Pavle Davitković • 2nd
Software Engineer @TraceOne
36m •

Follow

What is vertical scaling?

And when should you consider it?

Let's say your application is performing heavy calculations, and at some point, its performance starts to degrade.

What would you do? Logically, you would increase the CPU power.

But what if that application begins to grow, or you have multiple applications running on the same server? You might increase SSD storage.

This is vertical scaling, also known as scaling up: the process of adding additional power to an existing server.

Some benefits you can expect include:

- Cost-effective: You are upgrading a single server with extra components that are cheaper than a whole new server

- Easy scaling: Performance improvements are visible immediately

- Easy maintenance: Since there's only one machine, maintenance requires fewer resources and less time

However, there is always another side to the coin.

Some disadvantages you may encounter are:

Single point of failure: If the server goes down, everything goes down

Downtime: It’s not possible to improve infrastructure while the machine is up

Limited scalability: Every machine has its limits, and you should be aware of those

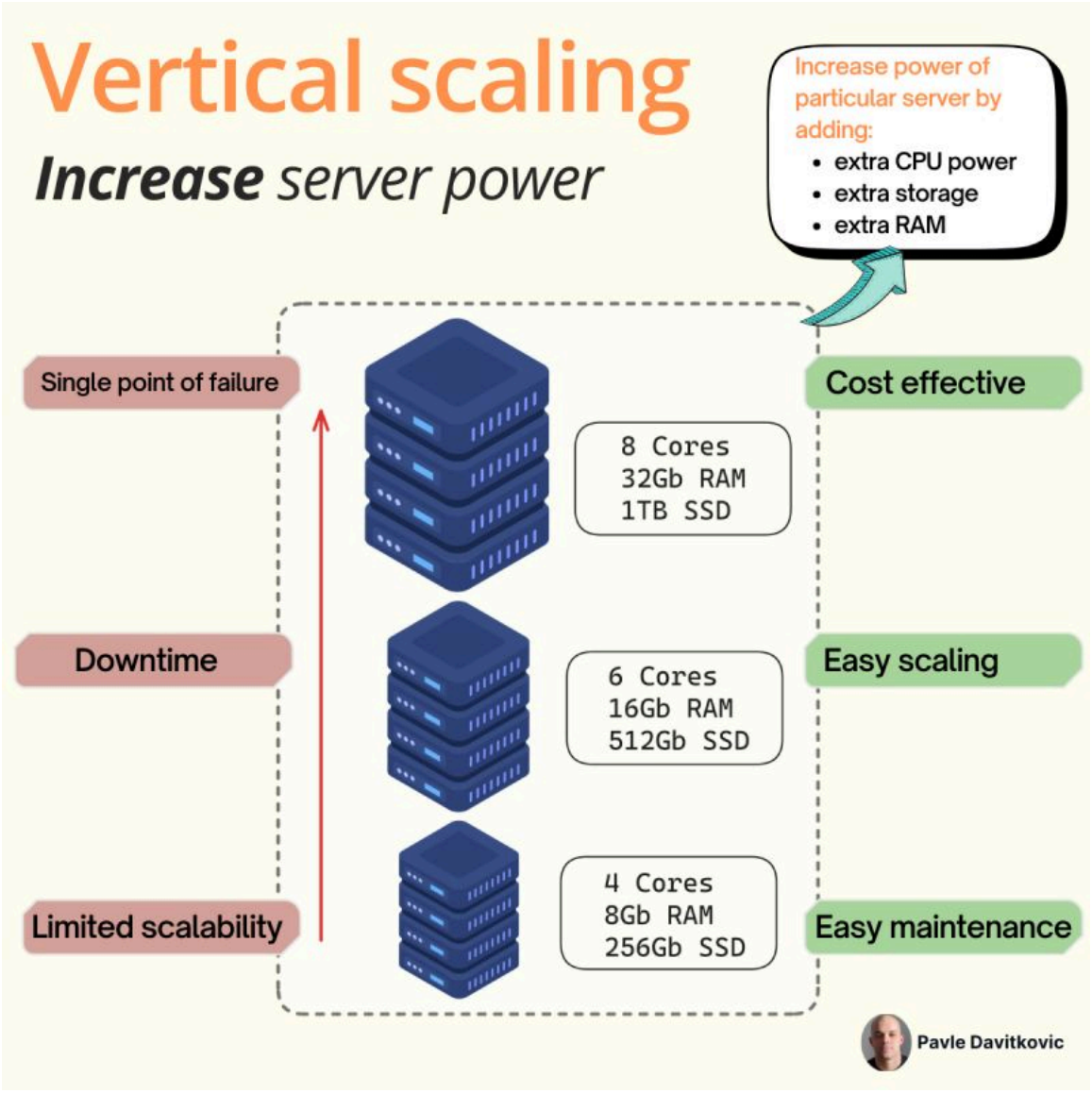
The harsh truth is that your application likely won’t hit 10 million users overnight, and fancy cloud services might be overkill at the start.

Instead, do this:

- Set up Open Telemetry
- Observe
- Scale up

Then, if needed, scale out.

Simple as that.



20 14 comments 2 reposts

Like Comment Repost Send