servers.com

A roadmap out of hyperscale cloud

How infrastructure as a service (IaaS) is helping optimize cloud costs in tough market conditions

Synopsis

This whitepaper explores the increasingly cost-prohibitive nature of hyperscale cloud environments and the various hyperscale infrastructure alternatives available today. Supported by insights from thought leaders interviewed exclusively by servers.com, this whitepaper is designed to help readers assess their own infrastructure and make steps towards achieving long-term cost and performance optimization.
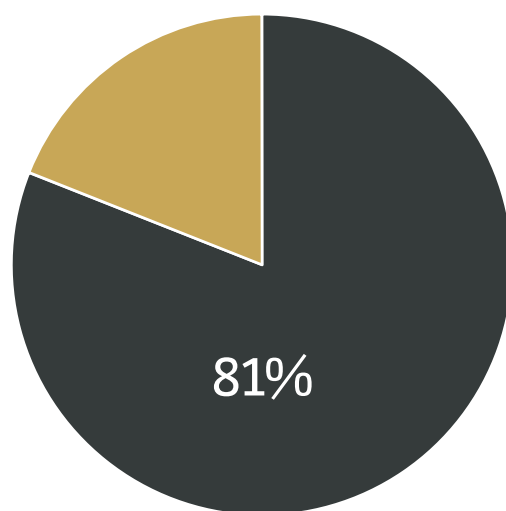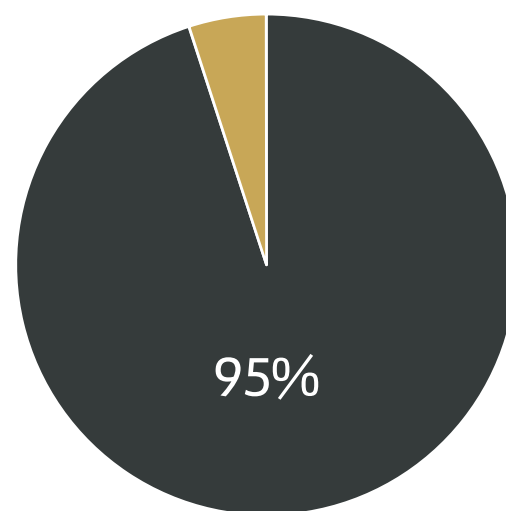
Contents

Introduction — Is the cloud bubble bursting?

The reign of hyperscale cloud is starting to crumble as companies become more conscious of their infrastructure spend. 81% of IT leaders[1] report that their c-suite has told them to reduce or take on no additional cloud spending. 95% want[2] increased visibility and control over their cloud costs and 80% are planning[3] to increase investment in cloud FinOps (the management and financial optimization of cloud infrastructure).
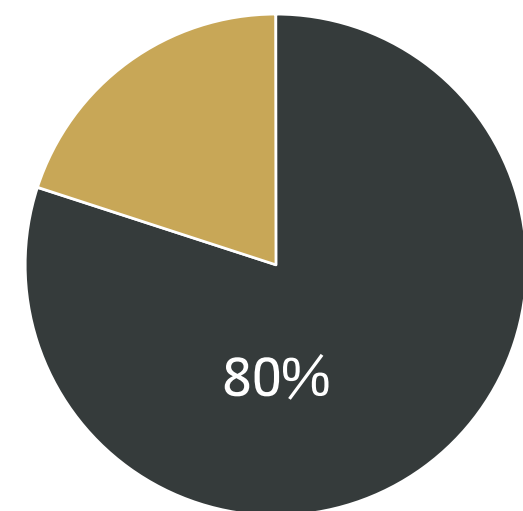
81% instructed to reduce or take on no additional cloud spending

95% want increased visibility and control over cloud costs

80% planning to increase investment in cloud FinOps

81%

95%

80%

The push to reduce costs is so prevalent that hyperscale cloud providers are already starting to feel the squeeze. In April 2023, AWS reported revenue growth[4] of 16% compared to 20% in the fourth quarter of 2022.

"Hosting on the cloud can be very expensive," explains Liam Brennan, Director of Games Technology and Services at Tencent. *"They know that they need to be cheaper to win new customers... I know of major cloud vendors giving up to 70% discounts on their list price and 50% discounts on their egress costs".*[5]

It all points to an overarching trend. Hyperscale cloud is proving too expensive for many. But to understand how it got to this point, we need to look back to the 1990s and where it all began.

A brief history of the internet and the birth of hyperscale cloud

The birth of hyperscale cloud is fundamentally tied to the birth of the internet. And so, to understand where it came from, it helps to look back to the origins of the internet itself. And that story starts in the 1960s.

1960s

The very first version of the internet was funded and developed by The Advanced Research Projects Agency (ARPA), an arm of the U.S. Defense Department. In 1968, ARPA contracted BBN Technologies to build the first routers enabling ARPANET (the Advanced Research Projects Agency Network)[6] to become operational. ARPANET was a revolutionary network capable of sharing digital resources between geographically separate computers. In 1969, the technology was demonstrated for the very first time and by 1972 the first public demonstration took place. That same year, Ray Tomlinson wrote the first email program.

1970s

By the late 1970s, Vinton Cerf and Robert Kahn[7] created the architecture for the Internet, allowing multiple networks and their computers to communicate with each other.
To implement the architecture, the pair designed a software known as the Transmission Control Protocol (TCP)/ Internet Protocol (IP) — the networking standard used to transmit data over the internet. By 1983, TCP/IP became the standard host protocol for ARPANET.

Did you know?   In 2004, Vinton Cerf and Robert Kahn won the Turing Award for their pioneering work on the Internet.

1980s

As computer and internet uptake soared, bookkeeping started to become a major problem. The first attempted solution involved manually assigning addresses to websites. Hostnames had to be added to a HOSTS.TXT file by contacting the SRI Network Information Center by telephone. But as the network grew, this became unsustainable. The world needed a simpler way to network.

Enter Paul Mockapetris and his team at UC Berkley who were tasked with coming up with a new system. In 1983 they created the Domain Name System (DNS)[8] and by 1984, the team developed the first Unix name server implementation which came to be known as BIND. Since then, DNS has been through numerous upgrades. Today DNS operates on a decentralized structure of multiple interconnected servers that collectively store DNS records.

1990s

In 1989, a British scientist by the name of Tim Berners-Lee invented the World Wide Web[9] (WWW) whilst working at CERN (the European Organization for Nuclear Research). Berners-Lee was working on a way to automate information sharing between scientists. In 1990, he wrote the first web client and server by creating the Hypertext Transfer Protocol (HTTP) and URLs — both of which are still used today. By 1993, CERN placed the World Wide Web in the public domain and shortly after gave it an open license. The following year, Berners-Lee founded the World Wide Web Consortium (W3C) to develop interoperable technologies for the furtherment of the Web and by 1995, commercial internet access was becoming available to consumers. With an estimated 16 million users[10] at the time, the scene was set for a boom in new internet-driven businesses.

Number of internet users (millions) 1995-2022

Year	1995	1996	1997	1998	1999	2000	2001
Users	16	36	70	147	248	361	513

Year	2002	2003	2004	2005	2006	2007	2008
Users	587	719	817	1,018	1,093	1,319	1,574

Year	2009	2010	2011	2012	2013	2014	2015
Users	1,802	1,971	2,267	2,497	2,802	3,079	3,366

Year	2016	2017	2018	2019	2020	2021	2022
Users	3,696	4,156	4,313	4,536	5,053	5,252	5,544

Source: Internet Growth Statistics 1995-2022 [11]

2000s

In the world of business, companies' servers were located in a designated server room or basement. These servers were usually IBM mainframes and desktop computers that had to be connected by a local area network (LAN). This meant that everything had to be accessed on premises.

As the internet continued to boom, more businesses started setting up their own websites. These websites needed hosting, and at the time that meant one of three things. Hosting your own servers on-premises, via a third-party infrastructure provider, or by using one of the newly emerging web hosting companies. The basic concept of cloud (using somebody else's computer) was starting to emerge.

But then things got a bit more complicated. As businesses started creating a whole range of new services (from ecommerce stores to digital content streaming services to multi-player gaming platforms) these platforms needed infrastructure. Colocation met this need head on, allowing businesses to buy servers and place them in a third-party data center.

Next came the move from CAPEX to OPEX. Businesses could do away with owned infrastructure altogether and rent servers over the internet. For the first time you didn't need the capital to own your own infrastructure to get your businesses online.

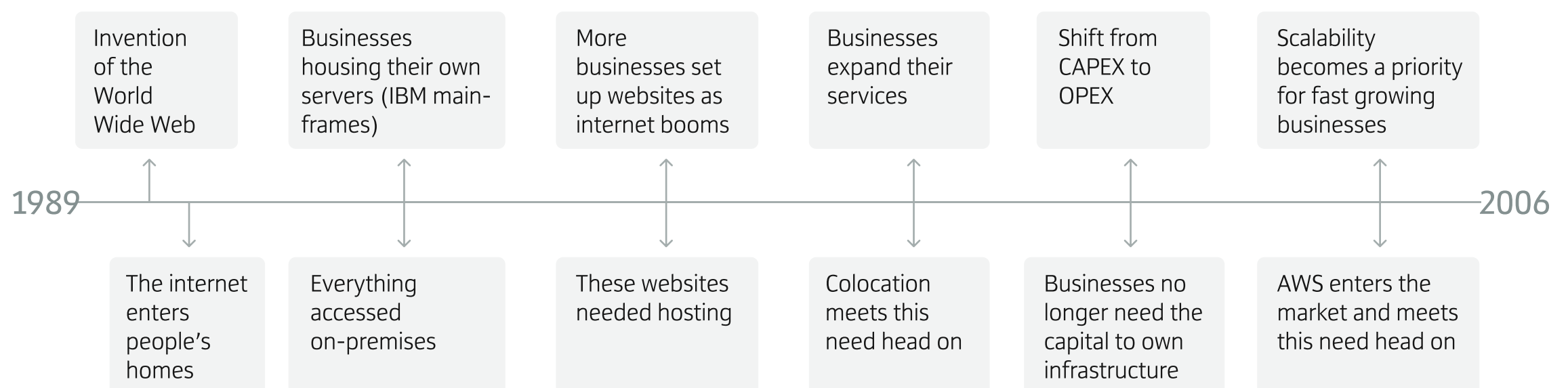| **Capital expenditure (CAPEX)**	**Operational Expenditure (OPEX)**
Money spent on fixed assets	Money spent on an ongoing basis to run a business
buy	rent
For example: buying servers	For example: renting servers from a hosting provider

OPEX quickly became the infrastructure model of choice for startups. They could now get their businesses online with just a couple of hundred dollars. But it wasn't all smooth sailing. Hosting options at the time had their limitations. Early hosting lacked flexibility whilst long term contracts and slow delivery times for infrastructure meant scaling was a slow process.

One of the fastest growing ecommerce stores at the time, Amazon, knew it. Amazon had already invested in developing tools and server architectures capable of scaling up and down in line with massive peaks in demand caused by seasonal holidays and ecommerce events like Black Friday (a key date in retail calendars since the 1980s and long before ecommerce). As a next step, Amazon decided to monetize these capabilities and sell them to other companies. Amazon Web Services (AWS) was born. Enter billing by the hour (these days by the second), and the ability to spin up new virtual machines in minutes.

AWS was the first hyperscale public cloud provider, launching its first infrastructure service — Simple Queue Service (SQS) - in 2006[12]. No one else was offering these services at the time so AWS gained a huge market share and built out a vast range of services. And with AWS operating under an IaaS (Infrastructure as a Service) model, businesses no longer had to think about compute or storage. Today, cloud is synonymous with AWS, with other hyperscale cloud providers Google Cloud Platform (GCP) and Microsoft Azure not far behind.

Invention of the World Wide Web	Businesses housing their own servers (IBM main-frames)	More businesses set up websites as internet booms	Businesses expand their services	Shift from CAPEX to OPEX	Scalability becomes a priority for fast growing businesses
1989					**2006**
The internet enters people's homes	Everything accessed on-premises	These websites needed hosting	Colocation meets this need head on	Businesses no longer need the capital to own infrastructure	AWS enters the market and meets this need head on

The expectation vs reality of hyperscale cloud

The birth story of hyperscale cloud providers the likes of AWS, GCP, and Azure is impressive. But as this business model matured, cracks started to appear. And, increasingly, hyperscale cloud customers have found themselves struggling with a discrepancy between an expectation of cloud and the reality of what it delivers.

The promise		**The reality**	
Efficiency	Simplicity	Unmanageable complexity	Prohibitive costs
Lower infrastructure costs	Scalability	Platform lock in	Lack of support

Take AWS, for instance. As its services have expanded over the years, AWS environments have become increasingly difficult for customers to manage. Navigating the array of products and services on offer, and identifying the most cost-effective solutions, is no simple task.

"One of the prevailing concerns with cloud solutions is the unpredictability and non-transparency of costs".

Ievgen Nagorny, Chief Technology Officer, MGID

There are even AWS courses dedicated to learning how to best manage an AWS environment. And some customers end up hiring AWS certified professionals to help them make the best buying decisions. Monthly invoices are accompanied by vast Excel documents that the average human cannot be expected to comprehend. And it requires a specialized internal tool to translate these bills into something that a human being can understand.

"One of the prevailing concerns with cloud solutions is the unpredictability and non-transparency of costs," comments Ievgen Nagorny, Chief Technology Officer at MGID. *"The fact that there are numerous companies providing cloud cost optimization services is testament to this problem. Businesses often find themselves paying for hidden costs or unnecessary resources, making budget management challenging"*[13].

The level of support provided by hyperscale cloud providers is negligible and break-fix only support is standard. It can take weeks to get a response after raising a ticket and often queries are closed with nothing more substantial than a link to a knowledge base article. Better service is available, but at an additional charge.

"You rarely get proper attention for things that matter to your business but are not critical for your cloud provider", explains the Global Director of Technology at an iGaming platform, who wished to remain anonymous. *"For example, a bandwidth interconnection between two regions is not important for them [the cloud provider] just because it's causing some delay. Even if this delay is causing you a loss of revenue, you are not going to get traction on understanding what happened and get the fix (or even a hot fix) for the problem"*[14].

Hyperscale cloud is perceived as the simple and economical choice, but the reality is far more complex. Many businesses are drawn into hyperscale cloud environments by the promise of simplicity and free credits only to find themselves locked into financially unsustainable contracts by the time those free credits run out. *"Cloud vendors tend to try and lock you into their ecosystem. If you do get locked in, then that's great for them, but it's not great for you. You're just going to be limited in how you can scale"*[15], comments Andrew Walker, Head of Business Development at Gameye.

Compute isn't infinite for any provider. Even AWS runs out of compute sometimes, so being able to access different sources of compute is important for scalability and redundancy. But for companies locked into these services, migrating away (even in part) can be exceptionally difficult. As Nagorny explains *"as businesses integrate deeply with a particular cloud provider's services, migrating to another platform can be costly and time-consuming"*[16].

This is something that Gameye found themselves all-too familiar with when working with game developers. *"When it comes to vendor lock-in, we've had developers come to us going 'we just ran out of credits, and our bill went through the roof. How can you help us keep this down?'"*[17], shares Walker.

This issue is exacerbated further by the addition of data egress fees that many hyperscale cloud providers charge their customers to move data out of the cloud. The cost of these fees can vary depending on the volume of data, its origin, and intended destination. AWS, for instance, charges[18] $0.09 per gigabyte of data for the first 10 terabytes and $0.02 per gigabyte for transferring data between EC2 instances in AWS in different regions.

Recent OFCOM investigations into the UK public cloud market also found[19] egress fees to be significantly higher amongst the major players, posing significant barriers to customers wishing to migrate their data. And whilst there is movement on this issue (as of January 2024 Google Cloud became the first major cloud provider to eliminate egress fees for customers choosing to migrate data away from the Google Cloud Platform) it's yet to be seen whether egress fees will be made reasonable across the board.

These obstacles to migration severely limit an organization's access to alternative compute sources. The best approach is to avoid lock-in in the first place.

Why you need to be realistic about your scaling needs

Hyperscale cloud is a great option for some businesses. For large-scale enterprise organizations that regularly experience huge peaks and troughs in demand, auto scaling (an AWS feature that continuously monitors an application and automatically increases resource capacity when demand spikes) is a significant advantage.

Businesses like these need to be able to spin up additional virtual machines near instantaneously in accordance with daily usage curves. And that's exactly what hyperscale cloud does best.
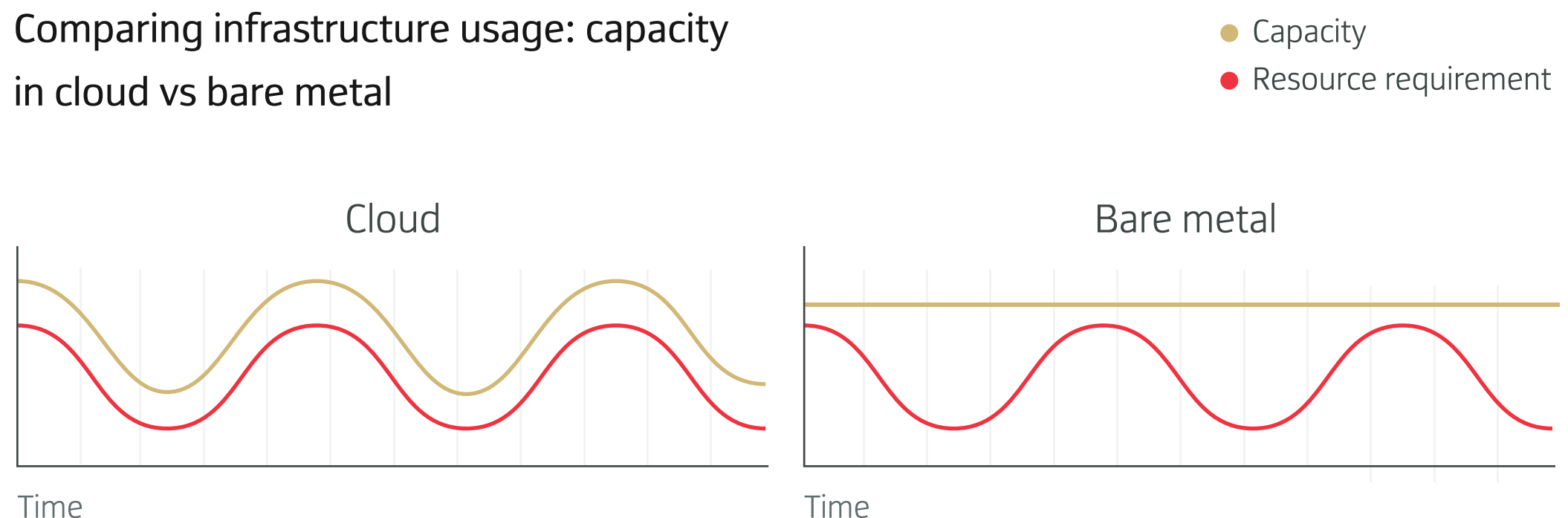
The issues arise when businesses that do not have these requirements still think they need hyperscale cloud. And misconceptions around scalability are to blame. *"They see a freedom there – unlimited resources"*[20], explains the Global Director of Technology at an iGaming platform. But it can be difficult to understand whether that freedom is *"fit for purpose"*.

Every business has small peaks and troughs in daily usage, but the majority require incremental scaling at most.

So, whilst the cloud sells itself on the ability to scale in a near-instant manner whilst optimizing spend, for anyone who isn't experiencing volatile demand, it's probably not worth the premium.

And when you consider that the cost of compute can be 10x compared to bare metal servers, added to the complexity of managing auto-scaling, that's a lot of money to invest in something that wasn't needed in the first place.

Comparing infrastructure usage: capacity
in cloud vs bare metal

- Capacity
- Resource requirement

Cloud

Bare metal

Time

Time

Industry trailblazers — 37signals

37signals David Heinemeier Hansson, Co-owner and CTO of 37signals has publicly documented the company's decision to migrate away from cloud in his blog 'Why we're leaving the cloud'[21] (published in October 2022).

In writing, Hansson describes a situation whereby *"the savings promised in reduced complexity never materialized"* and compares being in the cloud as *"like paying a quarter of your house's value for earthquake insurance when you don't live anywhere near a fault line"*.

In other words, unless your application is extremely low traffic or experiences highly irregular swings in usage (two conditions under which cloud excels), there is absolutely no cause for investing in cloud. And doing so means paying in excess for features your business simply does not need. That's exactly the situation 37signals found themselves in. Paying an *"absurd premium"* for the possibility that the business might be cloud compatible when 37signals has and will continue to be more compatible with owning hardware.

In his follow-up blog, "We have left the cloud"[22], (published June 2023) Hansson reveals *"after six months of effort, it's done. We're out."* The company that last year had a collective cloud budget of $3.2 million is now saving at least $1.5 million per year by owning its own hardware rather than renting it from Amazon.

So, why do so many organizations fall into this trap? For many there's a temptation to use hyperscale infrastructure, like AWS, as a kind of insurance policy. It's familiar, reliable, and many startups simply don't have an in-depth knowledge of the various alternative infrastructure hosting options out there. And as Brennan explains, for those that do, *"many still perceive bare metal as the inflexible option"*. And more fundamentally, *"the majority of indie studios don't know about the various bare metal providers and what they offer for gaming. They know about AWS so they just sign up for that"*[23].

"The majority of indie studios don't know about the various bare metal providers and what they offer for gaming".

Liam Brennan, Director, Games Technology and Services, TENCENT

The problem isn't that cloud is bad and bare metal is good or that cloud is good and bare metal is bad. It's simply that there exists a chronic lack of understanding around how to optimize infrastructure for cost and scalability. And where there is the understanding, then it's typically a lack of time that gets in the way.

"Most people don't know how to cost optimize from an infrastructure perspective", furthers Brennan. And in the gaming industry *"the server engineers' first priority is getting the game to scale, then cost-optimized"*[24].

How to balance cost and scalability

Understanding the type of scaling events your business experiences is fundamental to optimizing your infrastructure stack and finding a sustainable balance between cost and scalability. It's true to say that this can be challenging. But whilst nobody has a crystal ball to predict scaling events with 100% accuracy, there are ways to determine your general scaling patterns. And the first step is to ask the following questions:

1. What causes scaling events in my business?
2. How often do scaling events occur in my business?
3. What causes downturns in demand in my business?
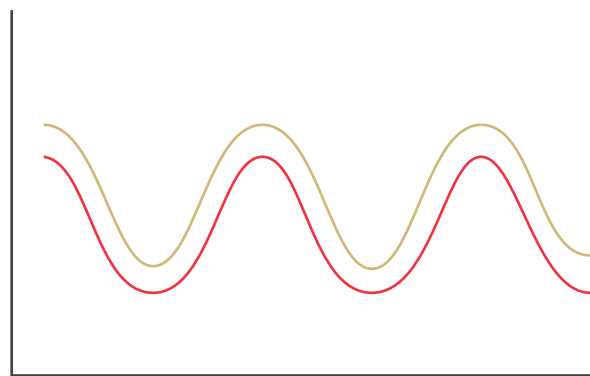4. How often do downturns in demand occur in my business?

Then identify which pattern best aligns with your organization's scaling events.

The diagram below illustrates three common scaling patterns and the best infrastructure solutions for each.

3 common scaling patterns

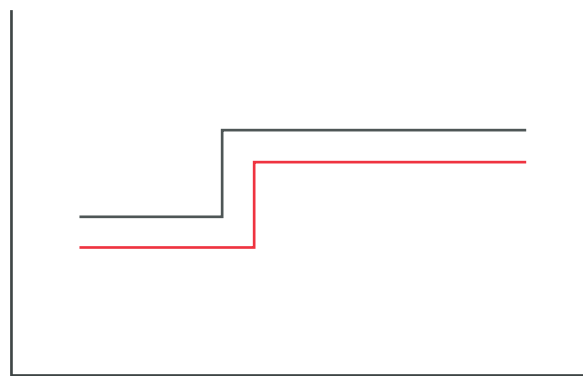● Demand　　● Capacity (cloud)　　● Capacity (bare metal)

Erratic scaling events	**Predictable scaling events**	**Incremental scaling events**
Your business experiences big highs and lows in demand.	Your business experiences predictable increases in demand that can be planned for in advance.	Your business is growing but demand is not volatile.
You need an IaaS solution that can be scaled up and down near-instantly.	You need an IaaS solution that can be scaled up and down periodically.	You need an IaaS solution that can be scaled up and down incrementally.
Best choice: **cloud**	Best choice: **bare metal**	Best choice: **bare metal / combined cloud and bare metal**

"The cloud is good when you don't want to think about data structure and you can solve all data or computation issues by paying more".

Ilia Varvaluk, Chief Technology Officer, Bidease

Of course, not everyone is looking to cost-optimize. For some, unlimited scalability will always be the top priority. And this is precisely where cloud comes into its own. According to the Global Director of Technology at an iGaming platform, there are three key scenarios when a cloud-first approach makes sense: *"If you are a consumer or producer of unlimited data and you don't know what your dataset is going to look like a year from now. If you need to host thousands of terabytes without caring too much about costs [or] if you just need to spin up massive infrastructure with massive unlimited potential"*[25].

For others it's convenience that wins out. *"The cloud is good when you don't want to think about data structure and you can solve all data or computation issues by paying more"*[26], comments Ilia Varvaluk, Chief Technology Officer at Bidease.

Understanding the hyperscale alternatives

Infrastructure tends to be split into two camps. On-premises infrastructure on the one hand and hyperscale cloud on the other. But the options are not as binary as it might seem. For businesses struggling in hyperscale cloud and looking to migrate to an alternative, there are several options.
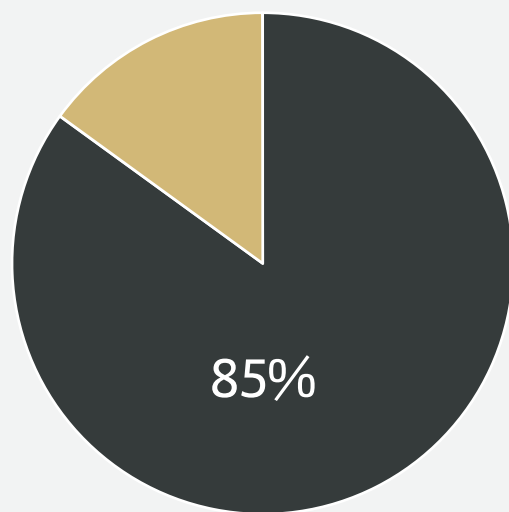
Option A: An on-premises data center

An on-premises data center consists of servers and computing infrastructure located and managed on a business' premises and using that business' native computing resources.

On-premises infrastructure comes with many benefits including:

- Continuous access to business data

- Full control over data storage

- The ability to track the location and movement of data

- Reliable and consistent connections

- Reduced latency and network connectivity

- Full control over any security protocols deployed

85% of enterprises will
shut down their traditional
data centers by 2025

85%

However, for many businesses today, running an on-premises data center is not considered a primary business activity.

In fact, 85% of enterprises[27] will shut down their traditional data centers by 2025. The capital outlay required to set-up and run an on-prem solution is substantial as is the labor required to maintain it.

Businesses need to factor in the cost of hardware (servers, racks, routers, switches, firewall, and load balancers), multiple contracts with internet service providers (ISPs), server maintenance contracts, hiring network engineers, system administrators, procurement managers, and legal personnel.

Then there's insurance (which can be up to 50% of the total cost of an on-prem solution), a ticketing system, and a service level agreement (SLA).

Prior to the hyperscale cloud boom, an on-prem solution was the standard method for storing data and running applications. And for some businesses today (37signals being a prime example) on-premises infrastructure might still be the best option.

On-prem is a good choice for:

- Businesses only using a few servers located in the same building
- Businesses using connected tools requiring local server hardware (this is common of manufacturing facilities and labs)
- Businesses that fall under data protection laws dictating that application and data storage must be on-premises
- Businesses with the financial capital and expertise to run a data center in house and for whom the additional performance benefits justify the amount spent

"The downside of having your own infrastructure is you need to have your own team of experts managing that".

Global Director of Technology, iGaming platform

However, for many businesses already struggling in hyperscale cloud an on-prem solution is going to be just as (if not more) expensive and complex to manage. As the Global Director of Technology at an iGaming platform explains, *"the downside of having your own infrastructure is you need to have your own team of experts managing that [and] good resources, we know, are not easily found in the marketplace"*. However, as they counter, in some cases it's still worth it. *"For an organization that values IT as an asset to deliver software that creates revenue, that's easier to achieve [on-prem] than achieving the same results without having control of the different aspects of technology"*[28].

Organizations without access to the capital or resources required to manage infrastructure in-house will have their needs better met through colocation or bare metal server hosting.

Option B: Colocation

Colocation allows a business to own its own servers but store them in rented rack space within a colocation hosting facility (an off-site data center).
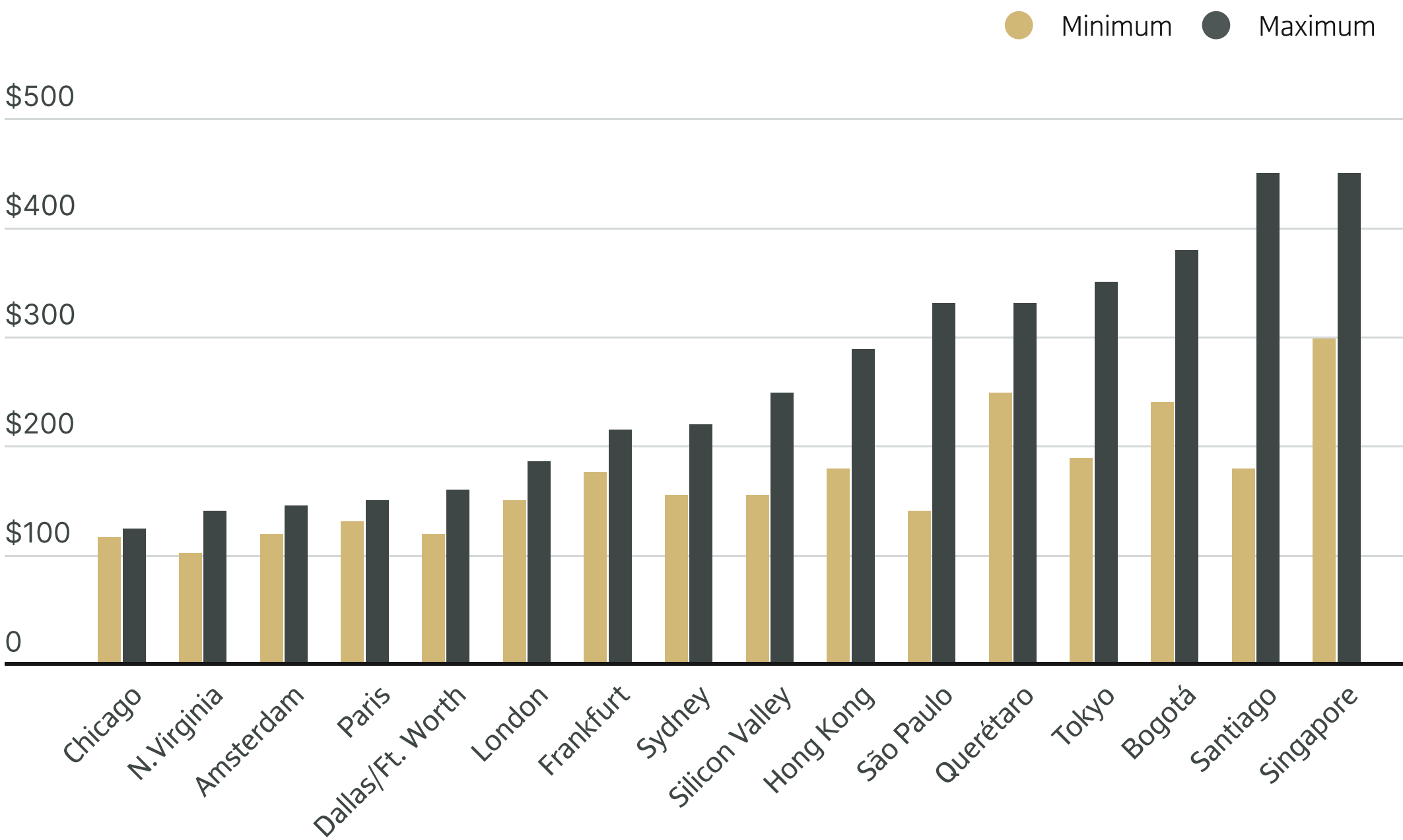
Colocation comes with benefits such as:

- It's more cost-effective than hosting a server on-premises
- Servers are stored in secure locations
- Colocation hosting facilities maintain optimum room temperatures
- Servers do not have to be moved if a business relocates
- Businesses maintain full control over their servers
- Excellent bandwidth availability
- Redundant power and internet

But even though colocation is usually cheaper than on-prem, it's still capital and resource intensive. Aside from handing over the management of rack space, businesses in colocation are still required to manage everything else they would in an on-prem solution. So, colocation doesn't offer much in the way of reduced complexity.

Before choosing colocation, businesses need to factor in:

- High initial start-up costs

- Purchasing equipment and updates

- Fluctuating monthly colocation costs due to bandwidth use

- Finding a colocation facility near business premises

- Server depreciation (typically between 3-5 years)

Monthly colocation rental rate range for 250-500kW without electricity cost (USD)

Source: CBRE Research, Q1 2022 & Q1 2023 [29]

For small-to-medium sized businesses, colocation can be a great solution. They benefit from more bandwidth per pound spent than in an on-prem equivalent. And there's also the added support of a colocation hosting provider that will manage server maintenance and updates on the business' behalf.

However, for businesses already struggling in hyperscale cloud, colocation isn't going to offer a significant reduction in spend or complexity. As a business grows it will need the capital to continue buying new assets whilst funding inactive servers sitting in reserve. This is expensive and even big tech has started extending server lifespans[30] in a bid to reduce costs.

For both colocation and on-prem solutions there's also the possibility of relocation to consider. Anything from regulatory change to unplanned incidents can force businesses to relocate. An eventuality that promises to be extremely costly for anyone physically managing their own infrastructure.

Option C: Bare metal servers

The third option is bare metal-as-a-service: renting dedicated servers from a bare metal hosting provider. For businesses using hyperscale cloud, but without the financial capital and resources necessary for a full reverse cloud migration to either on-prem or colocation, bare metal servers are a fantastic option — and can also be accessed as an IaaS solution.
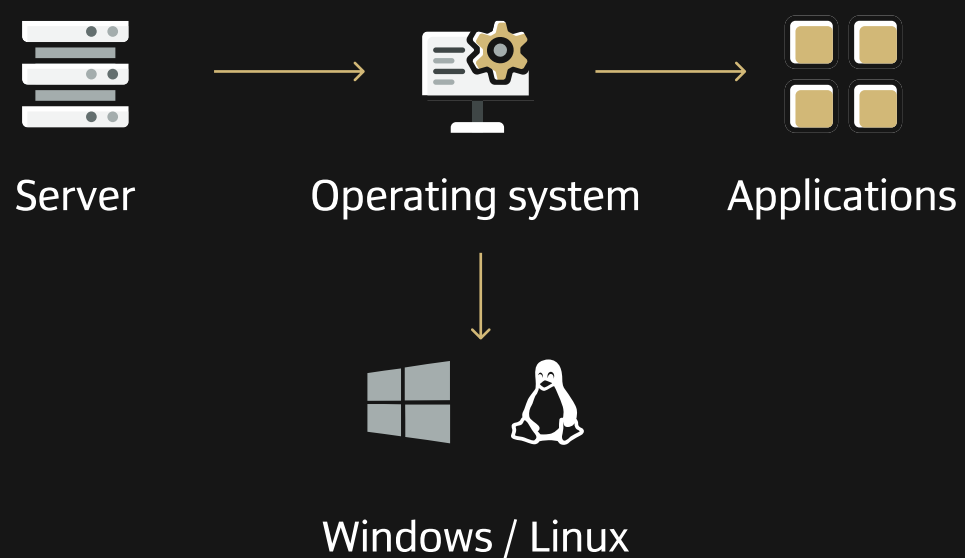
If we return to the fundamental definition of cloud computing — using somebody else's computer — bare metal hosting is essentially that. Businesses lease somebody else's machines but have sole tenancy over the server and its resources and typically enter mid-to-long-term contracts.
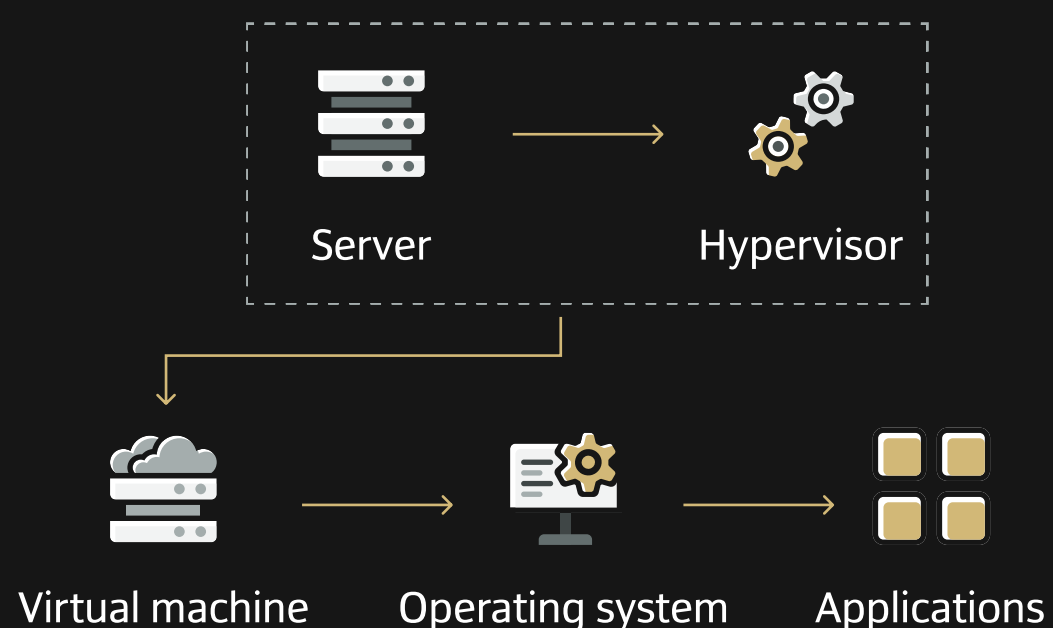
IaaS explained

IaaS (Infrastructure as a Service) is often associated with hyperscale computing services like AWS and GCP. But the definition of IaaS is simply leasing computing resources from a third-party supplier. So bare metal hosting also falls under an IaaS model. Only, unlike cloud which provides virtualized computing resources over the internet, bare metal IaaS hosting providers lease out single tenant dedicated server infrastructure solutions instead.

Bare metal

Server → Operating system → Applications

Windows / Linux

A **bare metal server** is a physical server that sits within a data center or rack. These are single tenant servers that serve just one client. The operating system (OS) is installed directly on the server and standardizes access to the compute resources of that server.

Cloud

Server → Hypervisor

Virtual machine → Operating system → Applications

A **cloud server** is a bare metal server that has been virtualized. The OS does not sit directly on the server. Instead, cloud servers have a hypervisor which creates multiple virtual machines on one server. The hypervisor splits up the server resources between the virtual machines. Each virtual machine has an operating system placed on top of it with applications on top of the OS.

As Brennan frames it, *"it's all machines inside of data centers. The key difference between bare metal and cloud is the business model. The business model for cloud is to charge for usage on a per-second basis and bare metal is to rent out predictably and long term"*[31]. With bare metal-as-a-service, businesses don't have to take on the capital outlay and resource intensiveness that comes with on-prem and colocation. Nor the cost premiums and complexities that have become synonymous with hyperscale cloud providers.

"Bare metal is becoming more flexible with hourly pricing models and hourly cancellation policies".

Liam Brennan, Director, Games Technology and Services, TENCENT
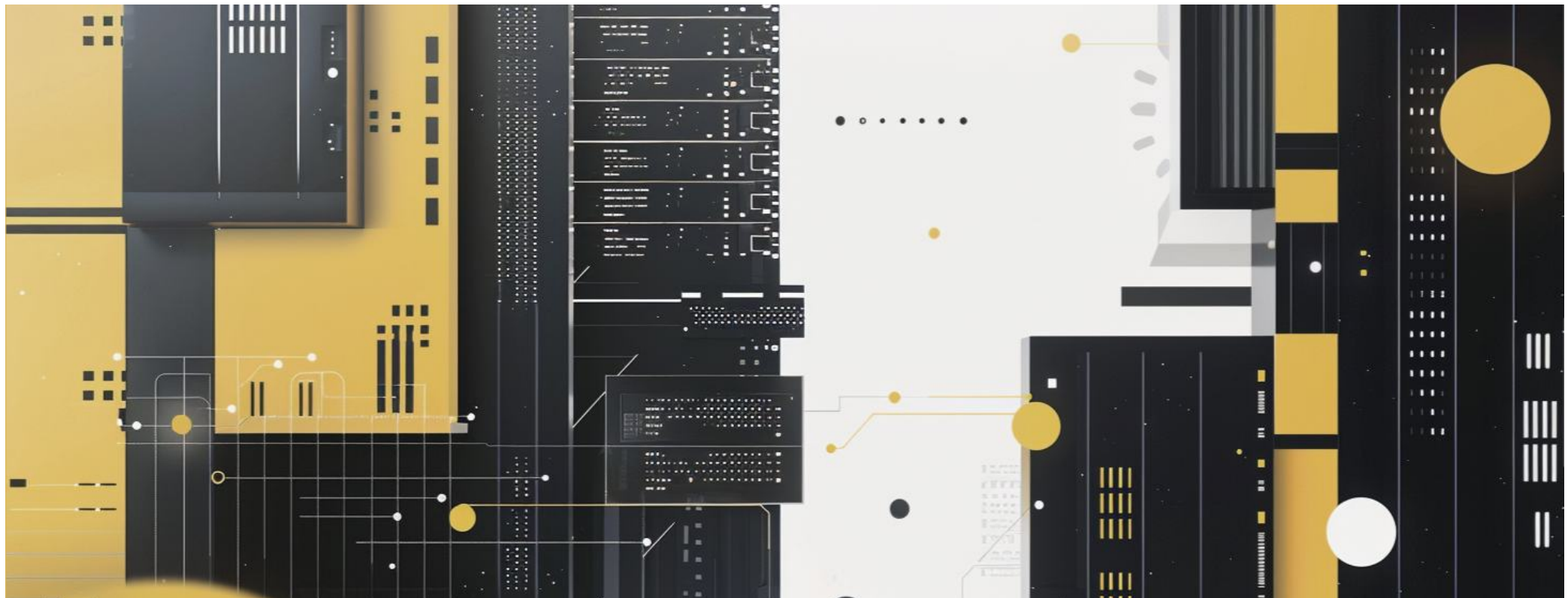

The benefits of bare metal

Bare metal server hosting offers a middle ground between two infrastructure extremes. Combining the scalability and performance of cloud with the control and customization of bare metal machines. For any business struggling in hyperscale cloud, shifting to bare metal or adopting a hybrid approach combining bare metal with cloud, on-prem, or colocation can help to strike a balance between cost, complexity, and performance.

Benefits of bare metal include:

Cloud like scalability

Businesses often assume that it's difficult to scale bare metal servers. *"It all started with cloud (which was very expensive but very flexible) and bare metal (which was very inexpensive but very inflexible)"*, explains Brennan. But scalability in bare metal has improved tenfold in recent years. *"Now these two things are coming closer together...cloud vendors are struggling to some degree because they are very expensive. They know they need to be cheaper to win new customers... bare metal is becoming more flexible with hourly pricing models and hourly cancellation policies"*[32], furthers Brennan.

These days provisioning additional servers in line with incremental scaling is quick and straightforward. The physical deployment of servers is the responsibility of the hosting provider and servers can be deployed according to a business's unique scaling patterns.

High performance

With bare metal hosting, servers are dedicated to individual tenants. This is different from hyperscale infrastructure in which server resources are split between multiple tenants and where, instead of renting sole access to a single server, businesses rent access to virtual machines (VMs). This sharing of resources can result in a 'noisy neighbor' effect where one tenant's virtual machine uses up most of the available resources, causing network performance issues for other tenants.

"With cloud, you're always dealing with other larger parties that have stolen the cloud capacity from you"[33]*,* comments Walker. *"We've had reports of noisy neighbor problems on cloud where badly behaved processes in multi-tenant situations are impairing the game session that's running elsewhere".*

It's a common problem for cloud users. Something that MGID also experienced first-hand. *"With cloud solutions we might encounter the 'noisy neighbor' effect where other users on the same infrastructure can affect our performance...dedicated server solutions, however, provide a predictable and consistent performance environment,"*[34] shares Nagorny.

"With cloud, you're always dealing with other larger parties that have stolen the cloud capacity from you".

Andrew Walker, Head of Business Development, Gameye

Because server resources aren't split between multiple businesses, bare metal is ideal for intensive workloads and offers reduced lag, high levels of server uptime, low latency, and fast application loading speeds and response rates.

Tight security

Bare metal hosting providers maintain the highest data center security standards. Because dedicated servers are leased out on a single-tenant basis, they're more secure than shared environments. *"By using dedicated servers, we have a clearer understanding of where our data is stored and managed"*[35], comments Nagorny. *"The risks associated with multi-tenancy in the cloud are alleviated, thus ensuring a more secure environment for our data".*

That said, it's important to remember that no hosting provider can protect against human negligence so it's also essential that businesses maintain best practices internally.

*"Dedicated server solutions offer more customizable options
when it comes to hardware and software configurations".*

Ievgen Nagorny, Chief Technology Officer, MGID

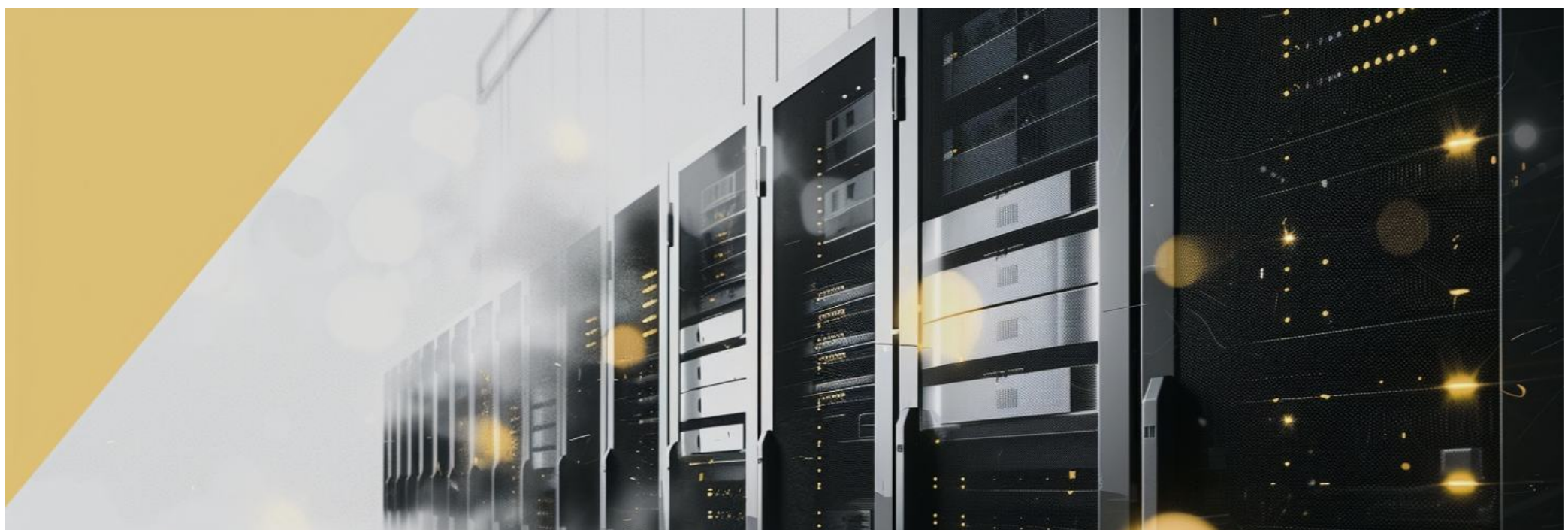
Control and customization

With bare metal servers, businesses can create a bespoke infrastructure solution. *"Dedicated server solutions offer more customizable options when it comes to hardware and software configurations"*[36]*,* explains Nagorny. *"This means MGID can optimize the server environment to our exact needs without being constrained by the limitations of standardized cloud offerings".*

There's no virtualization process and only one tenant per server, so each business gets full administrative control over its machines. Some components will be determined by the hosting provider (server model and CPU), but everything else (disk storage, operating system, software and network connectivity, RAM, security software) comes custom.

Data center locations

Contrary to popular belief, cloud is not everywhere. *"Believe it or not, it's not as present as you might imagine. AWS and GCP don't really have a PoP in Eastern Europe"*[37]*,* comments Walker. And when it comes to mitigating latency, location is everything. By working with one, or multiple, bare metal hosting providers it's much easier to deploy servers across various data center locations as close to end users as possible.

Made with AI

Reduced costs and complexities

With bare metal servers, businesses don't have to manage server maintenance. The hosting provider will do everything from provisioning, to maintenance, to updating the hardware. That means the complexities and expenditure associated with on-prem and colocation are much reduced. And, unlike hyperscale cloud providers, many bare metal hosting providers operate off simple month-to-month contracts with minimal administrative complexities.

"Through our evaluations, we've found that cloud solutions can be 2-3 times more expensive than dedicated server solutions over extended periods"[38], comments Nagorny. *"Although cloud solutions come with the allure of scalability and pay-as-you-go models, over time and with consistent heavy workloads, the costs can spiral. Dedicated server solutions...offer predictable pricing models, which in the long run, proves to be far more economical"*.

"Cloud solutions can be 2-3 times more expensive than dedicated server solutions over extended periods".

Ievgen Nagorny, Chief Technology Officer, MGID

Support that actually supports

As hyperscale cloud providers have grown, the level of support being offered has diminished. Lack of support when things go wrong is one of the biggest challenges businesses face in hyperscale cloud. A good bare metal hosting provider, on the other hand, will work with each customer to optimize the cost and performance of its infrastructure in line with business goals and help its clients find solutions when things go wrong. So, there's more control over your technology.

Quick-fire infrastructure comparison

	Architecture	Advantages	Disadvantages	Best if...
On-prem	Servers housed within a business' premises in a dedicated server room	Full control over servers Full control over server security Reduced latency Full access to your infrastructure in the event of a technical fault or outage	Steep upfront and ongoing investment Specialist expertise required to set up and maintain	You only need a few servers and most of your users are located in the same building Your business can withstand a small amount of downtime Your business uses connected tools requiring local server hardware Data protection laws dictate that application and data storage must be on-premises
Colocation	Businesses own servers but place them in a third-party data center	Full control over your servers Redundant power and internet No need to manage rack space	Steep upfront and ongoing investment (though marginally cheaper than on-prem) Specialist expertise required to set up and maintain	You need full control over your servers but don't have the capital or resources to run your own data center
Bare metal	Dedicated servers leased out by a hosting provider	Incremental scaling as and when needed Sole access to your server and its resources Easy to customize Support from your hosting provider Cost-effective (your hosting provider fronts the cost of running and maintaining the data center)	No auto-scaling Some specifications will be determined by your hosting provider	You need the convenience of a managed service but don't want to operate in a shared environment You have incremental scaling requirements You need cloud-like scalability without the hyperscale pricing You want infrastructure that can be customized to your unique requirements
Hyperscale cloud	Bare metal servers that have been virtualized (virtual machines are leased out by a hosting provider)	Near-instant scalability Follow your daily usage curve with cloud auto-scaling Near-instant provisioning Easy to set up	Extremely complex to optimize for cost and performance Expensive Risk of vendor lock-in	You're a large-scale enterprise business with massive peaks and troughs in your scaling needs

Why a hybrid approach might be the answer

Choosing between bare metal and cloud isn't a case of choosing between good and bad. Both have their benefits under specific scenarios. And often the best way to optimize your infrastructure stack for cost and performance is to adopt a hybrid approach combining both. By choosing a vendor-agnostic approach to infrastructure, businesses can take what they need from cloud, what they need from bare metal, and (importantly) avoid lock-in.

That's exactly the approach Gameye takes. *"We have two different kinds of compute types"*, explains Walker[39]. *"So yes, we do have cloud, but we also use metal-as-a-service from other providers. Each compute type for us is about solving a particular issue. Do we need to scale rapidly? Do we need to cost-effectively manage a baseline? Do we need to backfill some expensive cloud with some more cost-effective bare metal? We just use the right tool for the right job. That's our philosophy"*. Gameye's agnostic approach means they can choose the right solution at the right time. *"We can swap in and out underperforming hardware, locations or providers without interrupting our customers' access,"* furthers Walker.

"Our agnostic approach means that we can choose the right solution for the right area or the right place".

Andrew Walker, Head of Business Development, Gameye

Other use cases calling for hybrid infrastructure include prototyping new projects and serving latency-sensitive content around the globe. *"When, for example, you have a busy infrastructure team...you can start prototyping things in the cloud...then get the permission of the infrastructure team to provision the right infrastructure for you to set up your project,"* examples the Global Director of Technology at an iGaming platform. *"Then there is also the case where, for example, you might be serving content which is performance-critical across the globe"*, they add[40].

"It's going to be difficult to set up infrastructure across the globe as close to the customer as it can get. So, then you might think about adopting an architecture where there is some last mile application serving component...as close to the customer as it can get. And I think, in that case, cloud would be a good candidate to complement your existing setup".

For others still, the best approach could be starting out in cloud and then diversifying once resource requirements become clearer.

Case study — Remoby DSP

Demand-side platform, Remoby DSP started its infrastructure journey in hyperscale cloud before shifting to a hybrid approach. In an interview with servers.com, Chief Technology Officer, Sergey Chernukhin, explains why.

"We started a new project and didn't know what resources [we needed] and to what extent we would need them. Additionally, it was important to have the ability to scale resources quickly within an hour or two in case of rapid growth or increased loads. Hyperscale cloud solutions are perfect for this. In fact, we chose the most popular and familiar one. AWS is a convenient service that covers almost all our needs. However, there were difficulties that we encountered and couldn't resolve with AWS.

It's critically important for us to respond to our partner's advertising requests promptly. Our services respond within the range of 5ms to 20ms; beyond that, network performance plays a significant role. It turned out that network delays between our partners and us were quite substantial. We tried various approaches within AWS, but we couldn't achieve acceptable response time. As it turned out, our partners were hosting their services with a bare metal provider, so we decided to migrate to the same provider to get closer to them. After migrating, the network ceased to be a problem for us.

The cost of traffic in AWS turned out to be unreasonably high for us. On the other hand, with [bare metal], we stay within the limits allocated for the public network on servers, essentially paying only for the channel.

We have already established a baseline configuration and a clear understanding of the hardware we need and when we need it. The need to provision resources 'here and now' has practically diminished, but if it arises, we still utilize AWS"[41].

Your roadmap out of hyperscale cloud

Every business is different. But if you're struggling in hyperscale cloud because of rising costs and complexities stifling your ability to scale your business then bare metal could be the answer. If saving money is your goal and you need to free up capital to invest into other revenue generating activities for your business then migrating to bare metal all, or in part, will help you regain control of your time and finances.

There's very little information and support out there to help businesses who've decided to migrate away from hyperscale cloud through the process. While the route out of hyper-scale cloud will be unique for each business, there are some essential steps that can be followed to ensure a smooth and effective migration.

1	2	3	4	5
Take time to understand your technology stack	Research hosting providers	Price up each hosting provider	Calculate a sum of possible migration costs	Compare current spend vs projected ROI

Step 1: Stop. Take time to understand your technology stack

Assess your cloud technology stack and identify where the inefficiencies lie.

- **Are you using too many instances and could you stop any?**
  Stopping some instances can help to reduce your monthly cloud bill.

- **Are you wasting resources?**
  Reducing your baseline resources and adding to them only when needed
  will help you spend more efficiently.

- **Did you over-engineer your cloud solution during set-up?**
  Many startups fall into this trap only to find that they are overspending
  on over-engineered resources that aren't being used.

Once you've identified any pitfalls or inefficiencies in your current cloud environment, consider whether you can make any changes to improve the cost and performance efficiency of your solution sufficiently. If the issues cannot be rectified, move on to step 2.
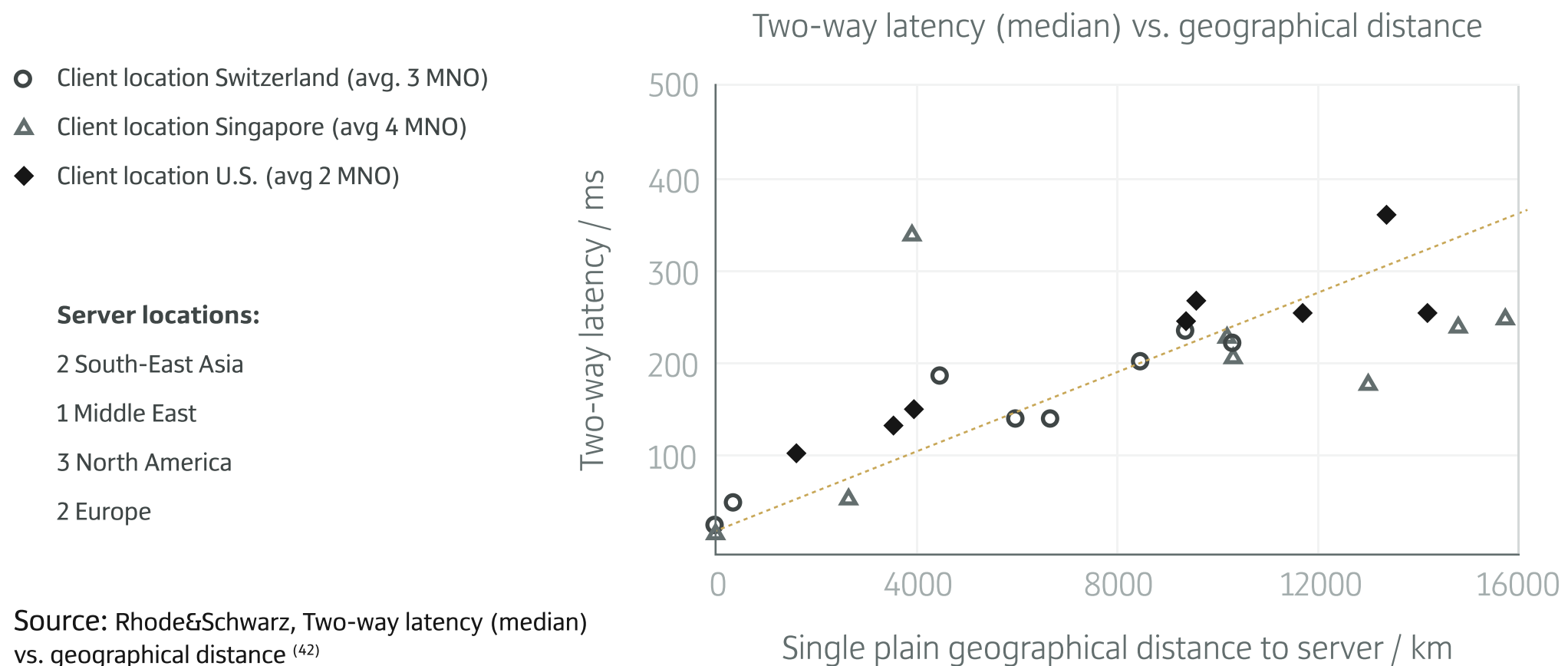
Step 2: Research providers

Start researching bare metal hosting providers, assessing each against the following criteria:

- **Data center locations.** It's important to pick a hosting provider with data centers in the right geographies. To reduce latency and improve user experiences, data centers should be located as close to end-users as possible.

  But there are also industry-specific reasons as to why location is so important. A streaming platform, for instance, may need a connected network of geographically dispersed servers. A content delivery network (CDN) will help to maintain optimum performance when streaming to global audiences.

  For the Web3 industry, on the other hand, European data centers are becoming a critical lifeline amidst continued regulatory uncertainty in the U.S. And for iGaming platforms, location is also critical. Operators must adhere to the regulations stipulated by their respective iGaming authority (some of which strictly mandate data center locations).

Two-way latency (median) vs. geographical distance

○ Client location Switzerland (avg. 3 MNO)

△ Client location Singapore (avg 4 MNO)

◆ Client location U.S. (avg 2 MNO)

Server locations:

2 South-East Asia

1 Middle East

3 North America

2 Europe

Source: Rhode&Schwarz, Two-way latency (median) vs. geographical distance [42]

- **Server provisioning time.** How quickly can servers be spun up (or down) along with your scaling requirements? Even if you don't require cloud auto-scaling, efficient provisioning is still critical. Bare metal scalability has improved tenfold in recent years so a good hosting provider should be able to provision and deploy new servers in 40-minutes or less.

- **Network performance & round-trip times.** To ensure network redundancy, servers should be connected to multiple networks - the internet, an out-of-band management network, and a global private network (protected at the physical and logical levels to ensure fast and safe data transfer). Along with server locations, a modern and redundant network is critical to ensuring fast round-trip times and low latency. Your hosting provider should be using top carriers, smart traffic routing, and give the option to geo-replicate your data if required.

- **Level of support.** Support can be make or break when it comes to forging a successful working relationship with your hosting provider. It's a good idea to try and meet with prospective providers to get a feel for the level of support on offer. Look for 24/7 access to the support team with additional account management and support from solutions architects. It's often overlooked, but high availability and highly personalized support makes a huge difference. It means that when there's an issue, you'll have the peace of mind that a known person is ready to help.

Step 3: Get quotes

Once you've narrowed down your preferred vendors, it's time to think about cost. Start by checking online pricing pages for a general indication of cost, but don't leave it there. As a next step get in touch with each vendor to discuss your specific needs and get a personalized estimate.

A good bare metal hosting provider will work with you to create a bespoke infrastructure environment optimized both in terms of cost and performance. And remember that the best choice won't necessarily be the cheapest — it's important to choose the vendor or vendors that will help you strike the right balance between affordability but also performance, customization, and support.

Step 4: Calculate a sum of all possible migration costs

Once you've picked your preferred infrastructure provider, calculate the total possible migration costs your business could incur by migrating to that provider. This number should reflect the cost of your new infrastructure and all possible internal costs incurred for the migration process itself.

There are many factors to consider when determining the end-to-end cost of a migration, so it can be challenging. But there are some universal costs that you should be factoring into your calculations. These include:

- **The cost of data transfer.** You'll need to work out (a) how much data you need to transfer, (b) how quickly that data needs to be moved, (c) how long the data transfer will take. It can help to use a cost calculator to generate an estimate.

- **Price of using a bare metal provider.** This includes the cost of servers in use and any redundant servers sitting in reserve.

- **Labor costs.** If your business will be hiring somebody to help execute the migration then you will need to factor this into the cost of the overall migration process.

- **Overlap.** To keep your applications running during the transition process, you may choose to plan in a period of overlap between your former and new infrastructure providers. During this time, you will continue to incur chargers from your former infrastructure provider until the old service is fully terminated.

  This is exactly the scenario forex brokerage FxGrow found itself in when migrating to a new hosting provider:

"One thing that we faced when migrating was paying two times on the migration costs. We were paying our previous hosting provider and paying our new hosting provider until the migration finished. It took us two weeks to fully migrate to the new hosting provider"[43].

Mohammad Mazeh, Head of Brokerage and Platform Administration, FxGrow

---

- **Post migration expenses.** This includes things like ongoing costs for licensing, maintenance, compute fees, and administration (if not included as part of the monthly hosting bill).

Server migration costs

1	2	3	4	5
Cost of data transfer	Price of using a bare metal hosting provider	Labor costs	Overlap	Post migration expenses

Step 5: Compare current spend vs. projected ROI

Once you have a total sum of your migration costs, compare what you're currently spending with the projected long-term ROI of your migration. The purpose of this step is to figure out if a migration is a smart financial decision for your business.
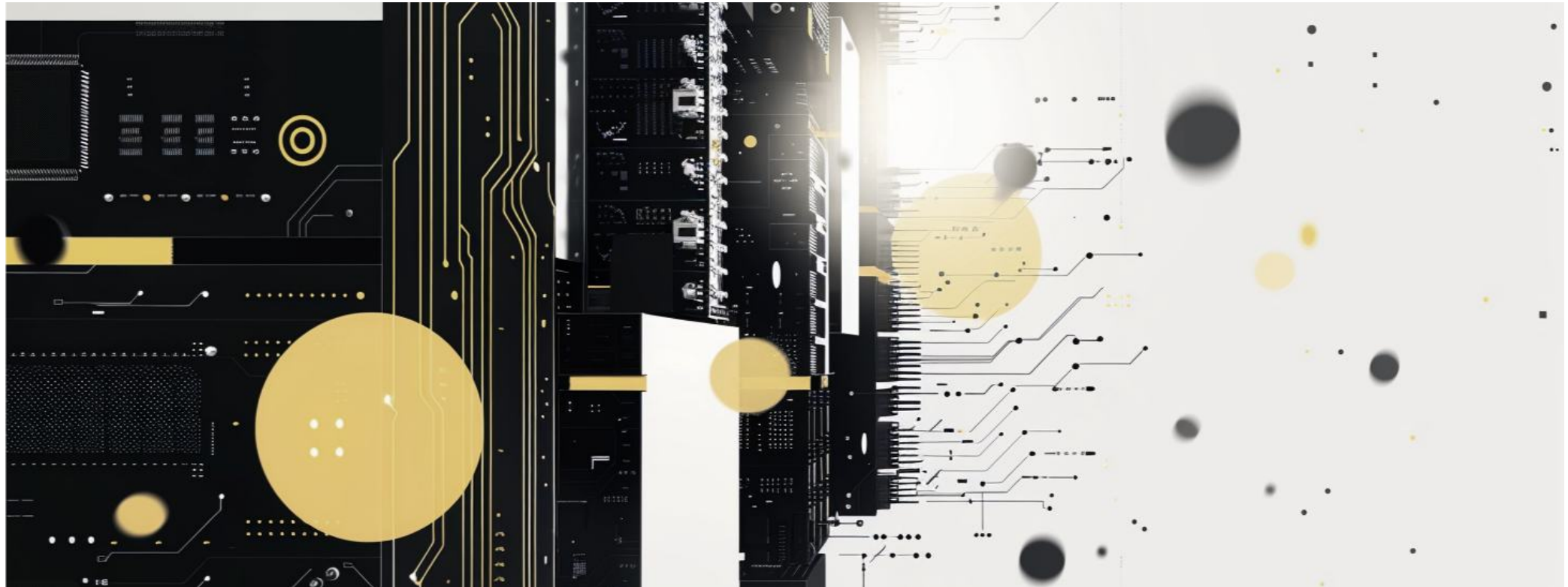
When setting your goal, you'll need to determine a minimum saving that would make migration worthwhile. For example, you may conclude that you need to achieve a minimum 20% total cost saving to make the migration a viable project (but bear in mind that this number will be different for every business).

This stage is particularly important if your migration is driven by cost-optimization. And the road to cost-optimization will be different for everyone. Say, for example, you're a game studio and you've recently launched a new video game. As Brennan explains, in order for a game studio to cost-optimize effectively *"you have to work out how many hours a day a bare metal machine needs to be in use for it to equate to cloud. Then calculate your player curve and understand what your player curve looks like on a per-region basis"*[44].

Step 6: Be prepared for a few hiccups along the way

When you start your migration, it's normal to face some challenges along the way. Especially if you'd previously built your entire infrastructure stack around a single vendor. You'll likely find that some of the products or services you were making use of aren't available elsewhere — and you'll need to work with your new hosting provider(s) to find suitable alternatives.

That's precisely the situation Remoby DSP found themselves in after deciding to migrate to bare metal from AWS. Chernukhin explains: *"Due to the absence of a block storage equivalent on the platform (e.g. Amazon EBS), we had to change our approach to deploying stateful applications and modify configurations"*[45].

Made with AI

But Remoby were able to overcome this initial hurdle and achieve their desired outcomes: improved response times and significant savings on both traffic and server resources. *"At the beginning, there was some chaos and confusion before we aligned with our managers",* furthers Chernukhin.

"There was a moment when we were ready to abandon the migration idea. Fortunately, we managed to work together to come up with a solution that allowed us to progress rapidly and pleasantly".

Sergey Chernukhin, Chief Technology Officer, Remoby DSP

Migration is a serious decision and teething problems are to be expected. The best way to ensure that these issues can be resolved as painlessly as possible is to work with a hosting provider that you can trust. Your vendor should be able to guide you through the migration process with consistent support from technical experts to create a bespoke infrastructure solution that works for your business.

Conclusion — It's time to question the status quo

The history of cloud is a story of massive innovation. As the internet boomed, cloud offered businesses a way to get online with easy access to hosting. Cloud is amazing. But it's not perfect. And that's the problem. Hyperscale computing has outgrown itself and for many of its customers, managing and maintaining infrastructure within these environments has become unsustainable. To the extent that for any business with incremental scaling requirements, it's probably doing more harm than good.

As the issues intensify, for many, a cloud repatriation (either part or full) is becoming the only viable solution. For some, it means engineering a full reverse cloud migration to an on-premises solution. But for businesses without the necessary resources or desire to take on the additional management and maintenance of hardware, there are other options available.

Migrating to bare metal hosting or incorporating bare metal hosting as part of a hybrid infrastructure approach can help organizations maintain cloud-like scalability and performance alongside greater scope for customization, predictable billing, and frontline support.

About servers.com

Trust sits at the heart of hosting. Which is why we design bespoke hosting solutions for every customer. Tailored to you, in consultation with our experts, our solutions marry the best of traditional hosting and cloud scalability to deliver powerful bare metal performance that you can rely on.

That quality extends to the exceptional service provided by our team of specialists, who work in trusted partnership with customers to support them throughout their time with us.

Available from prime locations worldwide, we are defining the next era of hosting with an uncompromising dedication to excellence that you can trust in.

Sources

1. VB Staff 2022, Report: 81% of IT teams directed to reduce or halt cloud spending by C-suite, VentureBeat, accessed 23 April 2024, <https://venturebeat.com/data-infrastructure/report-81-of-it-teams-directed-to-reduce-or-halt-cloud-spending-by-c-suite/>

2. Moore, J 2023, Cloud cost management takes center stage, TechTarget, accessed 23 April 2024, <https://www.techtarget.com/searchitchannel/news/365532532/Cloud-cost-management-takes-center-stage>

3. Ibid.

4. Teal, K 2023, 'Cautious' Org' Cloud Cost Optimization Efforts Dent AWS Revenue, Channel Futures, accessed 23 April 2024, < https://www.channelfutures.com/cloud/-cautious-orgs-cloud-cost-optimization-efforts-dent-aws-revenue>

5. Brennan, L. (2023). Interviewed by Frances Buttigieg and Hannah MacLeod [Skype], 18 August.

6. ARPANET n.d, Defense Advanced Research Projects Agency, accessed 25 April 2024,<https://www.darpa.mil/about-us/timeline/arpanet#:~:text=Its%20initial%20demonstration%20in%201969,enabled%20ARPANET%20to%20become%20operational>

7. Imagining the Internet n.d., ELON University, accessed 25 April 2024, <https://www.elon.edu/u/imagining/time-capsule/150-years/back-1960-1990/>

8. Petrova, B 2023, DNS history. When and why was DNS created?, CloudDNS, accessed 25 April 2024, <https://www.cloudns.net/blog/dns-history-creation-first/>

9. The birth of the Web n.d, CERN, accessed 25 April 2024, <https://home.cern/science/computing/birth-web>

10. Internet Growth Statistics n.d, Internet Growth Stats, accessed 25 April 2024, <https://www.internetworldstats.com/emarketing.htm>

11. Ibid.

12. Surdival, O 2020, Birth of the Public Cloud — The History, How & Why, LinkedIn, accessed 25 April 2024, <https://www.linkedin.com/pulse/birth-public-cloud-history-how-why-oliver-surdival/>

13. Nagorney, I. (2023). Surveyed by: Frances Buttigieg. 25 September.

14. [Global Director of Technology, iGaming platform]. (2023). Interviewed by Frances Buttigieg and Jamie Daniel [Skype], 8 September.

15. Walker, A. (2023). Interviewed by Frances Buttigieg, Hannah MacLeod and Jarrod Palmer [Skype], 26 September.

16. Nagorney, I. (2023). Surveyed by: Frances Buttigieg. 25 September.

17. Walker, A. (2023). Interviewed by Frances Buttigieg, Hannah MacLeod and Jarrod Palmer [Skype], 26 September.

18. Amazon EC2 On-Demand Pricing n.d., AWS, accessed 25 April 2024, <https://aws.amazon.com/ec2/pricing/on-demand/#Data_Transfer>

19. Chervek, E 2023, *Hyperscalers push anticompetitive practices, ignore industry cries for multicloud*, sdxcentral, accessed 25 April 2024, < https://www.sdxcentral.com/articles/analysis/hyperscalers-push-anti-competitive-practices-ignore-industry-cries-for-multicloud/2023/10/>

20. [Global Director of Technology, iGaming platform]. (2023). Interviewed by Frances Buttigieg and Jamie Daniel [Skype], 8 September.

21. Heinemeier Hansson, D. (2022). Why we're leaving the cloud. [Blog] David Heinemeier Hansson. Available at: https://world.hey.com/dhh/why-we-re-leaving-the-cloud-654b47e0 [Accessed 25 April 2024].

22. Heinemeier Hansson, D. (2023). We have left the cloud. [Blog] David Heinemeier Hansson. Available at: https://world.hey.com/dhh/we-have-left-the-cloud-251760fb [Accessed 25 April 2024].

23. Brennan, L. (2023). Interviewed by Frances Buttigieg and Hannah MacLeod [Skype], 18 August.

24. Ibid.

25. [Global Director of Technology, iGaming platform]. (2023). Interviewed by Frances Buttigieg and Jamie Daniel [Skype], 8 September.

26. Varvaluk, I. (2023). Surveyed by: Frances Buttigieg. 19 September.

27. Earls, A 2023, Will data centers become obsolete?, TechTarget, accessed 25 April 2024, < https://www.techtarget.com/searchdatacenter/feature/Will-data-centers-become-obsolete>

28. [Global Director of Technology, iGaming platform]. (2023). Interviewed by Frances Buttigieg and Jamie Daniel [Skype], 8 September.

29. Global Data Center Trends 2023, CBRE, accessed 25 April 2024, <https://www.cbre.com/insights/reports/global-data-center-trends-2023>

30. Maurer, M 2023, 'Big Tech Expects Some Assets to Last Longer. But the Boost to Profit Is Temporary', The Wall Street Journal, 4 May, accessed 25 April 2024, <https://www.wsj.com/articles/big-tech-expects-some-assets-to-last-longer-but-the-boost-to-profit-is-temporary-66ce9f98>

31. Brennan, L. (2023). Interviewed by Frances Buttigieg and Hannah MacLeod [Skype], 18 August.

32. Ibid.

33. Walker, A. (2023). Interviewed by Frances Buttigieg, Hannah MacLeod and Jarrod Palmer [Skype], 26 September.

34. Nagorney, I. (2023). Surveyed by: Frances Buttigieg. 25 September.

35. Ibid.

36. Ibid.

37. Walker, A. (2023). Interviewed by Frances Buttigieg, Hannah MacLeod and Jarrod Palmer [Skype], 26 September.

38. Nagorney, I. (2023). Surveyed by: Frances Buttigieg. 25 September.

39. Walker, A. (2023). Interviewed by Frances Buttigieg, Hannah MacLeod and Jarrod Palmer [Skype], 26 September.

40. [Global Director of Technology, iGaming platform]. (2023). Interviewed by Frances Buttigieg and Jamie Daniel [Skype], 8 September.

41. Chernukhin, S. (2023). Surveyed by: Frances Buttigieg. 6 September.

42. Berger, J 2021, Interactivity test: Distance to server impact on latency and jitter (part 4), Rohde & Schwarz, accessed 25 April 2024, <https://www.rohde-schwarz.com/se/solutions/test-and-measurement/mobile-network-testing/stories-insights/article-interactivity-test-distance-to-server-impact-on-latency-and-jitter_255082.html>

43. Mohammad, M. (2023). Interviewed by Frances Buttigieg & Michael Sparshott [Skype], 16 November

44. Brennan, L. (2023). Interviewed by Frances Buttigieg and Hannah MacLeod [Skype], 18 August.

45. Chernukhin, S. (2023). Surveyed by: Frances Buttigieg. 6 September.

The information provided in this whitepaper is for general informational purposes only. While we strive to present accurate and up-to-date information, we make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability with respect to the content contained herein. Any reliance you place on such information is therefore strictly at your own risk.

Any quotes or testimonials included in this whitepaper are provided by customers voluntarily and may not represent the views of their respective companies.

This whitepaper may contain references to third-party sources for informational purposes. Views expressed in such sources are the opinion of the author, and not necessarily those of servers.com. We make no representation regarding the content, accuracy, or completeness of any third-party information and shall not be liable for any errors, omissions, or delays in this information or any losses, injuries, or damages arising from its use.

Certain content, including but not limited to quotes, excerpts, and images, may be sourced from and/or owned by third parties and are used under fair use principles or with permission. All rights to such material belong to their respective owners, unless stated otherwise, and inclusion in this whitepaper does not imply ownership or endorsement by servers.com.

Images used in this whitepaper may be obtained from various sources, including stock photo websites, and are used for illustrative purposes only and are not owned by servers.com unless otherwise stated.