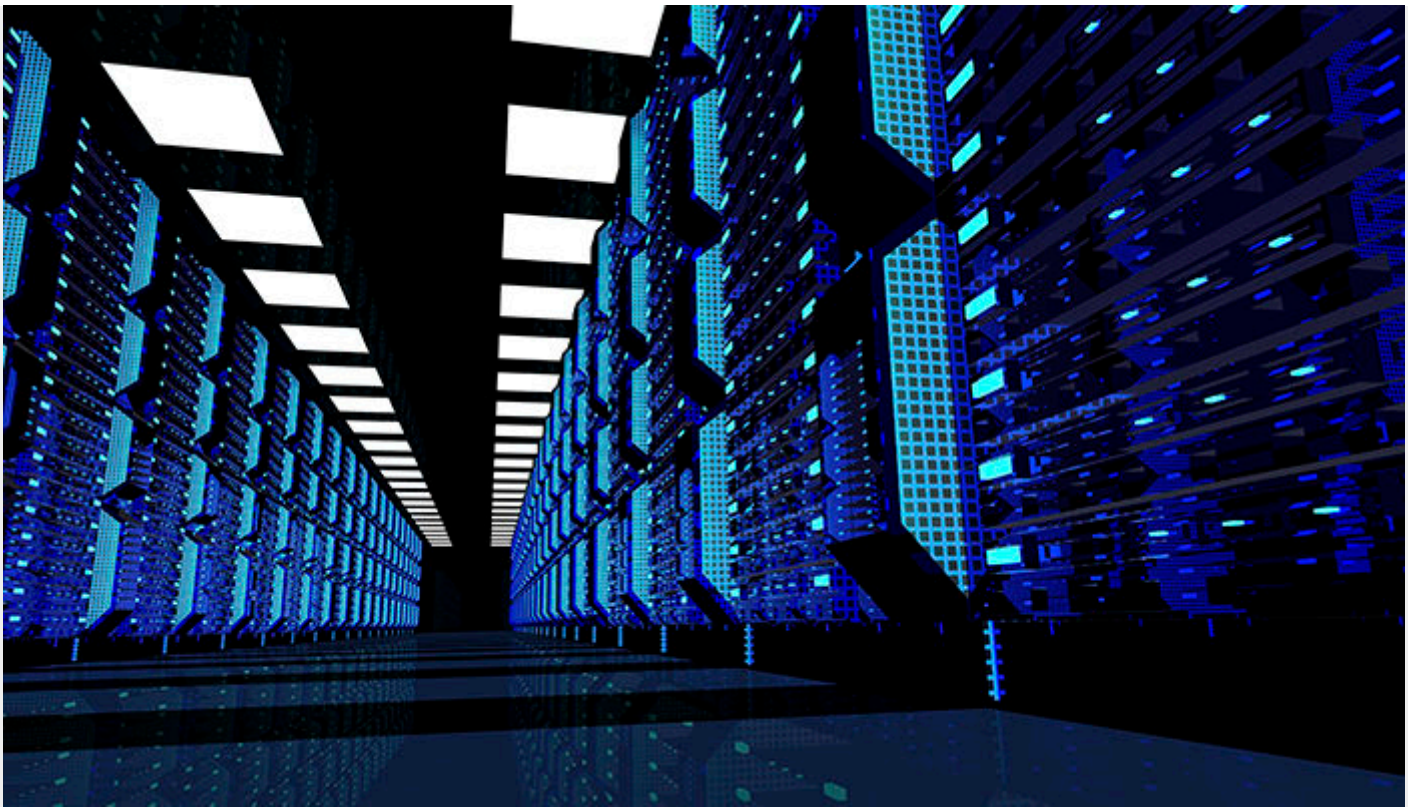
WHAT IS A HYPERSCALE DATA CENTER?



We can all name a few of the biggest technology companies in the world, like Google, Amazon, Microsoft, IBM, even Facebook. They all developed technologies that millions, sometimes billions, of people take for granted on a daily basis. They have made technology inherent to our daily routines.

But how do these technologies work? What fuels these companies? How can they maintain a significant global presence, increasingly existing in the cloud, while maintaining the speed and reliability required by life in the 21st century?

The answer is **hyperscale data centers**.

Industry intelligence firm Synergy Research Group recently studied hyperscale data centers (HDCs) around the world. By the end of 2017, nearly 400 hyper data centers were in operation. More interesting, perhaps, is that only 24 companies are responsible for building and maintaining these, averaging about 16 HDCs each.

Skewing that average, though, are U.S. behemoths who make up 44 percent of the world's HDCs. That's because American companies Amazon, Google, IBM, and Microsoft each have at least 45 data centers around the world – nearly half of the HDCs in the world. And they aren't stopping any time soon. SRG estimates that by the end of 2019, close to 500 HDCs total will be in operation or planning stages.

Let's take a look at hyperscale data centers, particularly how HDCs are more than just hypersized versions of enterprise data centers.

The size of hyperscale

The term "hyperscale" refers to a computer architecture's ability to scale in order to respond to increasing demand. Computers rely on resources within a given node or set of nodes. Scaling some part of our computer architecture typically means increasing computing ability, memory, networking infrastructure, or storage resources.

The goal of scaling is to continue building a robust system, whether that system revolves around the cloud, big data, or distributed storage, or, as is increasingly likely these days, a combination of all three. Of course, "hyper" indicates an extreme or excess. Hyperscale is not just the ability to scale, but the ability to scale hugely and quickly.

We often think of scaling "up", which is a vertical approach to architecture, often adding more power to the existing machines. But another way to increase capacity is by "scaling out", a horizontal approach, such as adding to the overall machines in your computing environment.

In general, companies deal with hyperscaling in three arenas: in the physical infrastructure and distribution systems that support data centers, in the ability to scale computing tasks (magnitudes of order, both in overall performance and in advancing the company's status quo), and the financial power and revenue sources of companies that require such hyperscale systems.

Data centers: Enterprise vs. hyperscale

At the core of most companies is a data center, any facility that houses computer systems and related components, like storage systems and telecommunications. Redundancies are also built into these environments, in case power, environmental factors, or security systems go down. The size of your company, and the size of your computing power, determines how large or how many data centers are necessary, but it is common for a single data center for a large-level enterprise to use the same amount of electricity as a small town. It's also common that such an enterprise may only require one or two of these data centers.

Maintaining an enterprise data center is no small feat. You are constantly managing the data center's environment to ensure consistent machine behavior, developing a patching schedule that allows for consistent fixes and minimizes downtime, and rushing to fix inevitable failures of any kind.

Contrast an average-size enterprise with a Google or IBM. While there isn't a single, comprehensive definition for HDCs, we understand that, at their most basic, HDCs are significantly larger facility than a typical enterprise data centers. Market intelligence firm International Data Corporation generally defines a data center as "hyperscale" when it exceeds 5,000 servers and 10,000 square feet. Of course, these are the locations that just qualify as an HDC. Some hyperscale data centers house hundreds of thousands, even millions, of servers.

But IDC says there is more that sets these hyperfacilities apart. Hyperscale data centers require architecture that allows for a homogenous scale-out of greenfield applications – projects that really have no constraints. Add to that an enormous infrastructure that is increasingly disaggregated, higher-density, and power-optimized.

Hyperscale companies who rely on these data centers also have hyperscale needs. While most enterprise companies can rely on out-of-the-box infrastructures from tech vendors, hyperscale companies must personalize nearly every aspect of their computing environment. Building in specific capabilities at massive scales, controlling every aspect of the computing experience, and

manipulating every configuration. At a scale this size, no one can do it better than the company can do for itself. It is the very cost of these demands who limits exactly who can join the hyperscale club.

Of course, for companies who operate hyperscale data centers, the cost may be a barrier to entry, but it isn't the issue. Automation is.

Companies who run hyper data centers instead focus on automating, or, "self-healing", a term that describes an environment where inevitable breaks and delays occur, but the system is so controlled and automated, it will adjust to correct itself. This self-healing automation is important because it encourages significant efficiency from the data.

Inside a hyperscale data center

A small town in the rural center of Washington State, Quincy is home to several hyperscale data centers for companies including Microsoft, Yahoo, and Dell. These companies like the region for the same reasons its farmers do: moderately cool temperatures, affordable land prices, and wide-open spaces. The wide-open spaces are significant when you think about how big these are. Microsoft's HDC in Quincy, one of more than 45 the company operates, comprises 24,000 miles of network cable. For comparison, that's just barely under the total circumference of Earth or six Amazon Rivers, for something a little more human-scale.

Another Microsoft HDC, which supports Azure, is located in Singapore, itself not a big place. That hyperscale data center has enough concrete in it to build a sidewalk between London and Paris, which would total about 215 miles in length. It is precisely this magnitude and complexity of scaling that makes a data center one of hyperscale.

Unlike many enterprise data centers, which rely on a large full-time staff across a range of disciplines, many HDCs actually employ fewer tech experts because so much of the technology is automated. In fact, it's quite possible that hyperscale companies employ more people at HDCs as security staff than as tech experts, because the data in these centers is the product.

As we build our reliance on technologies that make the world seem like a smaller place, we must concede that our data gets to see much more of the world than we may – and a lot faster. Talk about globalization.