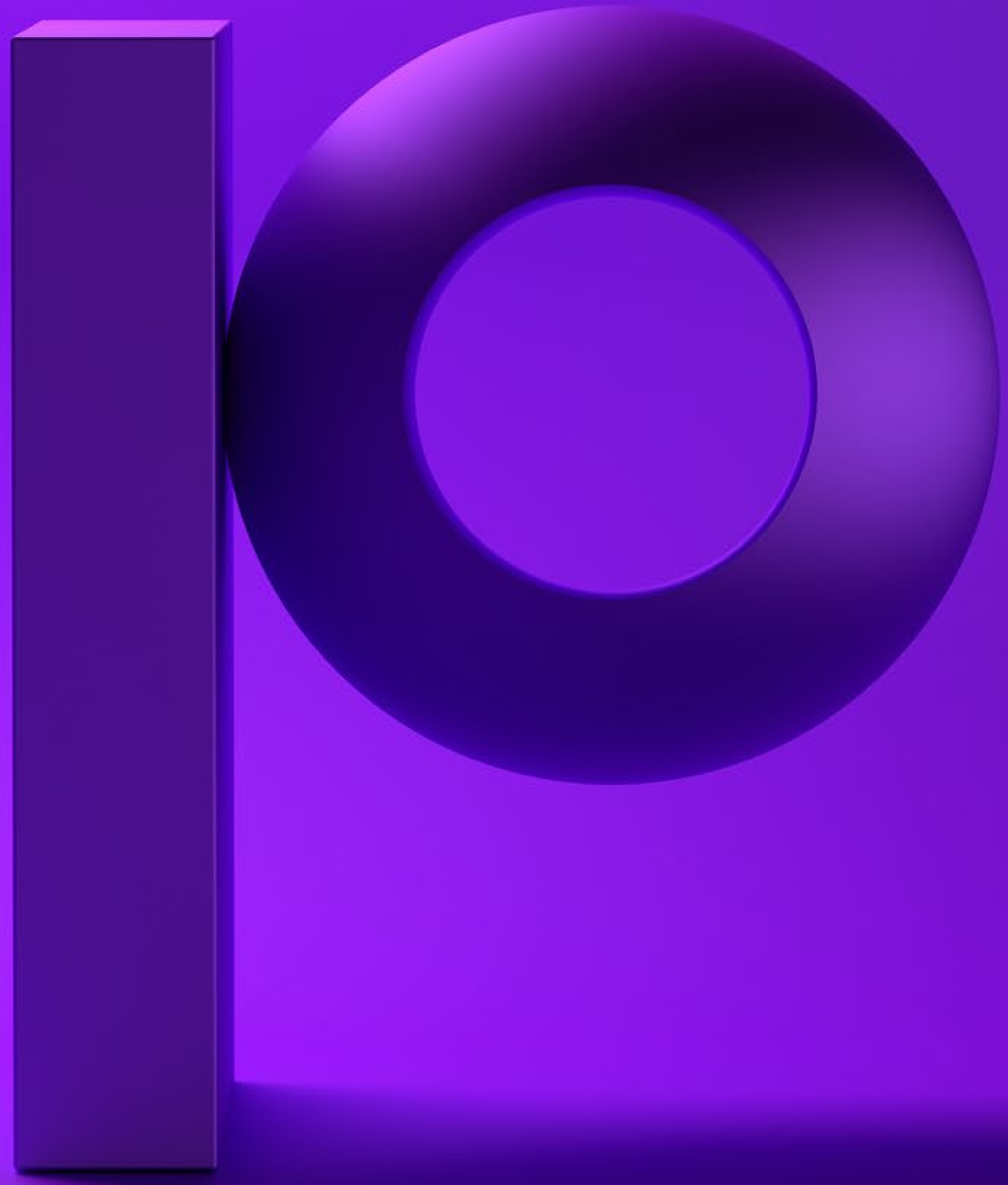
precisely

Controlling Cloud Costs with Capacity Management



Introduction

For many organizations the move to a cloud-first IT strategy has already taken place. In fact, today's typical IT department has a minority of their apps and platforms residing in on-premises data centers. With the increasing use of cloud environments, it's key to ensure you're managing and optimizing your cloud resources as you would with on-premises resources. Across both cloud and on-premises resources, capacity management informs forecasting and planning by helping you determine the capacity levels for your environment, including compute configurations, storage, database, and network bandwidth – as well as the most cost-effective way to provision them.

This eBook looks at what it means to extend capacity management to the cloud and how it differs from traditional on-premises capacity management.

What Is Cloud Capacity Management?

Capacity management has been used for decades to optimize resources on-premises. Now, as cloud environments transform IT, this practice is being extended to enable holistic planning, management, and optimization of all your resources — both cloud and on-premises — in one place and at the same time.

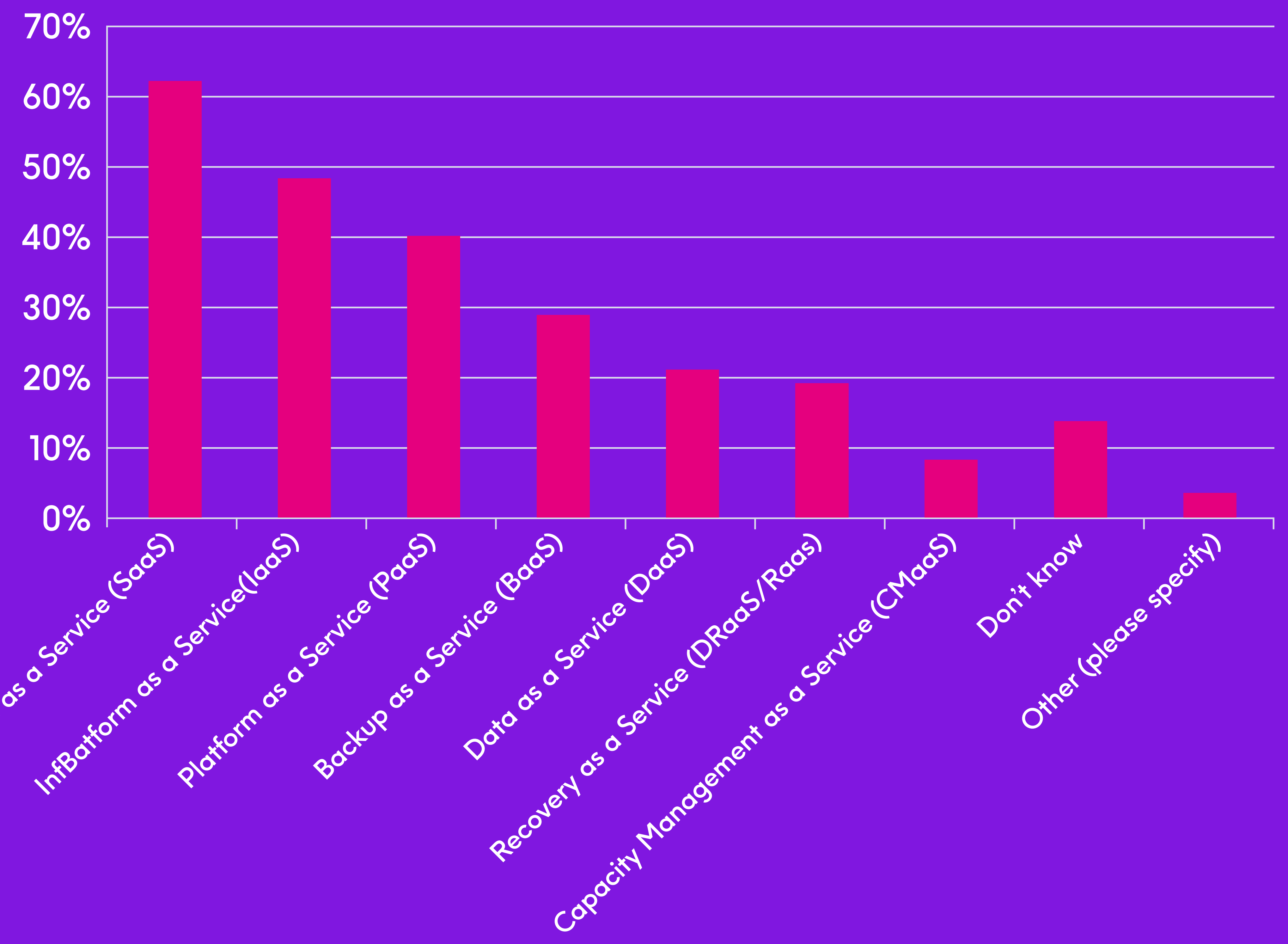
For today's businesses, capacity and cost management are essential to ensure adequate resources and budget, whether on-premises or in the cloud. This discipline is necessary to successfully support new, existing, and growing business services. If you have not yet moved workloads to the cloud, optimizing and right-sizing resources before the move to cloud helps prevent overprovisioning, unnecessary operating expense, cloud sprawl, and excessive management complexity. Another important pre-cloud practice is performance benchmarking. You want to ensure that cloud resources will provide the same or better performance as on-premises resources.

In a recent Precisely survey, customers listed Capacity Management as one of the top challenges they were struggling with related to their cloud environments. According to Gartner, approximately 28 percent of server capacity currently goes unused, as well as 40 percent of storage. As applications move to the public cloud, capacity management can help you understand what on-premises resources can be decommissioned and how to optimally restack on-premises workloads on the resources that remain.

Preventing waste is a key goal of capacity management — but it's also essential to ensure adequate capacity for the applications and services that run on cloud resources. The objective is to get the most for your cloud spend while ensuring a good experience for customers and business users. In order to accomplish this, your capacity management tool needs to scan your current environment usage for configuration corrections that can be applied to improve the performance of cloud-based services. Additionally, it needs to identify possible configuration remediations to enable performance improvements. Your capacity management solution should help you identify additional opportunities for efficiency or performance enhancement, such as identifying resources not properly decommissioned or resources still available but not in use. Finally, you need to implement new polices based on the data you are collecting to improve sizing, handling unused or overprovisioned capacity, and so on.

What Are Companies Doing in the Cloud?

Does your company use any of the following Cloud services? Select all that apply.



Why Capacity Management is Different for Cloud Environments?

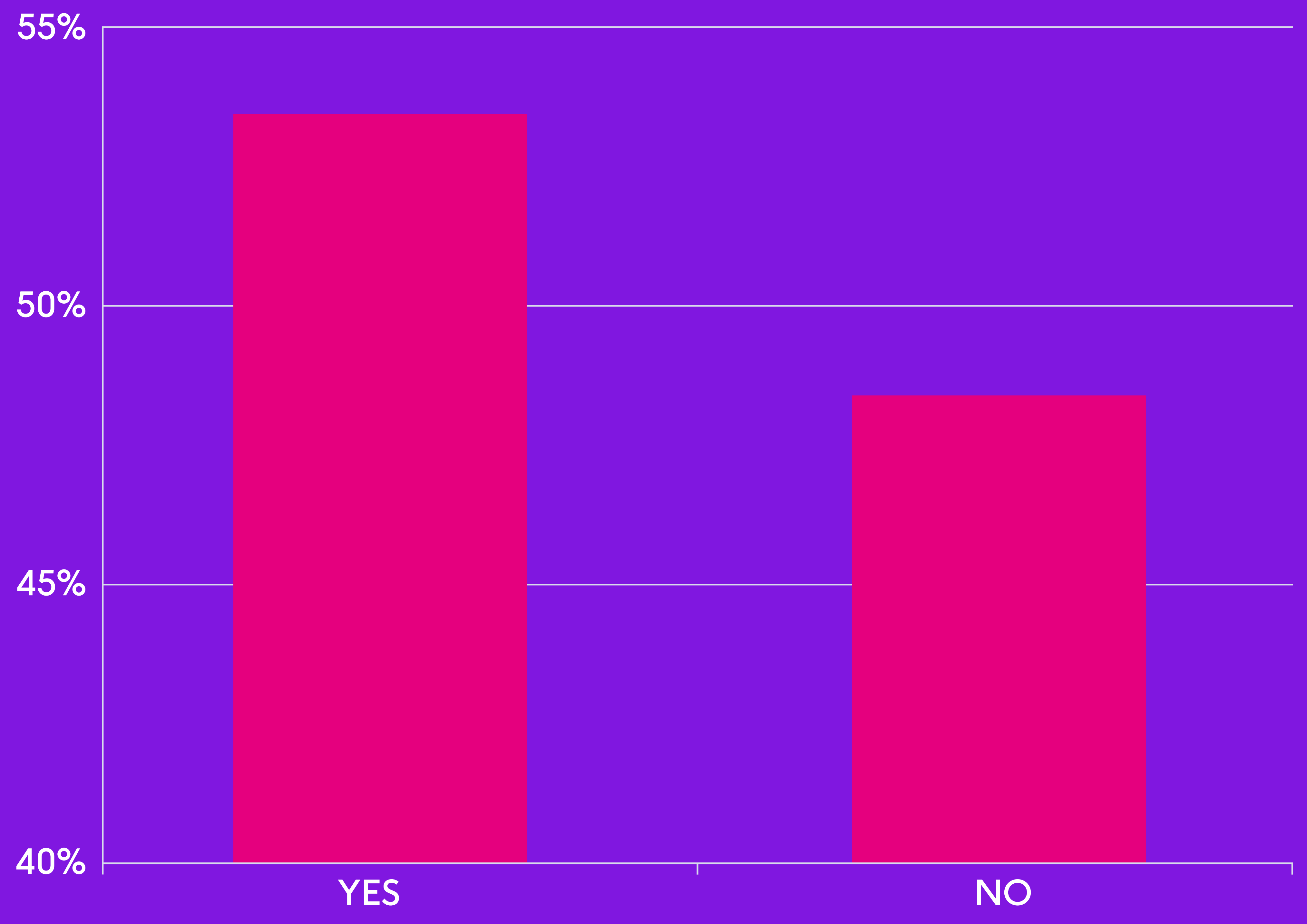
One of the first advantages we generally think about with cloud computing is the flexible usage of resources. The cloud requires less hardware than traditional computing structures which can mean greater flexibility and a lower upfront cost. It can be a simple, quick process to purchase and deploy additional resources. However, this elasticity is only useful if your implementation strategy is good. Even though increased capacity can be easily acquired, it is important to know what your company is currently using and will need in the future. You need to know how to use that power effectively because everything in a cloud computing environment is shared using some sort of multitenant model which complicates capacity modeling and planning. The ease of creating capacity on demand creates the scenarios of idle capacity—for example, the spawn of some infrastructure or test and the failure to turn it off ("by the time of day" or "shut down"). An example of underutilized capacity is having 100 TB storage when 30 TB is needed now, 50 TB next year and possibly more the next year.

You can use cloud computing systems as needed to cost-effectively provide temporary capacity. Often called "cloud-bursting," the cost of this type of architecture has been difficult to justify until cloud computing provided a cheaper "public" option. The ability to provide additional resources with short notice is an efficient option to providing capacity needs – but the downside is that it can be very expensive to rely on cloud-bursting as a regular option instead of having a well-defined management process of resources.

On the surface, it may seem like some aspects of capacity planning can be less important for cloud deployments. Since capacity can be allocated as needed, one could be lulled into thinking that you can just respond when the demand is needed. You may think there is less need to pay attention to capacity needs – however, cost is a core driver for the use of cloud computing. Having more capacity than needed at any time costs money and reduces the cloud's value.

Cloud Costs Are Frequently Higher Than Expected

Has the cloud cost your company more than anticipated?



Understanding Resource Usage in the Cloud

In order to ensure consistent performance, you must be able to understand resource usage patterns over time. It is not enough to just be alerted when you have a critical issue or when a threshold gets exceeded. You need a method to identify trends and growth patterns that will give you an early warning when you are headed toward a problem when you still have enough time to respond. This predictive analysis is even more crucial in the flexible, frequently changing world of cloud computing. An understanding of resource usage patterns over time helps you determine the capacity levels needed to ensure consistent performance for both your on-premises and cloud environments.

Understanding when changes in workloads occur is essential to efficient use — especially in the cloud, where you're paying for resources on a daily, hourly, minute, or second-by-second basis. You need to be able to differentiate between regular periodic behaviors and one-time spikes in workload. By understanding these behaviors, you can make better and more informed decisions on how to handle and resource applications to ensure performance without wasting resources.

Effectively organizing and optimizing your use of resources is also critical. This may mean making configuration changes in workloads such as adding memory or CPU. Ideally, some level of automation can be applied here. This type of configuration optimization often requires automation to be done effectively. Manual efforts like importing data into spreadsheets to drive your analysis can be as much as 30 days out of date. You will not be successful in keeping abreast of the pace of change in the modern enterprise without an automated solution.

Cloud Capacity Management Questions to Address

- How many additional VMs can we still deploy?
- How can we increase the efficiency of our virtual hosts?
- Which is the most constrained resource and the most likely to impact our services based on current trends?
- How much spare capacity do we have?
- When will we saturate resources based on business growth?
- Do we need to buy more VMs, increase or decrease the size of the ones we're currently using, or change the type?
- Do we need to increase or decrease storage?
- How much will these changes cost?
- Would it be cheaper to move to a different cloud vendor?

Don’t Neglect Reporting

A key sign of capacity management maturity is the ability to automatically generate reports and dashboards that can be distributed to stakeholders. No matter where you are in your cloud capacity management maturity, effective reporting practices are vital to your success. Effective reporting is imperative for you to understand what is going on in your environment today and can help you see into the future to predict the challenges coming down the road.

Even if you are managing to stay on top of the capacity requirements of your cloud environment, inevitably, you are going to need to get management support for changes and investments you will need to make. Your capacity management solution should provide the ability to report on these metrics in an easy to create, repeatable format that can be shared regularly with management. Time spent focusing on your reporting strategy and process will pay dividends for you in being more effective and having stronger management support for your efforts.

Effective reporting can highlight the ROI benefits to management, such as:

- Growth in resource utilizations
- Count of hardware/software assets
- Cost of hardware/software assets
- Maintenance costs for hardware
- Capacity related incidents
- Average cost of incidents
- Risk of capacity related incidents

Conclusion

Cloud computing continues to grow. Today, the cloud services multiple organizations, users, services and applications. Effectively managing cloud capacity to avoid extra spending must be a focus for today's companies. Most organizations are switching to the cloud and enjoying its low-cost advantages, but at the same time, they need to establish processes and controls to ensure they are efficiently utilizing all the benefits of the cloud, which means capacity planners will need to update their skills. Today, environments are far more complex and there is an increasing need for capacity planners to keep up-to-date with how the cloud has changed the IT landscape.

precisely