Navigating the Shift from Hyperscale to Microclouds

Contents

Introduction

During the early 2000s, Amazon's retail business was expanding rapidly, driving the need for a reliable and scalable infrastructure to handle the increasing online traffic and sales volume. Realizing the potential in monetizing its infrastructure expertise and underutilized data center capacity during non-peak times, in 2006, AWS officially launched its first services, including Simple Storage Service (S3) for storage and Elastic Compute Cloud (EC2), for scalable computing without the overhead of managing physical infrastructure. The services were available on a pay-as-you-go basis, revolutionizing the way companies handled their infrastructure forever.

Following Amazon's vision, Google started offering cloud computing services with the launch of Google App Engine in 2008, and Microsoft launched Microsoft Azure in 2010. Around the same time, other significant players, like IBM, Oracle, and Alibaba, also entered the cloud computing space. Over the years, these platforms continued evolving, adding services and features to cater to diverse needs, such as IoT, machine learning, serverless computing, and more. Competition between these major providers intensified, leading to price reductions, geographical expansions, and service innovation.

Today, the key initiators have evolved into hyperscale giants with new vendors cropping up and maintaining their own niche markets. As cloud models evolve, extending from the vast hyperscale data centers to more compact, localized microclouds, CXOs and cloud teams face the crucial task of selecting the most fitting cloud model for their business needs. In this paper, emma explores hyperscale and microscale cloud computing paradigms to aid informed decision-making during this transformative phase of cloud adoption.

What’s Hyperscale?

Hyperscale is a system or infrastructure’s ability to rapidly handle growing demand and scale. In cloud computing, hyperscale refers to massive cloud data centers with an unusually large number of servers, storage systems, and networking equipment that can deliver cloud computing services on an extraordinary scale.

Leading cloud providers, like Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and IBM, have built data centers with thousands of servers to provide affordable, on-demand hyperscale computing. They are sometimes referred to as “hyperscalers”. Large distributed internet companies like Meta (Facebook), Apple, Netflix, and TikTok have also invested in hyperscale computing to address the challenges of scale, efficiency, global reach, and innovation inherent in their operations.

Background

The exponential growth of data and the increasing demand for real-time processing and analytics, fueled by trends like social media, IoT, big data analytics, and multimedia content, necessitated computing infrastructures and architectures capable of handling vast amounts of data. The demand for scalable, flexible, and high-performance computing infrastructure skyrocketed. Cloud pioneers like AWS, GCP, and Microsoft Azure responded by continuously expanding their infrastructure, building massive data centers across the globe.

In addition to the sheer scale of these data centers, the evolution of these CSPs (Cloud Service Providers) to hyperscalers involved developing highly efficient and scalable architectures using commodity hardware, advanced networking, and innovative software solutions to handle massive workloads, save energy, and offer competitive rates. It also involved building a diverse service portfolio beyond basic computing and storage, including machine learning, AI, IoT, serverless computing, analytics, and more.

Currently, there are over 900 operational hyperscale data centers, with several hundred more in development and expected to be operational by the end of 2024. Given the rapid, large-scale penetration of GenAI, analysts expect the global hyperscale capacity to almost triple in the next six years.

Characteristics of Hyperscale Cloud Data Centers

Hyperscale data centers operate on a vastly different scale and possess several characteristics that set them apart from traditional data centers.

Massive Size

Hyperscale data centers can span hundreds of thousands or even millions of square feet, hosting a vast number of servers, memory and storage systems, and networking equipment. A data center should span 10,000 square feet, accommodate 5000 servers, and provide 40MV of capacity counts to be considered a hyperscale data center. However, this is just the bare minimum. Google's 375-acre (16.3 million sq ft.) data center in Midlothian, Texas is one of the largest in the entire country.

Elastic Scalability

Hyperscale data centers utilize a modular design approach to scale horizontally, allowing seamless expansion by adding thousands of servers and storage nodes as needed. Because of this modular design. if a component fails or requires an upgrade, it can be replaced or upgraded independently without affecting the entire system.

Standardized, Commodity Hardware

Each component within the hyperscale data center is pre-fabricated and standardized to streamline installation, upgrades, repairs and replacements. Unlike standard data centers that house a mix of standard and specialized, proprietary hardware, hyperscale data centers use off-the-shelf, commodity hardware for cost-efficiency and flexibility.

These characteristics collectively enable hyperscalers to cater to the vast computational, storage, and networking requirements of modern enterprises and applications, offering scalability, reliability, and performance at a massive scale.

emma

Why Hyperscale Computing Matters?

Big data analytics involves processing and analyzing massive datasets, typically generated in real-time. It requires the computational power and distributed architecture of hyperscale computing. For instance, analyzing vast genomic datasets, conducting bioinformatics research, and supporting personalized medicine initiatives require the computational resources and storage capacity provided by hyperscale architectures. Training and deploying ML models also requires substantial computational resources and parallel processing. Hyperscalers can provide the infrastructure and resources needed to train ML models on large datasets to businesses that couldn't have otherwise invested in dedicated ML platforms. Essentially, complex workloads with dynamic resource requirements need the scalability and computational power of hyperscale data centers.

Use Cases

Companies that operate at a global scale and/or handle massive amounts of data and require extensive computing capacity need to tap into hyperscale infrastructure.

1. Social media platforms like Facebook (Meta Platforms), Twitter, Instagram, and Snapchat handle a colossal amount of user-generated content, interactions, and real-time updates.
2. E-commerce giants like Amazon, Alibaba, and eBay process a vast number of transactions and manage large product catalogs.
3. Streaming platforms like Netflix, Hulu, and Disney+ deliver high-quality video content to millions of users simultaneously.
4. Search engine providers like Google process billions of search queries.
5. Hyperscale cloud providers themselves, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), support millions of businesses across industries and regions.
6. Big data and analytics companies analyze vast datasets for their distributed clientele.
7. Online gaming platforms, such as Xbox Live, PlayStation Network, and massive multiplayer online games (MMOs) support real-time interactions and gaming experiences.
8. Telcos handle large volumes of data traffic and support resource-intensive and latency-sensitive services, such as video conferencing and VoIP.
9. Banks and trading platforms process enormous volumes of transactions and perform real-time analytics, risk management, and fraud detection.
10. Healthcare and life sciences companies conduct complex bioinformatics research, analyze vast genomic datasets, and support personalized medicine initiatives.

Use cases for hyperscale computing often involve companies and businesses operating on a massive scale.

What’s a Microcloud?

Microcloud typically refers to a limited-scale, hyper-localized cloud infrastructure, often within a specific organization, department, or a limited geographical area. It is strategically designed to meet limited-scope, specific consumer demands, offering a customized solution. Unlike the massive scale of hyperscale clouds, such as AWS, GCP, and Azure, microclouds operate on a smaller scale, providing organizations with greater control over their cloud environment.

The concept of microcloud spans different implementations — it can be implemented within the organization (on-premises) or anywhere else, often close to the organization and/or users. In an on-premise setting, the microscale enterprise data center embodies key cloud characteristics, like virtualization, scalability, and resource pooling, albeit on a hyper-localized scale, which makes it a micro “cloud” as opposed to a traditional enterprise data center. This flexibility in deployment scenarios makes microcloud a versatile solution for organizations seeking tailored cloud services.

Who are Microcloud Providers?

While organizations can build localized cloud solutions using open-source tools or components from larger cloud providers, several alternate cloud providers have emerged to fill in the gap between hyperscalers and organizations requiring tailored cloud solutions. Some well-known microcloud platforms include OpenNebula, which is an open-source cloud computing platform that supports the deployment and management of virtualized infrastructure. It can be used to create private microclouds on a smaller scale, suitable for specific organizational needs. Nutanix also offers hyper-converged infrastructure solutions that can be utilized for building microclouds.

Not all microcloud providers may use the term “microcloud” specifically since the term is relatively new. However, the concept has been around for a while. More recently, in 2023, Canonical, the company behind Ubuntu, launched its “MicroCloud” which is a low-touch, open-source enterprise cloud solution that essentially enables enterprises to launch their own, fully functional microclouds.

Characteristics of Microclouds

Microclouds share similarities with hyperscale clouds in terms of cloud principles, but they differ significantly in scale, scope, and functionality. Below are the defining characteristics of a microcloud:

Limited Scale

Microclouds can vary in size from a Kubernetes cluster of at least three nodes in a private datacenter to a small edge cloud running with no on-site maintenance. Compared to hyperscalers with virtually unlimited, instant scalability, microclouds can scale by adding more hardware and software resources up to their specific capacity.

Localized Infrastructure

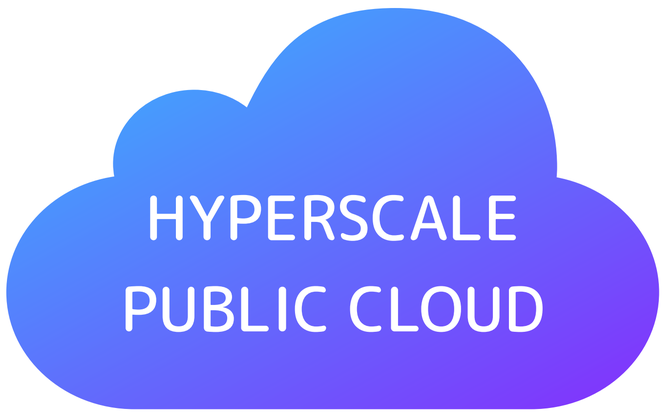
Microclouds can be remote, but they often establish localized infrastructure for specific organization or even a team or department within that organization. Organizations get the advantage of proximity and control, which is particularly beneficial for scenarios requiring low-latency access, data sovereignty, and compliance to specific regulatory requirements.

Tailored for Specific Needs

Organizations can often choose proprietary hardware components that align with their performance, cost, and scalability requirements. Generally, organizations have greater control over the configuration and management of their microcloud infrastructure. They can configure network settings, security policies, and other parameters based on niche requirements without the complexity of a hyperscale infrastructure.

Edge Computing

Sometimes, microclouds can be used to describe infrastructure for on-demand computing at the edge. In this case, microcloud can be a small group of compute nodes with their own storage and secure communication setup, replicated across thousands of edge locations to form a distributed edge environment. The use of standard and open APIs along with hardware abstractions enable edge microclouds to perform with the agility and flexibility of the cloud without its exponential scalability.



- At least, 10,000 sqft. area, 5000 servers, and 40MV of capacity
- Global search
- Virtually unlimited scalability
- Standardized hardware

- As small as a cluster of three nodes
- Hyper-localized
- Scalable up to defined capacity
- Proprietary hardware

Microcloud vs Small Cloud Providers

Microclouds may sometimes be used interchangeably with small-scale public clouds, but both have some nuanced differences. They both aim to provide specialized services, customization, or personalized support that might be challenging for larger providers to offer. However, small-scale clouds may refer to regional or industry-specific cloud providers that support multi-tenancy but serve a limited customer base. Whereas, depending on the configuration and use case, microclouds may or may not support multi-tenancy. Microcloud can refer to an on-premises private cloud tailored for specific organizational requirements, edge computing deployments, or multiple smaller cloud environments within a company.

Why Microcloud Matters?

The concept of microcloud emerged as a response to the growing realization that one size does not fit all when it comes to cloud computing needs. Diverse workloads have different requirements — some can benefit from the scalability and global reach of the hyperscale cloud providers, while others need the speed and performance of localized solutions. The rise of edge computing also influenced the development of microclouds, allowing resources and compute capacity to be placed at the network edge.

Certain industries, such as healthcare, finance, and government, have specific compliance and security requirements, and microcloud provides the level of control and customization that might be challenging to achieve in a larger, more standardized cloud environment. Technologies like virtualization and containerization, coupled with infrastructure automation and distributed storage, empower microclouds to offer cloud-like capabilities while still addressing unique needs.

Use Cases

Microclouds serve as an alternative for scenarios where a smaller but more localized or customizable infrastructure is required and where the scale and complexity of hyperscale cloud environments are not needed.

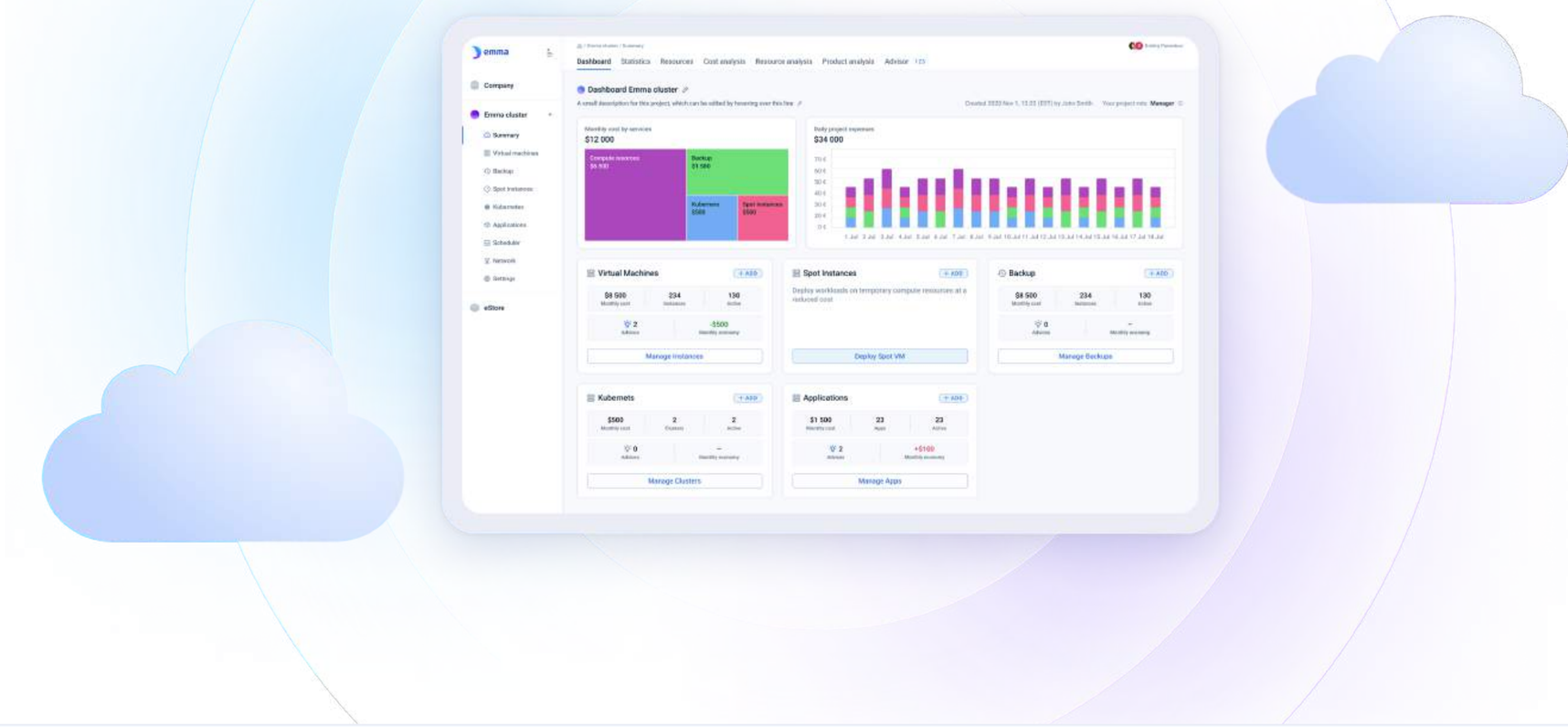
1. Edge computing use cases, such as IoT or real-time analytics, that need faster processing and reduced latency.
2. Organizations may deploy microclouds as on-premise private clouds to meet specific security, compliance, or performance requirements.
3. Organizations supporting flexible or hybrid working models need to host virtual desktop infrastructure (VDI) to allow users to access their desktop environments remotely.
4. SMBs need cost-effective cloud solutions tailored to their scale and without the complexities of managing and optimizing hyperscale cloud.
5. Custom content delivery networks (CDNs) that distribute content from localized servers to improve content delivery speed, reduce latency, and optimize resource utilization.
6. Retail chains need a customized and ultra-low latency platform for managing point-of-sale (POS) systems, inventory tracking, and customer engagement applications.
7. Different development teams or projects within an organization can have their own dedicated environments for resource isolation and reduced risk of contamination.
8. Developers can set up versatile dev environments with specific toolchains, libraries, and dependencies required for their projects and localized staging environments to mimic production environments for testing.

Managing Hyperscale and Alternative Cloud Providers with the emma Platform

The ever-expanding array of cloud computing options compels organizations to make strategic decisions between different kinds of cloud solutions based on their distinct needs, priorities, and objectives. Large enterprises with global operations often gravitate towards hyperscale clouds for their scalability and diverse offerings. Smaller enterprises, research institutions, or those with specific data residency and compliance requirements may choose microclouds for their customizations and control.

However, the choice is not entirely intertwined with scale and operational scope. Many enterprises are consciously choosing not to align exclusively with hyperscale cloud providers due to concerns related to cloud monopolization and vendor lock-in. Letting a few industry giants control the entire organizational infrastructure can raise issues such as data privacy, security, and the potential impact of centralized control on pricing and policies. Consequently, many are now finding a middleground — hybrid and multi-cloud strategies that integrate hyperscale cloud with alternative cloud providers.

This approach affords organization greater autonomy over their infrastructure along with tailored services for their needs. Organizations with diverse workloads, spanning global-scale operations to localized requirements, find this hybrid approach to be a strategic fit. However, orchestrating complex hybrid and multi-cloud architectures involves interoperability and integration challenges as well cost management and optimization issues. As such, these environments require advanced cloud management strategies and solutions.



Introducing the emma Cloud Management Platform

The emma platform is the first end-to-end, no-code cloud management application that enables organizations to unlock all the benefits of multi-cloud (private, public, and edge microcloud) without the typical complexities involved. Organizations can create, manage, and optimize all kinds of cloud infrastructure in their multi-cloud portfolio centrally and seamlessly.

Discover our unique features:

Unified Management Dashboard

A single platform for a centralized view of the entire cloud environment, which enables users to manage all aspects of their cloud infrastructure from a single location.

Cloud Agnostic

A truly vendor agnostic platform allowing organizations to integrate and manage different types of cloud platforms from virtually any provider, all from a centralized application.

No-code Provisioning and Management

A no-code platform enabling users to create, manage, scale, and optimize any cloud infrastructure with just a couple of clicks, no coding required.

Resource Optimization

Enabling real-time monitoring of resource utilization across all cloud platforms in a multi-cloud set-up for effective resource optimization, even in resource-constrained microclouds.

Cost Management

Providing advanced AI-based insights for finding least expensive spot instances and real-time offers across all integrated cloud platforms.

Multi-cloud Interconnectivity

Deploying a high-performing and reliable multi-cloud networking backbone to migrate instances and scale workloads seamlessly across multiple cloud providers.

Through its comprehensive features and capabilities, the emma platform empowers organizations to deploy applications in the most suitable cloud environment. It facilitates seamless integration of different cloud paradigms — hyperscale, small-scale, micro and edge — allowing organizations to choose strategically between scale and global reach, localization and customization, and everything in between!