Vertical vs. Horizontal Scaling: Which One to Choose

Find out the pros and cons of horizontal and vertical scaling, and choose the best one for your business needs.

MW Team September 27, 2021 Knowledge

Scalability is gaining traction and dynamism in the tech space.

It's a well-known concept among information technology (IT) and data professionals who often deal with changing application and system processing requirements.

Scalability is measured by the ability of applications or systems to handle simultaneous requests. If an application or system can't cope with increasing demand, it reaches its limit. Therefore, when system resources are exhausted and no free resources are available for further computation, scaling the application or system is the next logical step.

Table of Contents

What is scaling?

Let's assume your application can handle a maximum of 'x' requests simultaneously. As soon as this number exceeds x (say, x+1), critical hardware resources are exhausted, and the application cannot process further requests. This is when you need scaling.

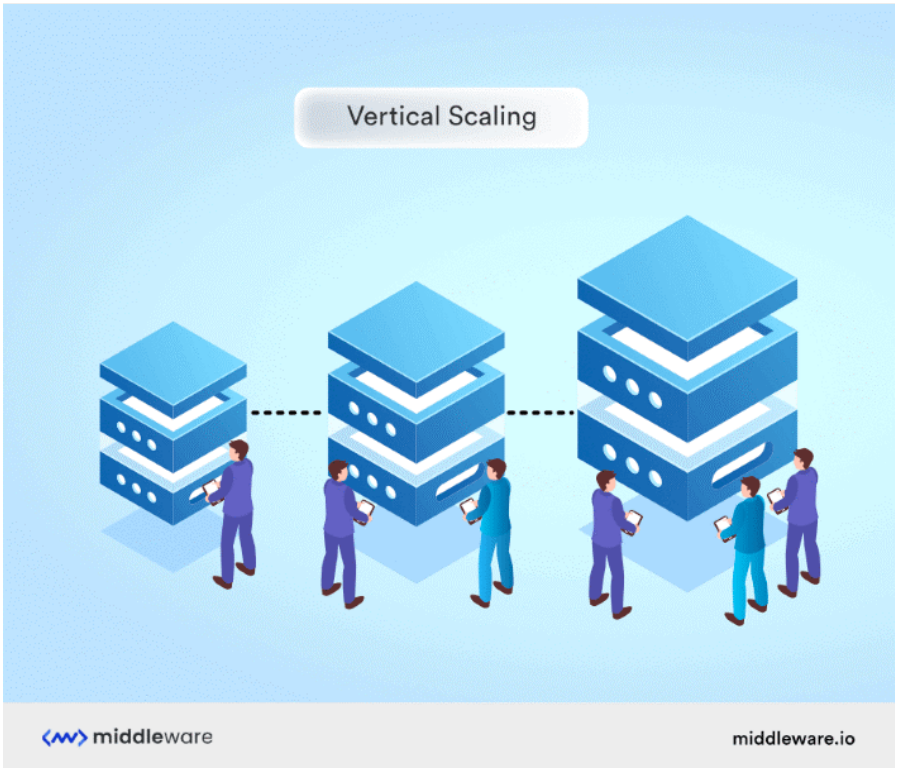
Scaling refers to the adjustments made to system hardware resources to help your business grow by enabling your applications and systems to handle increased traffic and requests and your business to grow with improved scalability. This could be in the form of:

- Adjustments to network bandwidth
- Upgrades to CPU capacity and physical memory
- Basic hard drive alterations

Alternatively, a business can simply replace the existing system infrastructure with bigger, better, and faster options. Based on your approach, your scaling efforts can largely be categorized into vertical (scaling up) or horizontal (scaling out). Let’s further understand the difference and their respective uses.

What is vertical scaling?

Vertical scaling or scaling up is the process of upgrading or adding resources to the existing system infrastructure on demand. Unlike scaling out, vertical scaling doesn’t involve adding more machines to the existing resource pool but rather meeting the increased demands by upgrading the existing system’s resources.



Vertical scaling can help add more processing power by upgrading the CPUs, increasing system memory, scaling storage or network speed, and so on. In some cases, it can also lead to a complete overhaul of the existing infrastructure, replacing the old server with a new and upgraded one capable of handling the increased demands.

What is horizontal scaling?

Horizontal scaling or scaling out refers to integrating additional server nodes or machines into your existing system infrastructure. This involves adding more nodes to an existing resource pool to distribute the workload across multiple servers.

Any business looking to scale out will add one or more new server nodes every time the existing servers can't handle the workload.

When you scale out, you're essentially delegating system functions to several units that need to work together to achieve the desired results. Unlike vertical scaling, you don't attempt to meet requirements by upgrading or replacing existing infrastructure.

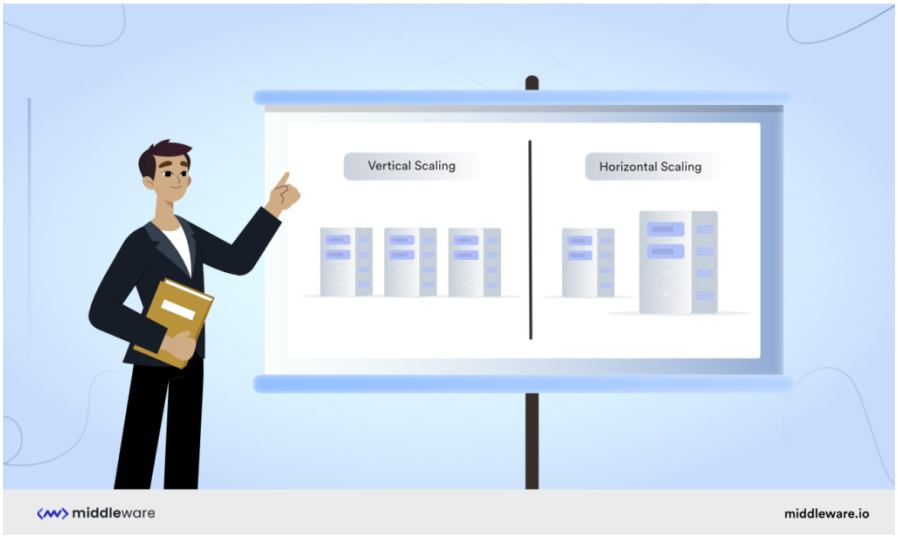
Vertical vs. horizontal scaling

There are many key differences between vertical and horizontal scaling to consider when deciding the best scaling model for a particular system or application. Let's discuss these in detail.

	Vertical Scaling	**Horizontal Scaling**
Databases	Vertical scaling involves multi-core system upgrade, and the information remains on a single node.	Horizontal scaling involves splitting databases and partitioning data, allowing information to exist on multiple nodes.
Downtime	Vertical scaling involves upgrading one machine or system on which you depend entirely. Since every machine has a fixed upper limit, scaling up might involve significant downtime.	Since horizontal scaling adds the server resources to the existing server pool, you're not dependent on one machine or system. You can scale out with lesser downtime.
Concurrency	The workload in vertically scaled systems is usually handled with the help of multi-core machines through in-process message passing and multi-threading of tasks.	The workload in horizontally scaled server pools is distributed over several nodes and servers. This requires the servers to work constantly and concurrently.
Message passing	Vertical scaling includes a shared address space, where all computing resources exist within a system or server. This makes data sharing and message passing easier and less complicated.	Horizontal scaling involves distributed computing that lacks a shared address space. This means that the machines need to communicate with one another and exchange data within the framework. As this involves copies of information, it can be costly.
Examples	Amazon RDS and MySQL	Google Cloud Spanner and Cassandra

How to choose between scaling up and scaling out

Once you have a sound understanding of the two scaling models, the next step is to choose the one that best suits your system. While this decision is largely influenced by the nature of the application and expected increase in server load, as well as other similar variables, there are three primary factors to keep in mind.



1. Performance

Horizontal scaling or scaling out allows you to combine the computing power of multiple servers and machines into one resource pool. This will no doubt offer a significant and drastic improvement in performance as well as the capacity of your application to handle more requests. At the same time, if your scaling requirements can be fulfilled by a single machine, scaling up or vertical scaling might prove to be a simpler and more efficient choice.

2. Flexibility

Flexibility in your scalability needs is important if you want to make your costs and performance more efficient. With vertical scaling, you're always bound by the minimum price predetermined by the machine your application is running on, which makes flexibility in terms of cost and performance optimization almost non-existent. By scaling out, you can access servers to pay for what you use. This allows you a lot more flexibility to control costs based on your fluctuating server bandwidth needs.

3. Cost

With horizontal scaling or scaling out, you need to add more servers to work with your existing ones to meet your performance needs. This is usually an expensive proposition. With more and more multi-core machines entering the market at lower prices, taking a vertical scaling approach and purchasing a single machine capable of fulfilling your requirements can prove to be the more cost-effective option.

Switching between the two scaling models

Realistically, it's unlikely to settle on one scaling model and stick to it perpetually. A growing business' needs can sometimes change quickly and unexpectedly. Therefore, maintaining a level of flexibility that allows you to switch between the two models – as and when required – would be the smartest option.

The industry is slowly shifting to a predominant horizontal scaling environment because of its flexibility for optimum resource utilization and being more cloud or SaaS friendly. However, there are two factors to keep in mind to help you seamlessly transition between the two scaling models.

1. Design your system architecture as a decoupled set of services from the start. This allows for easier code shifting during scaling.
2. Partition your data so that parallel server units do not exchange any information. Although this requires extra effort in server management, it offers flexibility and dynamism that'll pay off in the long run.

Choose the right scaling method to succeed

There is almost no company that goes into business without the goal of growing and increasing revenue. If a business successfully addresses a need, the chances of rapid growth are extremely high. More often than not, some form of regular scaling is inevitable.

This is why it's necessary for every business and organization to clearly understand the two scaling models. They need to zero in on the most suitable method depending on the functions and services offered to customers.

Both vertical scaling and horizontal scaling have their pros and cons, and there is no straight-jacket formula to determine which one is best for your

business. The ideal approach would therefore be to do a cost-benefit analysis of both models and then choose the one that best satisfies your business needs now and in the near future.

Leave a comment

Your email address will not be published. Required fields are marked *

Enter Name *

Enter Comment *

Post Comment

1 comment

CC December 6, 2021 at 4:59 pm

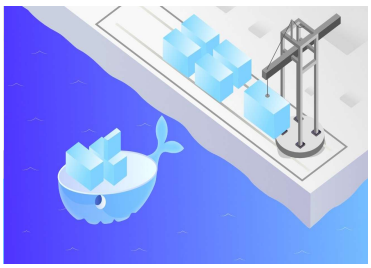
Great article! Really covers a lot in so little time. Thank you for sharing!

Reply

Knowledge November 2, 2021

6 Advantages of Cloud Infrastructure for Your

MW Team



Tutorials February 16, 2022

Docker Cleanup: How to Remove Images,

MW Team

Knowledge November 15, 2021

What Is Multi-cloud? How to Create a Multi-

MW Team