CLOUD COST MANAGEMENT

BRING YOUR COMPANY'S CLOUD SPENDING UNDER CONTROL

OVERVIEW

Businesses cite infrastructure cost savings as one of the main reasons for hosting applications in the cloud. They want to do more with infrastructure while spending less.

Cloud flexibility and on-demand pricing hold out the promise of bringing server spending under control.

- Infrastructure-as-a-Service (IaaS) platforms are elastic so server resources can be scaled up or down in response to business needs.
- Businesses only pay for the compute and storage resources they use.
- Billing increments are hourly rather than monthly or longer for other types of server hosting.
- Cloud platforms are post-paid: businesses pay for what they use, rather than paying up-front for resources they might not need.

But businesses often find it challenging to turn the theoretical cost benefits of the cloud to their advantage.

According to the RightScale 2019 State of the Cloud Report, as much as 35 percent of cloud spend is wasted.

- Many cloud platforms have ferociously complex pricing structures. It is difficult for businesses to predict how much they will be billed. Gartner found that 80 percent of organizations overshoot IaaS cloud budgets.
- Businesses that "lift and shift" applications into the cloud duplicate legacy infrastructure on cloud platforms They over-provision infrastructure and fail to take advantage of cloud elasticity.
- The flexibility of the cloud can work against businesses that don't implement and enforce infrastructure provisioning and teardown policies.

Infrastructure-as-a-Service gives businesses the tools they need to manage infrastructure costs, but businesses without a cloud cost management plan may not realize the full cost benefits of the cloud's agility and flexible pricing.
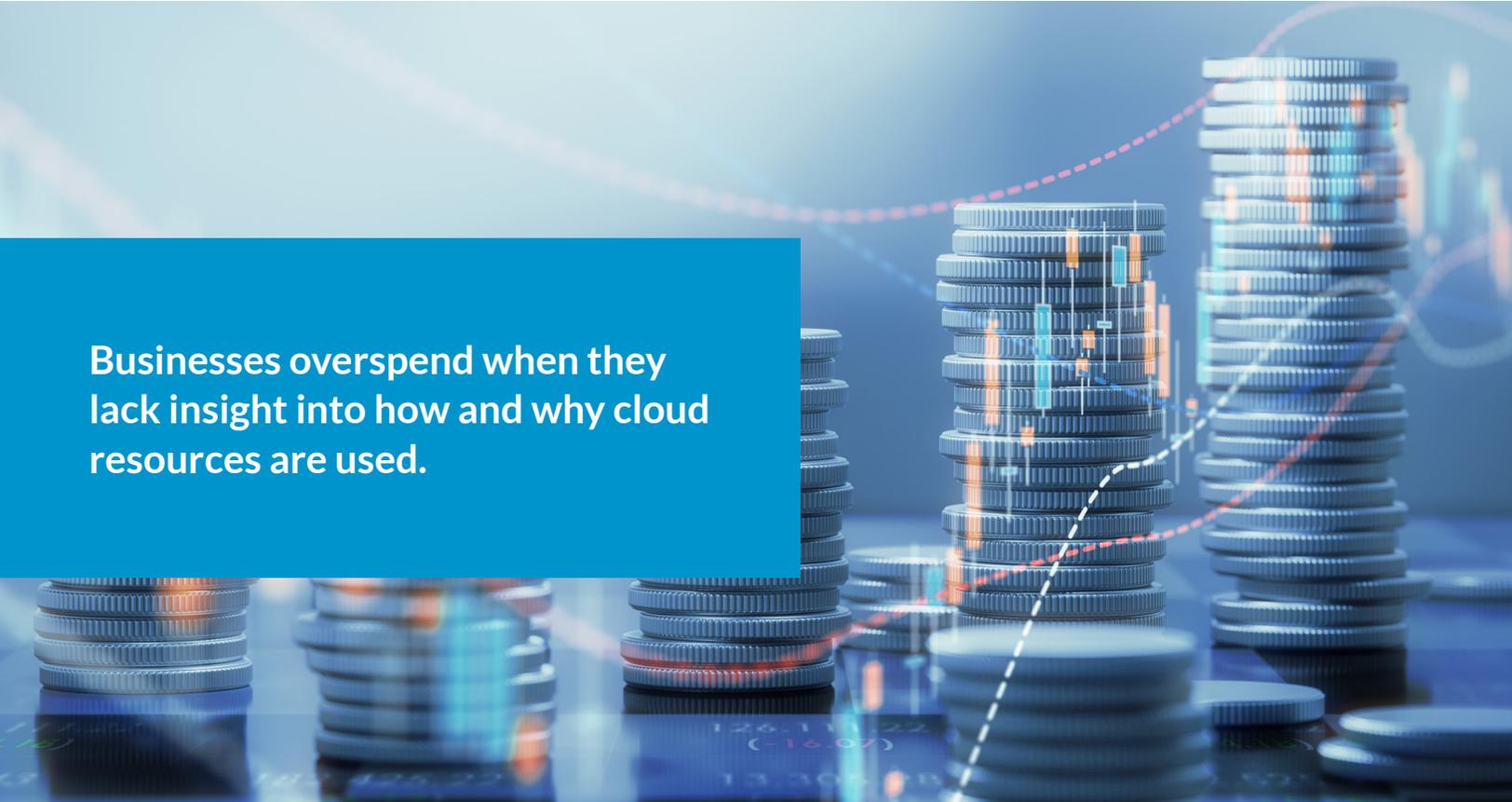
In this white paper, we're going to look at techniques that businesses can implement to maximize their cloud infrastructure value.

Businesses overspend when they lack insight into how and why cloud resources are used.

Provisioning infrastructure on the cloud is fast and easy. An employee can provision a cloud server in seconds. Teams can provision servers to meet a project's infrastructure needs without waiting for approval from managers or the IT department.

Easy on-demand server deployment fosters innovation and agility. The cost of experimentation is low. We are a long way from the old days of months-long lead times and complex authorization processes. But cloud agility has a potentially harmful side-effect: cloud infrastructure tends to grow and change so quickly that businesses lose track.

Businesses can see the infrastructure they have deployed in the cloud control panel, but they may have no idea why it was deployed or how it is used.

The first step in cloud cost management is to **identify which infrastructure the business has deployed,** why it was deployed, and what it is used for today. Once a business has an overview of its infrastructure, it can develop strategies and implement controls that optimize spending without disrupting operations.

In the previous section, we talked about developing an understanding of who is spending what on cloud infrastructure and why they are spending it. The next step is to design and implement a plan of action.

Awareness is only useful when it informs action. A cloud cost management plan describes the actions the organization intends to take to control cloud costs.

A cloud cost management plan should include:

Immediate cost remediation steps that focus on easy wins. For example, if your organization has identified a deployment of underutilized cloud servers, plan to rightsize or remove them.

Monitoring and responsibility. Who in your organization is responsible for monitoring cloud utilization to ensure that it remains cost-effective? If no-one is responsible, as is often the case with smaller organizations, cloud resource use will not be efficiently managed.

Process and automation. Make it easy for employees and managers to responsibly manage cloud costs by implementing clearly-defined processes for infrastructure provisioning and deletion. For example, the ServerMania Cloud offers server templates and an API that allows for the development of custom deployment and teardown scripts for specific scenarios.

A cloud cost management plan describes your organization's intention to implement cost management best practices, many of which are discussed in the remainder of this white paper.

Underutilization is the biggest cause of cloud waste. In fact, it is the biggest cause of waste across all infrastructure hosting modalities; it is notoriously difficult to align infrastructure provisioning with unpredictable real-world requirements.

However, it is far easier to accomplish cost-effective utilization on an Infrastructure-as-a-Service platform than on colocated or on-premise infrastructure.

- Cloud servers can be resized on the fly.
- Cloud servers can be removed in minutes with no penalty.
- Infrastructure is paid for after it is used, not in advance.

Rightsize New Servers

Before deploying new servers, make an estimate of the CPU, memory, and storage resources the server is likely to need. This should take into account:

- The CPU and RAM requirements of the software the server will host.
- The size of the software and the amount of data that will be stored locally.
- The expected bandwidth consumption of the server.

Select a server from the instance types available on the cloud platform that matches projected resource requirements. These are calibrated for a variety of use-cases and will include allocations of RAM, CPU, storage, and bandwidth at a predefined price.

NAME	RAM	CORES	STORAGE	BANDWIDTH	MONTHLY
S1-25	1 GB	1 CPU	25 GB	1 TB	**$5**
S2-50	2 GB	1 CPU	50 GB	2 TB	**$10**
S4-80	4 GB	2 CPU	80 GB	4 TB	**$20**
S6-100	6 GB	2 CPU	100 GB	4 TB	**$30**
S8-160	8 GB	4 CPU	160 GB	5 TB	**$40**
S16-320	16 GB	6 CPU	320 GB	8 TB	**$80**

VIEW ALL CLOUD OPTIONS

From a cost management perspective, it is better to rightsize servers accurately before they are deployed, but it is not critical. Your cloud servers can be resized at any point to more accurately reflect real-world resource consumption.

Rightsize Existing Servers

Rightsizing existing servers is a two-step process. First, identify underutilized servers. The ServerMania Cloud provides CPU, Memory, and Storage utilization metrics, and there are many open-source tools available for assessing server utilization, ranging from standard Linux tools such as "top" to comprehensive monitoring applications such as Nagios.

Second, adjust server resources to align them with demand. This can be achieved both via the cloud platform's web control panel and via the API.

Businesses more familiar with on-premise and colocated hosting platforms often overprovision in the cloud because they are used to overprovisioning bare metal servers. Cloud platforms are infinitely more flexible. There should be a margin of error—resource use is unpredictable—but servers can be quickly scaled back up should additional CPU, memory, or storage be required.
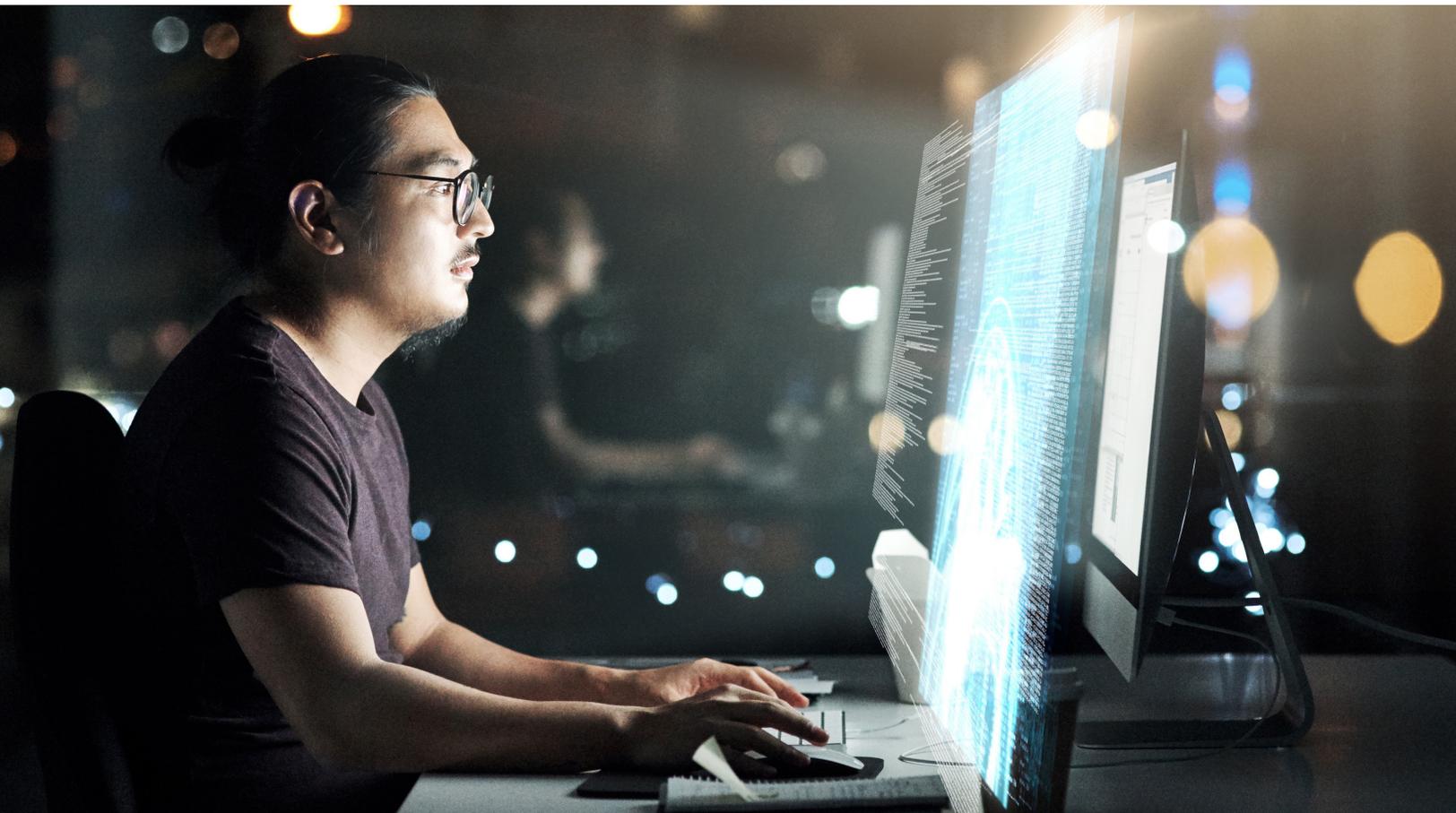
A zombie server is a cloud server that is provisioned but unused. The business pays for the server and its resources but gets nothing in return. An estimated 30 percent of servers are zombies. They cost businesses millions of dollars per year.

Zombie servers are the result of human process failures. A server is provisioned but never used. Or a server that once did something useful is retired without being removed from the cloud user's account. It is not unusual for a server to be provisioned, booted, and shut down but never deleted. Cloud users pay for servers even when they are not running because the resources are allocated.

As with all cloud spending optimization strategies, removing zombie servers is a process of identification and remediation. Audit the cloud servers that your organization has provisioned and ask how they are being used. Are they being used at all? If they aren't, shut them down and remove them from your cloud account.

Pay particular attention to unused storage. Pre-configured cloud instances allocate storage proportional to the amount of RAM and number of virtual CPUs. A high-end cloud server with many vCPUs and a large RAM allocation may also have several terabytes of storage. Applications that require lots of processing power don't always require this much storage, and it may benefit cloud users to scale back on storage that would otherwise be unused.

Furthermore, ensure that the data stored on a server is essential. It's easy to let a server's storage allocation fill up with unneeded logs and other data. As a general rule, if data isn't useful to your organization, delete it to free up space and reduce storage costs.

Rightsizing focuses on vertical scaling—growing and shrinking the resources available to individual cloud servers. But horizontal scaling, which is often referred to as scaling out, is often preferred on Infrastructure-as-a-Service platforms.

Instead of making a single server larger or smaller, horizontal scaling adds or deletes multiple smaller servers. An app is hosted across a cluster of servers. Horizontal scaling is often the more flexible and agile approach.

An alternative approach is to divide the application's components across multiple smaller cloud servers: several load-balanced web servers, a database server, servers dedicated to long-running processing jobs, file servers, and so on. Rearchitecting in this way allows the components of the app to be scaled independently.

As web requests rise, additional web servers can be deployed in seconds and torn down when they are no longer required. As the size of the database increases, the database server can be scaled. CPU-intensive tasks can be moved to high-CPU cloud servers, leaving web and database servers free to focus on their primary task.

There are many performance and availability benefits to rearchitecting for horizontal scaling, but there are also cost advantages.

- It is typically less expensive to provision multiple low-cost servers with precisely calibrated resource allocations than to (over)provision a single large server.
- With horizontal scaling, individual servers are ephemeral. They do not have to be managed and maintained with the same care as the single server on which the whole application depends.
- Horizontal scaling can be automated, reducing management costs and ensuring resource availability.

Consider an organization that hosts a business-critical application on a single high-performance dedicated server, which hosts the database, the webserver, and the application code. It is tempting to "lift and shift" this resource-hungry application to an equally high-powered cloud server.

Rearchitecting for the cloud increases the complexity of cloud migration and infrastructure management, but it brings numerous benefits for application and infrastructure performance, availability, automation, and cost-efficiency.
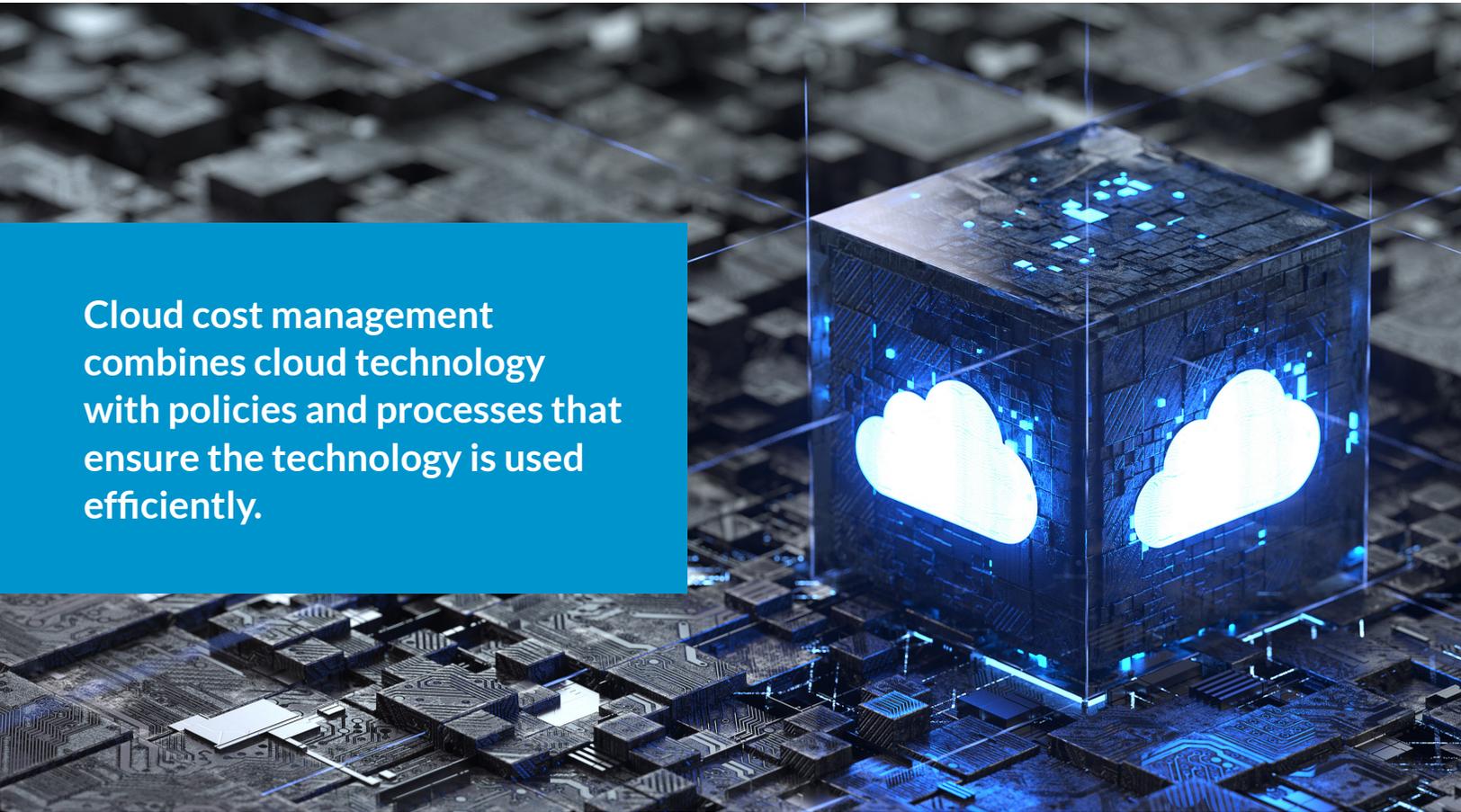
Cloud cost management combines cloud technology with policies and processes that ensure the technology is used efficiently.

Infrastructure-as-a-Service has the potential to be more cost-efficient than other hosting modalities, but an organization that does not manage its cloud infrastructure is liable to waste money on underutilized compute and storage resources.

Automation is one way businesses can ensure that cloud users follow cost management best practices, as we hinted at in the previous section. Cloud users can mandate the instance type and software their employees should use in specific scenarios, limiting the potential for over-provisioning and ensuring that resources are allocated efficiently.

The ServerMania Cloud provides templates for automation. A template represents a preconfigured operating system and software environment that will be installed on a newly deployed cloud server.

> Automation gives organizations some control over how their cloud platform is used, but enforcing policies and allocating responsibility are just as important. Your managers and teams should understand their responsibility for auditing cloud costs and implementing cloud management best practices.

CLOUD COST MANAGEMENT ON THE SERVERMANIA CLOUD

The ServerMania Cloud is a powerful Infrastructure-as-a-Service platform that simplifies cloud cost management. The ServerMania Cloud offers:

- A straightforward "no surprises" pricing structure to help customers to stick to their infrastructure budgets

- Hourly billing on all cloud servers.
- Fast scaling up and down.
- Automation via API and cloud server templates.

To learn more about how the ServerMania Cloud can help your business to reduce IT infrastructure spending, contact a cloud hosting specialist today for a free initial consultation.

VIEW SERVERMANIA CLOUD OPTIONS

serverMANIA
HOSTING EMPOWERED

A leader in web hosting solutions since 2002.

www.servermania.com

f facebook.com/servermaniainc

in linkedin.com/company/server-mania

🐦 twitter.com/servermaniainc

▶ youtube.com/servermaniainc