

What is Hyperscale?

eBook

Contents

- 03** **Introducing Hyperscale**
 - What is Hyperscale?
 - Features of Hyperscale

- 06** **Comparison to Other Types of Data Center**
 - Retail Colocation Data Center
 - Wholesale Colocation Data Center
 - Enterprise Data Center
 - Telecom Data Center
 - The Hyperscale Data Center

- 08** **What Makes a Hyperscaler?**
 - Technology
 - Maintenance
 - Power Consumption
 - Architecture
 - Connector Density and Fiber Count
 - Workloads

- 12** **What's Next for Hyperscalers?**
 - Technological
 - Connectivity
 - Commercial
 - Demand for Data

- 16** **The Internet of Things and Edge Computing**

- 18** **In Conclusion...**



Introducing Hyperscale

Hyperscale computing powers our on-demand world, yet few truly grasp the enormity of it. This guide introduces the concept and significance of hyperscale data centers.

In this first eBook, of the Hyperscale Rising series, we are going to unpack everything from the definition of hyperscale to how it enables so many households' digital brands to operate at the scale they do. We will also be exploring the past, present and potential future of some of the data center giants.

Contributors

Alan Keizer
Senior Technology Advisor

Keith Sullivan
Director of Strategic Innovation

What is Hyperscale?

hyper scale

Hyper - Prefix, excessive, greater than usual
Scale - Verb, to change in size or number

Every day, we effortlessly binge-watch on-demand TV shows, use apps to order cab rides, and chat with friends worldwide without a thought about the immense scale of apps like Netflix, Uber, and Facebook. To provide a view, Netflix has a viewership that's grown beyond 200 million subscribers¹, Uber facilitates millions of rides daily², and Meta's community has expanded to a remarkable 2.9 billion monthly³. So, what's the secret sauce behind this vast digital expanse? It's cloud computing.

Cloud computing is the foundational pillar for these digital giants, offering global resource availability and elasticity. Diving deeper, you will see it is the hyperscale data centers that stand as the true champions of the cloud.

The term 'hyperscale' encapsulates the capability of these data centers to automatically adjust and cater to fluctuating demands. Whether it is the release of a Netflix blockbuster or a trending X or Instagram hashtag, hyperscale ensures uninterrupted service, dynamically allocating resources as needed.

But hyperscale data centers are not just about sheer size. They represent efficiency, adaptability, and resilience. Crafted to manage extensive tasks, from streaming to social networking to cloud services, they adapt seamlessly to user activity.

Hyperscale data centers ensure our online experiences remain smooth and efficient, irrespective of the digital traffic.

1. [Netflix Fourth-Quarter 2022 Financial Results](#)
2. [Uber's Impact on Ridesharing](#)
3. [Facebook Users Worldwide 2022](#)

Features of Hyperscale?

Hyperscale has shifted IT networks from on-premise computer rooms to huge fleets of data centers. There are three key characteristics that really help to define hyperscale: the physical structure, how incoming traffic is processed, and how software is used to automate different functions.

01

Physical Structure

Hyperscale goes beyond a single data center; it represents a comprehensive network of interconnected facilities. These data centers have evolved into cohesive, low-latency clusters, some in close proximity and others spanning continents. This distributed architecture is strategic: it mitigates risks from natural disasters and power outages, and by situating in various regions, it ensures optimal service while adhering to regional data regulations.

02


Incoming Traffic

Hyperscale operators oversee vast networks, bridging the gap between their users and data centers. They first receive traffic at network edge locations, often shared with Internet Service Providers (ISPs)s and Internet Exchanges. This traffic is then directed to specific clusters, termed 'regions', within their global network. Each region can house several individual data centers, categorized into availability zones.

03

Software Automation

Software automation is the heartbeat of hyperscale data centers. It not only ensures efficient workload mobility, directing tasks to optimal resources, but also streamlines resource provisioning, dynamically allocating or de-allocating assets based on demand. This adaptability is crucial for resource management and load balancing, distributing tasks evenly to prevent overloads. Policy-driven networking enhances network agility, automating configurations.



Comparison to Other Types of Data Center

Fundamentally, hyperscale data centers are large private web facilities, colocation and wholesale data centers lease space and power. Hyperscalers sometimes lease extra capacity to expand faster. This symbiotic relationship enables rapid scaling benefiting both. Both provide critical digital infrastructure.

As of the end of 2023, there will be over 900 data centers operated by hyperscalers with an additional 312 already in the pipeline for the next year compared to colocation data centers of which there are over 9,900⁴.

Although hyperscale data centers represent just over 10% of the number of total data centers⁵, they have a disproportionately significant impact on the market. In terms of market share, hyperscale data centers are expected to account for 62% of the global data center market by 2032⁶. Non-hyperscale data centers are expected to account for the remaining 38% of the market in terms of their overall spend.

4. 451 Research
Datacenter
KnowledgeBase
5. S&P Global Market
Intelligence
6. Hyperscale Data
Center Market Size,
Trends, Growth,
Report 2032



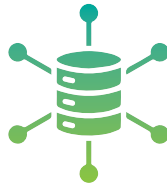
Retail Colocation Data Center

Colocation data centers consist of one data center owner selling space, power and cooling to multiple enterprise and hyperscale customers in a specific location.



Enterprise Data Center

An enterprise data center is a facility owned and operated by the company it supports and is generally built on site but can be off-site in certain cases also or leverage third party colocation facilities.



Wholesale Colocation Data Center

Wholesale colocation data centers consist of one owner selling space, power and cooling to enterprise and hyperscalers like a standard colocation data center. In these instances, interconnection is not really a requirement. These facilities are used by hyperscalers or large companies to hold their IT infrastructure.



Telecom Data Center

A telecom data center is a facility owned and operated by a Telecommunications or Service Provider company such as BT, AT&T or Verizon. These types of data centers require very high connectivity and are mainly responsible for driving content delivery, mobile services, and cloud services.

The Hyperscale Data Center



- A hyperscale data center is leased (or, in some cases, owned) and operated by the company it supports (this includes companies such as AWS, Microsoft, Google, and Meta)
- They offer robust, scalable applications and storage services to individuals or businesses.
- Hyperscale computing is necessary for cloud and big data storage.
- At least 100,000 _ sq. ft. in size they are usually upward of 3,000 racks of compute and storage servers, connected with an ultra-high-speed, high fiber count networking.
- There are 250,000 switch ports in the computing fabric of a typical hyperscale data center.

What Makes a Hyperscaler?

They invest in entire pre-configured server and switch cabinets, essentially operating in units of complete racks rather than individual servers. But this expansive vision isn't confined to hardware alone; it's about global reach. Hyperscale networks are swiftly erecting new data centers, extending their availability zones at a remarkable rate.

Technology

Hyperscalers represent the cutting edge of data center innovation and operations. Their sheer size provides them with unique advantages in deploying the latest technologies. Hyperscalers have unparalleled purchasing power and close relationships with component suppliers, allowing them to customize and optimize solutions to an extent unavailable to other data center operators. They walk a fine line between costs and efficiency, driving innovations that enhance performance while maximizing value. Though not all data centers can match the scale and buying leverage of hyperscale players, these giants provide a model of the potential of data center technology fully realized. Their push for continuous improvement serves as a rising tide that lifts the entire industry to new heights of efficiency and capability. While hyperscalers enjoy a privileged position, their pioneering spirit serves as a guiding light for all data center operators seeking to optimize infrastructure and future-proof operations.

Hyperscalers accounted for almost half of the global server shipments in 2022

Enterprise data centers use robust cooling systems to maintain server temperatures, serving a wide range of clients with different needs. On the other hand, hyperscalers have made notable progress in data center design. They have a comprehensive approach to thermal management, from hall layouts to airflow patterns and server configurations. Additionally, they've made strides in Layer 1 infrastructure technologies, improving connectivity and optimizing physical configurations.

Global Server Revenues to Grow 17% YoY in 2022.

Maintenance

There's a saying amongst those that operate large, virtualized data centers; treat your servers like cattle, not pets. For any small to medium business, server issues can halt operations in a heartbeat. But within the sophisticated world of hyperscale data centers, there's a backup plan.

They're equipped with smart management software that jumps into action when hardware stumbles. It reroutes tasks and processes, ensuring continuity and crucial uptime.

However, it's only during significant malfunctions that the need for crash carts arises, highlighting the robustness of their system. It's worth noting the astonishing operational efficiency and up time achieved by these sprawling centers.

Remember the days when mending a server took nearly an hour? Those days are behind us. Now, in many cases, it's a swift no-hands procedure that takes less than two minutes. To add a cherry on top, it's not uncommon for a single technician to handle an impressive load of 25,000 servers in a given shift.

Power Consumption

In 2023, the power consumption of hyperscale data centers has become a focal point in discussions about global energy usage. These massive facilities, designed to accommodate the ever-growing digital demands of our world, have a significant energy footprint. As per recent insights, data centers are estimated to account for about 3% of global electricity consumption. Specifically, the average hyperscale facility consumes between 30 and 60 Mega Watts, 24 hours a day, 7 days a week. To provide a tangible perspective, this amount of electricity is theoretically sufficient to power about 47,000 average US homes or about 125,000 average UK homes.

However, it's essential to understand that while these numbers seem staggering, hyperscale data centers are at the forefront of efficiency and sustainability. Their sheer size and scale allow them to implement advanced cooling, power distribution, and energy-saving technologies that smaller data centers might not be able to afford or justify. This efficiency is not just about saving costs; it's about reducing the overall carbon footprint and supporting global sustainability efforts.

A study by the Lawrence Berkeley National Laboratory suggested that transitioning 80% of US servers to optimized hyperscale facilities could result in a 25% reduction in energy use⁷

The most significant cost reductions stem from optimizing the utilization of cloud servers. Typically, enterprises utilize just a fraction, around 10%, of their actual server capacity. However, through the practice of sharing virtual servers, efficiency levels can be elevated to 30% or even higher. This presents a welcome advantage for businesses that depend on these data centers. By choosing to host their digital operations in these hyperscale facilities, they play a role in promoting more efficient energy utilization. Instead of numerous smaller data centers drawing power at suboptimal rates, these colossal centers efficiently consolidate the demand and manage resources more effectively, providing a silver lining for environmentally-conscious businesses.

7. [Recalibrating global data center energy use estimates](#)

Architecture

Traditional enterprise networks operate on a three-tier architecture: core, aggregation, and access. However, as data demands grow and applications require quicker responses, this structure shows its inefficiencies. Depending on the network's layout, data might need to travel just to the distribution layer or go up to the core and back down, leading to inefficiencies.

Remember the '90s? A simple web search took an age due to limited bandwidth. Fast forward to today, and we can instantly search text, images, videos, and more, with ads tailored to our queries, all thanks to advancements in connectivity and software automation.

Enter the leaf-spine architecture, the game-changer. Its strength lies in connecting vast numbers of servers efficiently. Regardless of where data starts or ends in this setup, it travels the same distance, ensuring consistent latency across all paths.

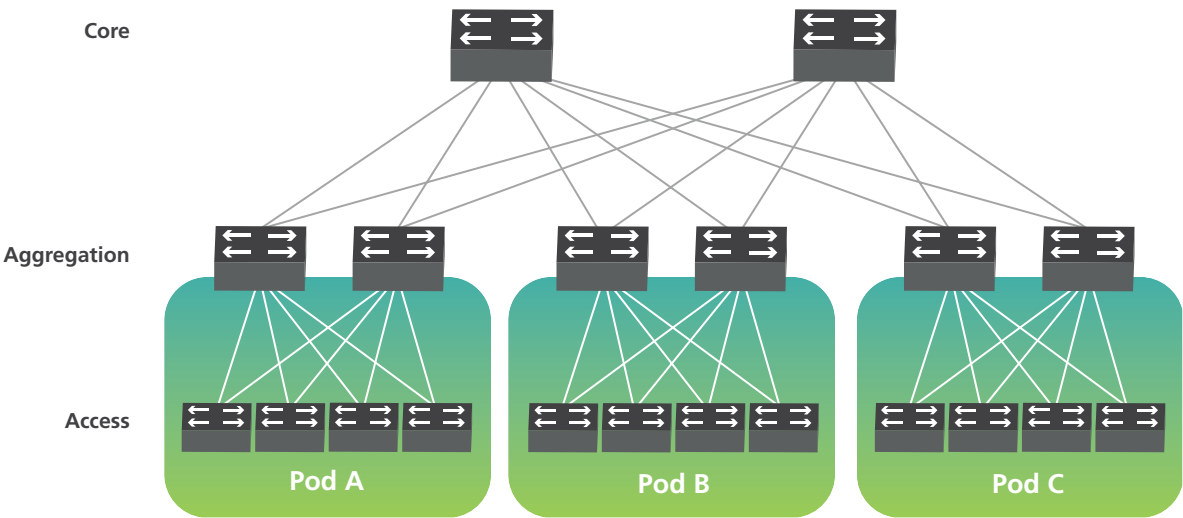
In hyperscale data centers, a staggering 90% of traffic remains internal, termed as East-West traffic.⁸

Contrast this with North-South traffic, which represents data entering or leaving the data center. Consider the intricacies of loading a Facebook page; that's East-West traffic in action.

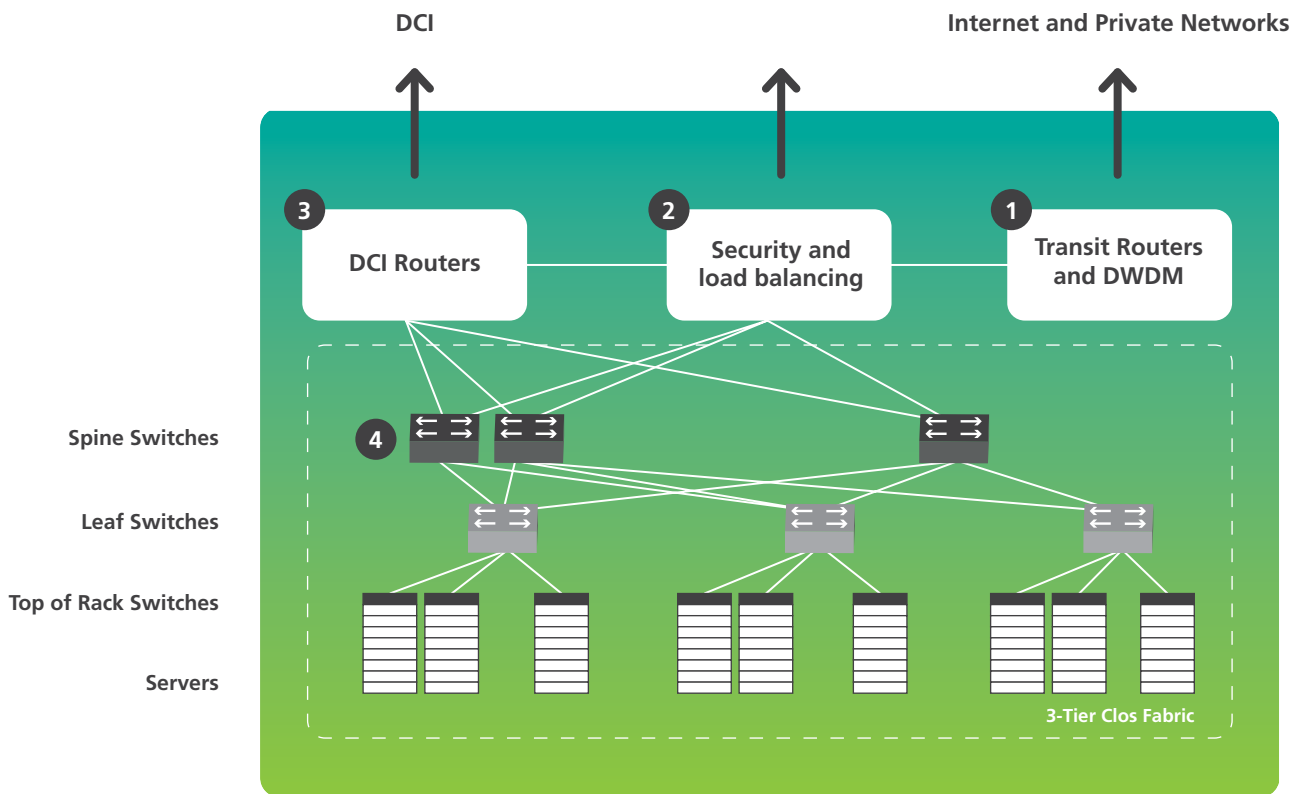
Leaf-spine excels in hyperscale settings, especially with distributed server architectures. Here, resources for a specific task spread across multiple servers. This design ensures minimal latency, avoids traffic bottlenecks, and guarantees unblocked communication even when all servers operate at full capacity.

8. [Cisco Global Cloud Index: Forecast and Methodology, 2018–2023](#)

Traditional 3 Tier Design



Typical Multi-tier Leaf-Spine Architecture



Connector Density and Fiber Count

The growing demand for increased bandwidth and cloud services has led to a surge in fiber deployment within data centers. Hyperscale data centers, in particular, heavily rely on high fiber counts due to substantial data storage and processing needs.

Connecting data center buildings requires substantial data capacity, however limited duct space is a challenge. High-fiber-count cables maximize duct space efficiency. External connections to neighboring data centers already comprise over 10,000 optical fibers. Ultra-high fiber count cables, with counts of up to 6,912 today and more on the horizon, are preferred to optimize space in hyperscale data centers.

Hyperscale data centers use an average of 12,000 miles of optical fiber.

Microsoft's Chicago data center boasts over 24,000 miles of network cable, almost sufficient to encircle the Earth.⁹

These statistics highlight the vital role of fiber optics in modern data center infrastructure.

9. [How Hyperscale Data Centers Evolve the Future of IT](#)

Workloads

Hyperscale demands and service level expectations surpass those of customers and end-users, encompassing uptime and load times. Hyperscale operations involve the automatic allocation of available resources for incoming tasks, utilizing provisioning (server setup for immediate use) and orchestration (workload management and allocation).

According to Gartner, by 2025, 95% of new digital workloads will use cloud-native platforms, requiring automation to manage complex workloads across distributed environments.¹⁰

These workloads extend beyond high-traffic websites and encompass complex tasks such as 3D rendering, advanced analytics, or AI training. These workloads may be hosted on higher performance dense systems, often leveraging GPU's or other specialized processors in some cases.

10. [Gartner Says Cloud Will Be the Centerpiece of New Digital Experiences](#)

Where next for Hyperscalers?

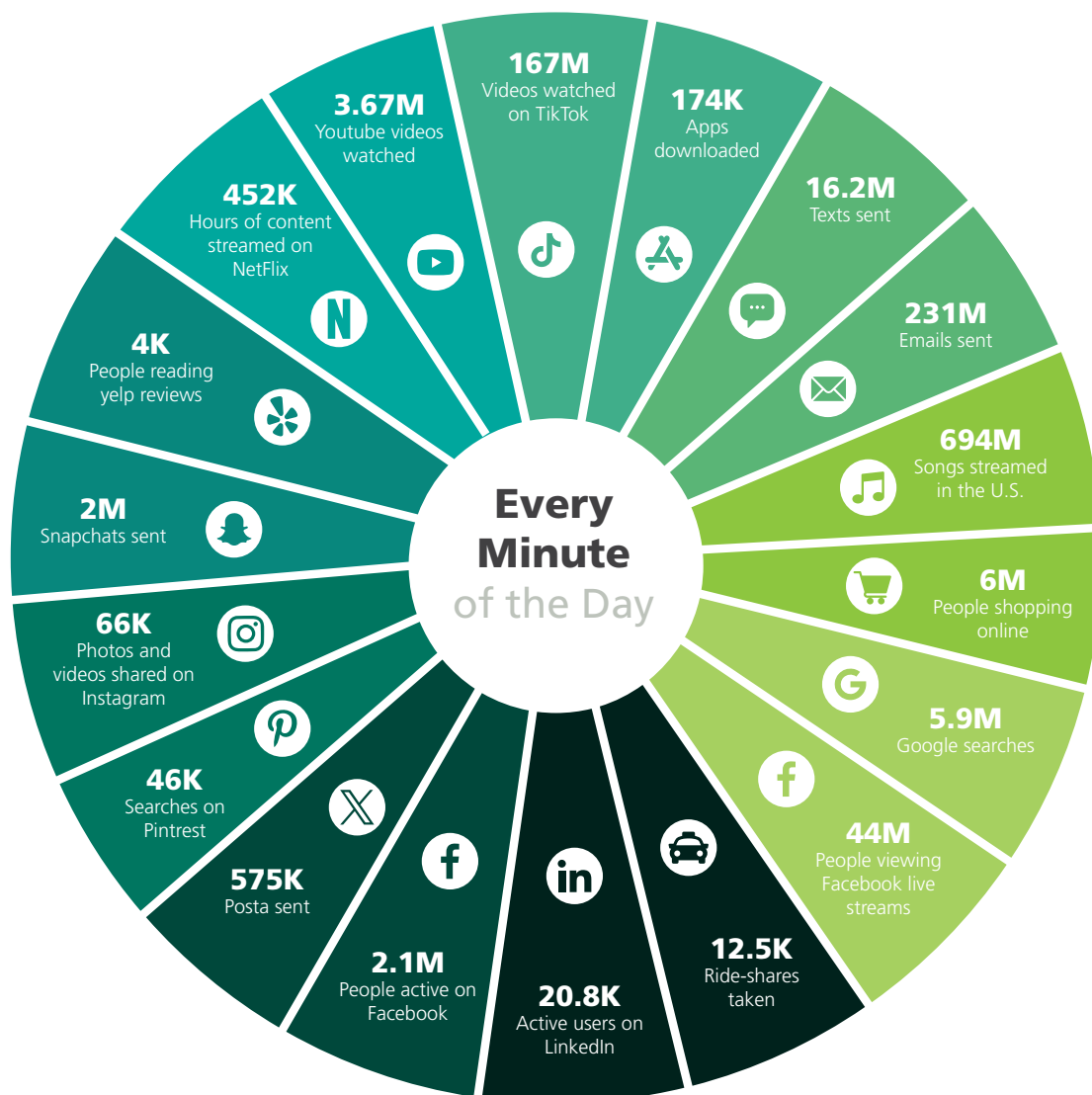
Hyperscale has reshaped business, commerce, and society through the delivery and availability of IT services and the future of hyperscale can be divided into two categories: technological and commercial.

Technological

Hyperscalers need to keep their eye on the speed of innovation and growth of deployment of new technologies. With the emergence of artificial intelligence, 5G, Internet of Things (IoT), augmented reality, and autonomous vehicles, the demand for data is set to grow steadily then take off fast.

All these applications will drive the need for more data, faster, simultaneously ensuring real-world suitability whether through personal safety in the case of autonomous vehicles, or user experience for augmented reality.

This growth in data consumption is evident if we simply examine the growth that has occurred in the past year. A lot can happen on the Internet in a minute.



11. This is What Happens in a Minute on the Internet, World Economic Forum

Connectivity

The ability for hyperscalers to be globally available and offer low-latency connections in any country is of critical importance. Hyperscale data center fleets are already racing to grow their footprints in Europe, Asia-Pacific and North America, with the focus now turning on establishing stronger footprints in South America, the Middle East, and Africa.

This brings with it the need for higher levels of fiber connectivity, and a stronger global circulation of subsea cables, cable landing stations, exchange nodes, and emerging edge data centers. This will all translate to higher fiber counts running into and out of the data center.

Commercial

As data center revenue grows and the demand for data increases, the expansion of the hyperscale data center infrastructure will continue. With that, commercial drivers, as always, will focus largely on speed and efficiency.

Demand for Data

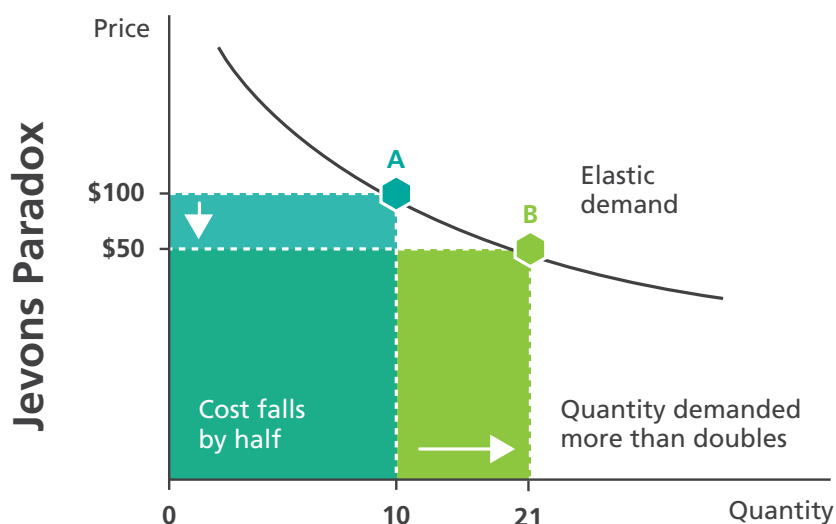
According to a prediction published on the Gartner Blog Network, 80% of enterprises will upgrade their traditional data centers by 2025 as hyperscale data centers rapidly become the norm.

By 2025, hyperscalers will account for 60% of the global datasphere. This is a significant increase from 2016, when hyperscalers accounted for only 19% of the global datasphere.

The hyperscale industry is set to expand, especially with the surge in data creation that demands storage and processing. With 5G rolling out and the Internet of Things (IoT) connecting more devices, we're seeing a rise in smaller, localized data centers, often referred to as "edge" data centers. These centers bring cloud resources and cached content closer to users. Now, you might think edge computing could slow down the hyperscale momentum, but in reality, all this data eventually finds its way back to the cloud for in-depth analysis when time isn't pressing.

Here's an interesting concept: the Jevons Paradox. In the world of economics, this paradox highlights a scenario where improved efficiency in using a resource doesn't lead to less consumption. Instead, as it becomes more available, its demand goes up. This idea rings true for data centers.

While 5G promises to cut down the time between a device and its edge data center, it doesn't necessarily mean we'll use less data. In fact, as data becomes more accessible, we'll likely use even more, underscoring the ongoing importance of hyperscale data centers.



Everything as a Service

Everything as a service (XaaS) encompasses SaaS, PaaS, and IaaS, the three main types of cloud computing.

SaaS: Software-as-a-Service

Software that's available via a third-party over the Internet.

PaaS: Platform-as-a-Service

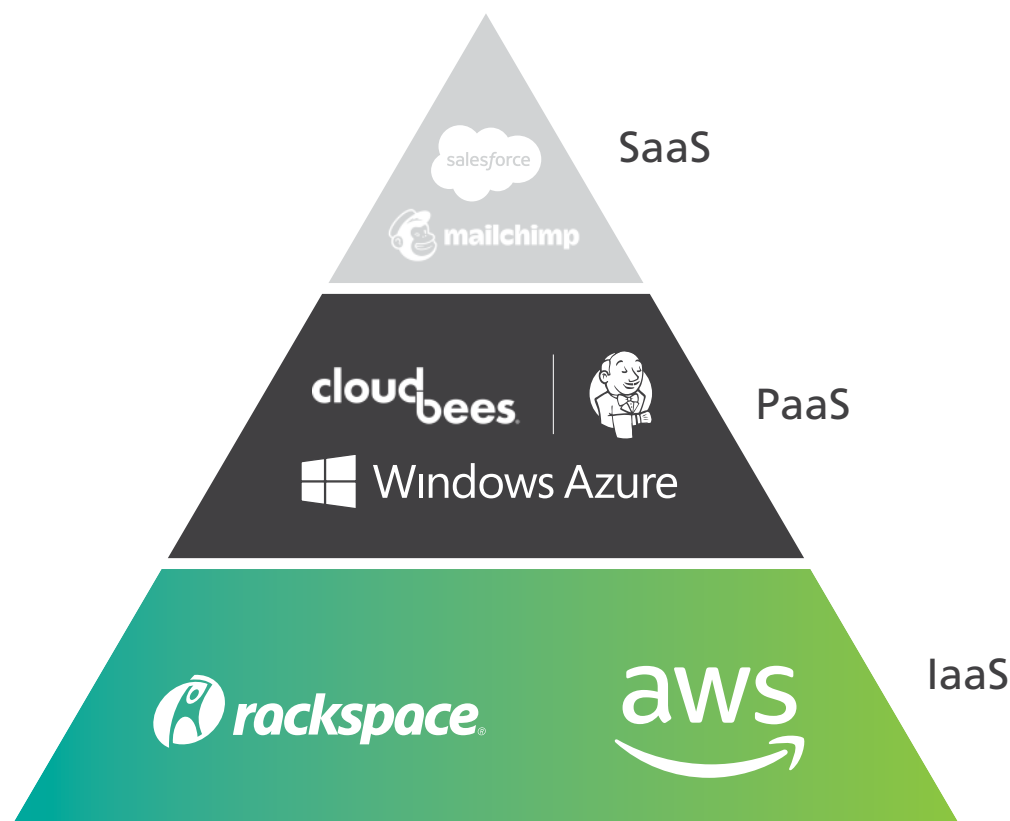
Hardware and software tools available over the Internet.

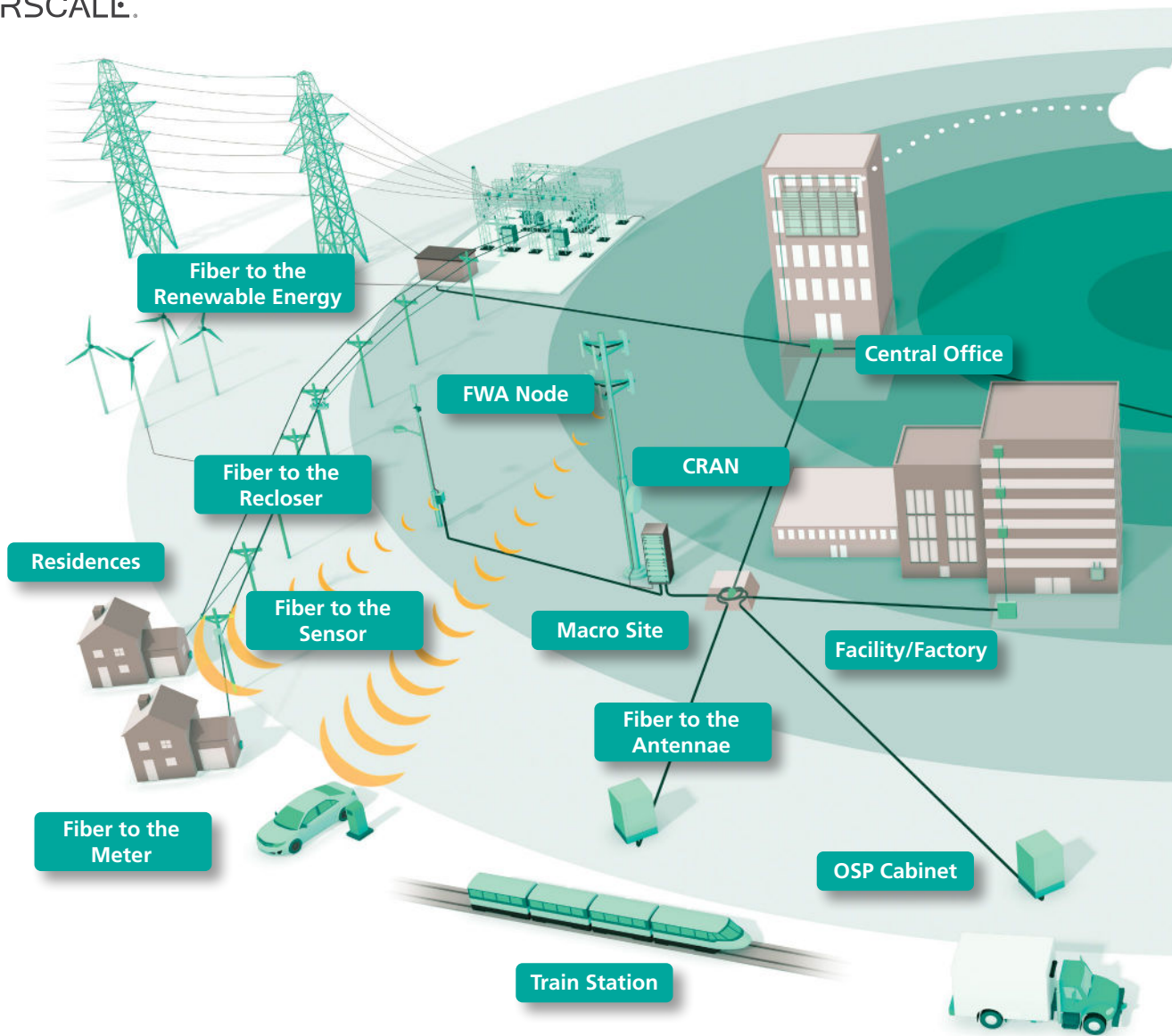
IaaS: Infrastructure-as-a-Service

Cloud-based services, pay-as-you-go for services such as storage, networking, and virtualization.

XaaS plays a huge part in cloud computing and in hyperscale data centers. This trend is set to continue as businesses move from a CapEx model to a subscription-based one with more and more in-house processes and services are being phased out in favor of outsourcing.

While XaaS as a whole is set to flourish, it is doubtful that there will be many more IaaS providers emerging as aspiring providers can buy space from established services such as AWS at a cheaper price than they can build their own – a very attractive prospect when faced with the daunting amount of resource and funds needed to carve out even a small market share. Instead, more SaaS providers will start to emerge and join the ranks of the above.





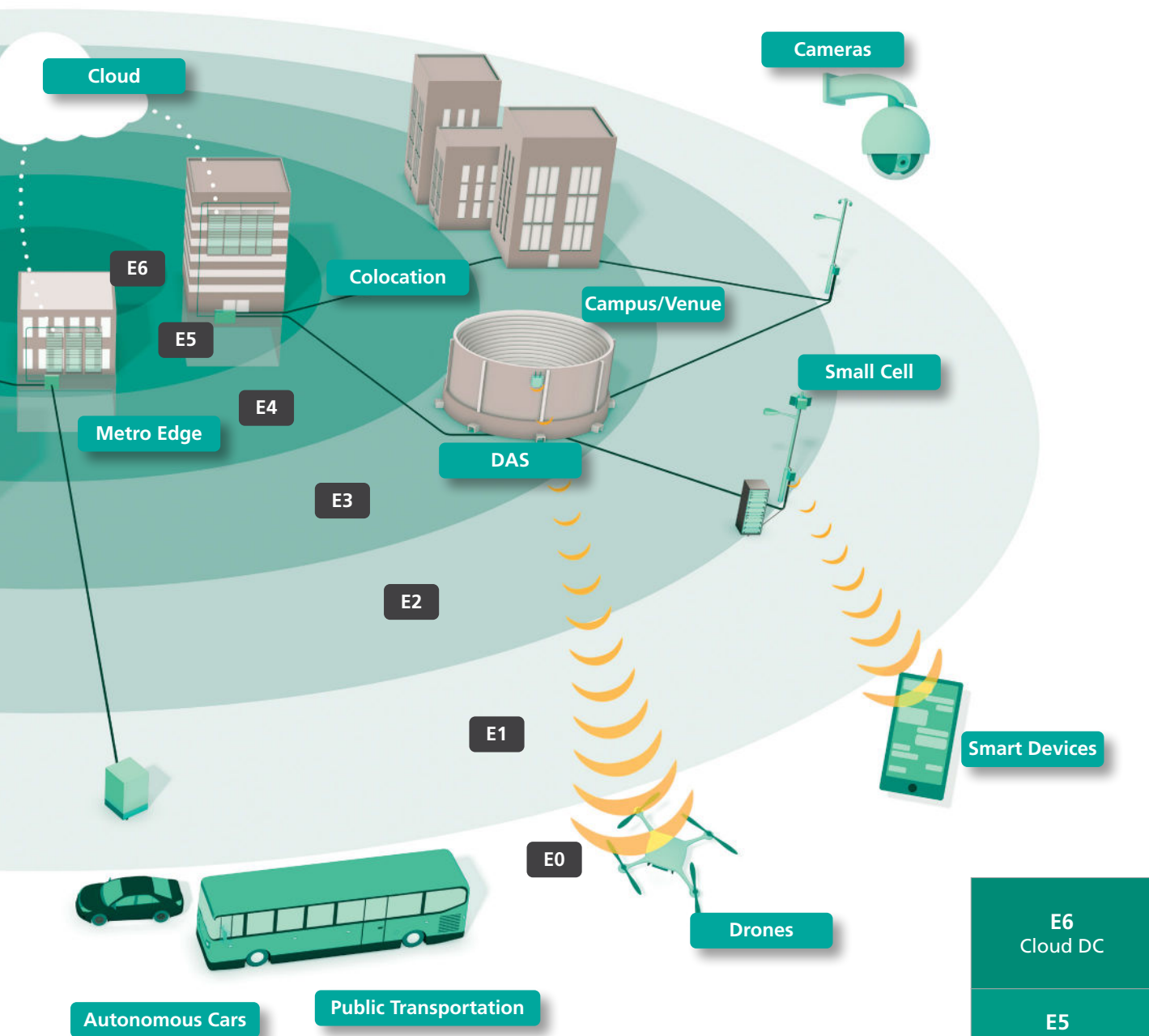
The Internet of Things and Edge Computing

One of the more notable trends in terms of technological advancement is the emergence and expansion of the Internet of Things (IoT). This refers to the interconnection of computing devices embedded in everyday objects, enabling them to send and receive data. As more and more devices are developed, the data generated and required to connect and interconnect these devices will develop accordingly. From smart home gadgets and wearable technology to automated vehicles and big data analysis in healthcare, IoT is ever-expanding, becoming an ever more crucial part of everyday life, making the reliability and performance of these devices and the networks that support them more important than ever.

Underpinning IoT is edge computing. Edge computing centers around the principle of bringing computing as close to the source of data as possible in order to reduce latency and bandwidth use. Put simply, edge computing takes processes that would ordinarily take place in the cloud and moves them to more local places such as a user's computer, an IoT device, or an edge server.

By moving these processes to the network's edge, communication between a client and server happens over shorter distances thus reducing latency and bandwidth use.

A good example of this would be an autonomous car with hundreds of sensors programmed to continuously gather data while the car is in motion.



According to Dell, connected vehicles will generate \$150 billion in annual revenue, grow to 100 million vehicles globally, and as a result transmit over 100 petabytes of data to the cloud per month by 2025, requiring 10 exabytes per month, approximately 10,000 times larger than the present volume.¹²

While IoT and edge computing are still in their infancy, there is no doubt that they will continue to grow, generating, processing and demanding more data than ever before.

12. Expanding Human-Machine Partnerships Through Connected Cars, AI, and IoT, DellEMC

E6 Cloud DC
E5 MetroEdge Colocation
E4 Central Office
E3 Facility, Factory, Campus, Venue
E2 Macro Site, Small Cell/FWA, CRAN, Private LTE
E1 OSP Cabinet/ Fixture
E0 End Device

In Conclusion...

In conclusion, hyperscale computing is a dynamic and transformative force that shapes our digital landscape in ways often taken for granted. It is the engine behind our seamless online experiences, from streaming movies to ride-sharing apps and connecting with friends on social media. Hyperscale data centers, with their massive scale, adaptability, and efficiency, play a pivotal role in ensuring uninterrupted service in our on-demand world.

As we've explored the key features of hyperscale, including its physical structure, incoming traffic management, software automation, and more, it's evident that hyperscale data centers represent the epitome of technological prowess and innovation. They are the biggest machines in the world. Their ability to handle enormous workloads, maintain uptime with smart management software, and optimize power puts them at the center of the ever-expanding digital ecosystem.

Looking ahead, hyperscale computing is set to continue its meteoric rise, driven by continuing technological advancements and growing data demands. The global expansion of hyperscale data center infrastructure is inevitable, with an emphasis on speed, efficiency, and connectivity across regions. Our next eBook, *Hyperscale Rising*, looks at this in more detail.



AFL HYPERSCALE®

Data Center Cabling and Connectivity Experts

AFL Hyperscale creates purpose-built fiber optic connectivity solutions for data centers.

As the first cabling and connectivity provider focused purely on hyperscale, colocation, and enterprise data centers, we intimately understand the unique infrastructure, performance, and scaling challenges facing these facilities.

With deep expertise in data center interconnection, white space, and emerging technologies like AI, we pioneer innovative fiber optic solutions engineered specifically for the demands of high-density data centers.

Leveraging 40+ years of optical networking expertise, our global team provides purpose-built connectivity solutions that enable rapid deployment, maximum uptime, and reduced costs for data centers worldwide. Backed by decades of experience, we offer invaluable guidance on optimizing network design to deliver future-proof solutions anywhere they are needed.

AFL Hyperscale. The World, Connected.

www.aflhyperscale.com

The information contained within this eBook is accurate and up-to-date to the best of our knowledge at the time of production. All graphs and visual representations are proprietary assets of AFL Hyperscale. These materials are intended for informational purposes only, and may not be used for commercial purposes without express permission from AFL Hyperscale.

Copyright © AFL Hyperscale 2023 All Rights Reserved E&OE AFLHSWHATISHYPERSCALEEB111023