheduled scaling, and automatic scaling.

Manual Scaling

Manual scaling is just as it sounds. It requires an engineer to manage scaling up and out or down and in. In the cloud, both vertical and horizontal scaling can be accomplished with the push of a button, so the actual scaling isn't terribly difficult when compared to managing a data center.

However, because it requires a team member's attention, manual scaling cannot take into account all the minute-by-minute fluctuations in demand seen by a

Scheduled Scaling

Scheduled scaling solves some of the problems with manual scaling. This makes it easier to tailor your provisioning to your actual usage without requiring a team member to make the changes manually every day.

If you know when peak activity occurs, you can schedule scaling based on your usual demand curve. For example, you can scale out to ten instances from 5 p.m. to 10 p.m., then back into two instances from 10 p.m. to 7 a.m., and then back out to five instances until 5 p.m. Look for a cloud management platform with Heat Maps that can visually identify such peaks and valleys of usage.

Automatic Scaling

Automatic scaling (also known as Auto Scaling) is when your compute, database, and storage resources scale automatically based on predefined rules. For example, when metrics like vCPU, memory, and network utilization rates go above or below a certain threshold, you can scale out or in.

Auto scaling makes it possible to ensure your application is always available — and always has enough resources provisioned to prevent performance problems or outages — without paying for far more resources than you are actually using.

Organizations can follow a number of best practices around scaling instances. Namely, they need to know when to use horizontal and vertical scaling and where they can automate their cloud scalability.

Horizontal and vertical scaling in AWS, for example, means paying attention to the *number* of EC2 instances provisioned as well as the *sizes* of those instances.

To scale horizontally in AWS, begin only with the resources you need and design your architecture to automatically respond to changes in demand. Amazon EC2 Auto Scaling can add or remove EC2 instances in response to changing demand. Two subsets of Auto Scaling in AWS include dynamic scaling, which can be configured based on policies you set, or predictive scaling, which can schedule the right number of instances based on predicted demand. Having a tool that can terminate instances automatically when they're not in use can also help organizations save money.

When it comes to vertical scalability, organizations should pay attention to their right sizing strategy. Right sizing instances, or choosing the correct instance sizes based on your actual application utilization, is one of the easiest ways to reduce cloud costs without affecting performance in any way. There are also some cost management strategies, like Reserved Instance purchases, that take away some of the ability to scale in or down, because you're committing to using certain amounts and types of resources for one to three years. When you're looking for ways to reduce costs, it's important to understand your current usage patterns and utilization rates to make the best decisions about how to strike a balance between total scaling flexibility and cost management strategies like RI purchases.

Managing Your Cloud Scaling Strategy

Managing scaling correctly is the key to ensuring you always have enough resources without over-provisioning and wasting your cloud budget. The ability to automatically scale is one of the most attractive parts of moving to a cloud environment. When used correctly, auto scaling can ensure you're only paying for the resources you actually use.

As you figure out the best strategy for managing scaling, it's important to understand your historical usage patterns, how RI purchases affect scaling, and whether manual, scheduled, or automatic scaling is best for your use case.

Paying attention to scalability is just one way to make the most of your cloud investment. You should also optimize your cloud technology to bring down high infrastructure costs, increase the speed in which you deploy applications, and improve observability.

Watch our on-demand webinar **Diving in Cloud-First: How To Modernize Your Cloud Architecture**. In this webinar, our own Chief Architect Bob Pease will show you:

- Ways to use your cloud-native architecture to give you more flexibility and efficiency
- How CloudCheckr modernized its own all-in-one cloud management platform
- Everyday tools you can use to analyze and monitor your modern technology

Watch the Webinar On Demand

Related Resources

Article

Spot and AWS Deliver an Integrated Deployment Experience with CloudCheckr Built-in Solution

Case Study

How inQdo Cloud Eliminated Billing Complexity and Uncovered New Revenue Streams

Article

Map Your Cloud Journey: 50 Essential Cloud KPIs to Guide the Way

Article

An Insider's Look at AWS re:Invent 2022

Products

CloudCheckr CMx

CMx Federal

What's New

Documentation

Training & Certification

Company

About

Partners

News

Careers

Support