Cloud Capacity Management

WRITTEN BY ROBIN REDDICK, SR. MANAGER, BMC

CONTENTS

WHAT IS CLOUD CAPACITY MANAGEMENT AND HOW DO I GET THERE?

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 modern digital businesses, capacity and cost management are essential to ensure adequate resources and budget, whether on-premises or in the cloud, to support new, existing, and growing business services. During cloud migration, right-sizing resources before the move to cloud helps prevent overprovisioning, unnecessary operating expense, cloud sprawl, and excessive management complexity. Performance benchmarking helps you ensure that cloud resources will provide the same or better performance as on-premises resources.

Capacity management continues to play a critical role for on-premises resources, as well. 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.

Across both cloud and on-premises resources, capacity management informs forecasting and planning by helping you determine the capacity levels that you'll notice across your environment, including compute configurations, storage, database, and network bandwidth, as well as the most cost-effective way to provision them.

In this Refcard, we'll look at what it means to extend capacity management to the cloud: what it takes, how it differs from traditional on-premises capacity management, and how to apply it in key use cases.

WHAT THE CLOUD MEANS FOR CAPACITY MANAGEMENT

Capacity management had a long history in IT well before the advent of cloud. In the mainframe era, capacity management was absolutely vital — you had to order your mainframe a year in advance, so you had to know exactly what you would need to ensure performance and availability. Otherwise, you risked maxing out your resources, getting caught short, and not having enough MIPS to support business-critical jobs that needed to be run.

With the rise of easy-to-deploy and cheaper distributed servers and virtual machines, many organizations allowed capacity management to lapse, or moved to a performance management approach where performance issues are used to flag capacity issues. Of course, this also meant accepting the high cost of inefficient provisioning, as VMs proliferated throughout the environment without a clear understanding of the capacity or utilization of each server. Over time, as this sprawl bloated capital expenses, many organizations returned to capacity management at some level, whether using a formal tool or informal spreadsheets, notes, and approximations.

bmc
MULTI-CLOUD

All your clouds now simplified.

Start Simplifying >

Now, the cloud has greatly increased the complexity of the IT environment at most organizations. According to ESG, 81 percent of enterprises use two or more public cloud environments, and 51 percent use three or more. Only 16 percent are limited to either on-premises resources or single public cloud environments. It's more challenging than ever — and more important — to manage capacity across this more complex environment, and to achieve complete and holistic visibility to ensure that each service has the capacity it needs. Capacity management also enables informed decisions about which apps, services, and workloads make the most sense to move to the cloud, as well as about the right way to make those moves. Visibility into what you have, what you're using, and what you're paying for makes it possible to manage costs and avoid bloat.

THE CAPACITY MANAGEMENT LIFECYCLE

STEP 1: IMPORT DATA

Data is power: there's not much you can do without it. A critical first step in capacity management is to import metrics for performance, capacity, and configuration, as well as business KPIs, for resources including:

- Physical/virtual/cloud infrastructure
- Databases
- Storage
- Networks
- Big data environments
- Facilities

There are a variety of ways to collect this data, including importing it from real-time monitoring tools, industry-standard ETL extraction, or direct API integration. You'll also need to determine the frequency and granularity by which you want to collect the data; most organizations do so every 24 hours. The more data you collect, the more accurate the underlying information becomes, in turn enabling better insights through complex analytics. This helps you make better business decisions and become more proactive as an organization.

Gathering performance data is only half the equation for a fully mature capacity management lifecycle. You'll also need business service models, likely populated from some kind of discovery solution into a configuration management database (CMDB). Discovery tools provide organizations with a full inventory of their assets, both known and unknown. Often, the discovery solution can map applications, as well. This allows necessary insight into which applications are using which infrastructure, as well as whether certain dependent applications require proximity for better performance. Best practices today leverage tags identified as configuration items (CIs) in the CMDB as filter criteria while building analyses, models, reports, and dashboards.

Using a tagging methodology is another way to get service views, and one that is being encouraged by cloud service providers. With a good tagging methodology, organizations can create custom views of data that meets the needs of various stakeholders with a need for visibility into on-premises and cloud resource usage and costs. Typical tags include categorization by department, data criticality, compliance, instance types, clusters, user groups, and so on. Tags can be applied as resources and provisioned, but you'll also likely need to define and apply additional tags over time using your capacity management application.

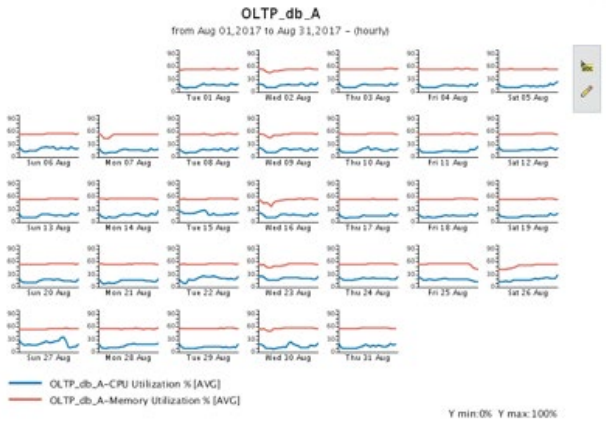
It is the responsibility of the capacity management application to marry the IT and business sides of the equation together. This will elevate a capacity management practice from simple siloed infrastructure capacity management to a more mature service-level capability, allowing for advanced modeling techniques such as modeling changes to service demand.

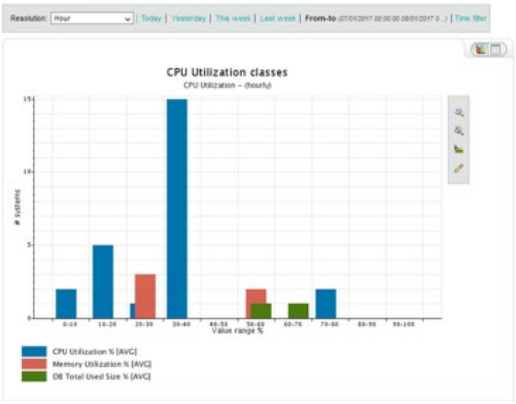
STEP 2: ANALYZE DATA

Now that you have data, you also need visibility into your assets to understand what is actually going on. Many organizations lack visibility into their business services because their business is organized into technology silos managed by multiple monitoring tools, each with its own user interface. Leveraging a solution that extracts and organizes this data in one location is critical. This brings us to the second step: data analysis.

Utilization analysis should be performed from several perspectives, as follows.

- **Visibility**: Visibility across your environment is the foundation of any capacity management process. Analyze your discovered data to gain insight into what assets you have, how they are configured, and where they are located.
- **Baselining**: Next, profile normal utilization profiles and baselines (this step requires machine learning). You'll need to understand usage patterns over time and identify the types of cyclical behavior that exist in the organization and their causes. The longer the period of time you analyze and the more data you collect, the more accurate your baselining and profiling becomes. Ongoing data collection and analysis is the key to proper profiling and baselining.



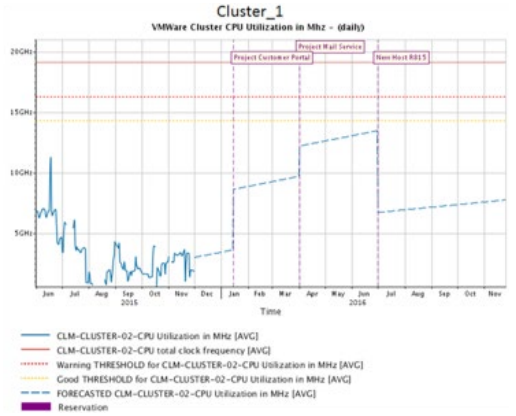


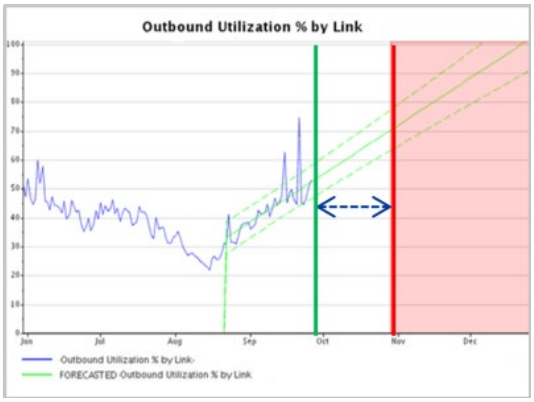
An understanding of resource usage patterns over time helps you determine the capacity levels needed to ensure consistent performance.

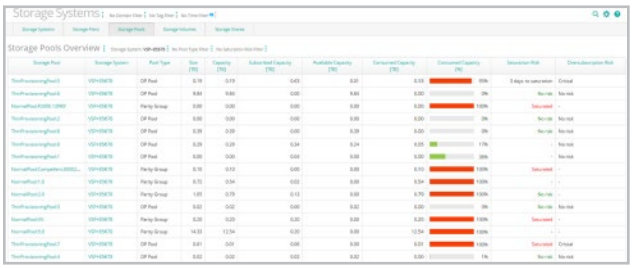
- **Peak analysis**: Identify periodic behaviors and busiest periods. Understanding when changes in workloads occur is essential to efficient use --- especially in the cloud, where you're paying for resources on a daily, hourly, minute, or second-by-second basis. By understanding these behaviors, you can make better and more informed decisions on how to handle and resource applications to ensure performance without wasting resources.
- **Optimization**: Look for opportunities to optimize your use of resources. This may involve adapting the configuration of compute to changes in workloads such as adding memory or CPU. This requires automation to be done effectively; manual efforts are typically 30 days out-of-date and can't hope to keep up with the pace of change in the modern enterprise.

STEP 3: FORECAST DATA

Armed with an understanding of what you have today and how your resources are being used, you can be more proactive in managing your environment by predicting future utilization, as well as potential capacity constraints or saturations. This knowledge can help you prevent service degradation and prevent potential outages. Forecasting also allows for an understanding of how future configuration changes will affect current and projected performance, another critical aspect of the capacity management process.







Forecasting allows you to anticipate the impact of future configuration changes on utilization levels and flag anticipated saturation points before they impact performance.

To proactively identify storage capacity saturation:

- Identify when storage pools may run out of capacity.
- Quantify the additional capacity required to meet allocation requirements.
- Verify whether there are enough unused disks in your storage systems to extend existing storage pools.

This process makes it possible to avoid under-purchasing and meet both current and future storage requirements so that you can prevent downtime. At the same time, accurate sizing helps you avoid over-purchasing and wasted storage capacity.

STEP 4: PLAN WITH DATA

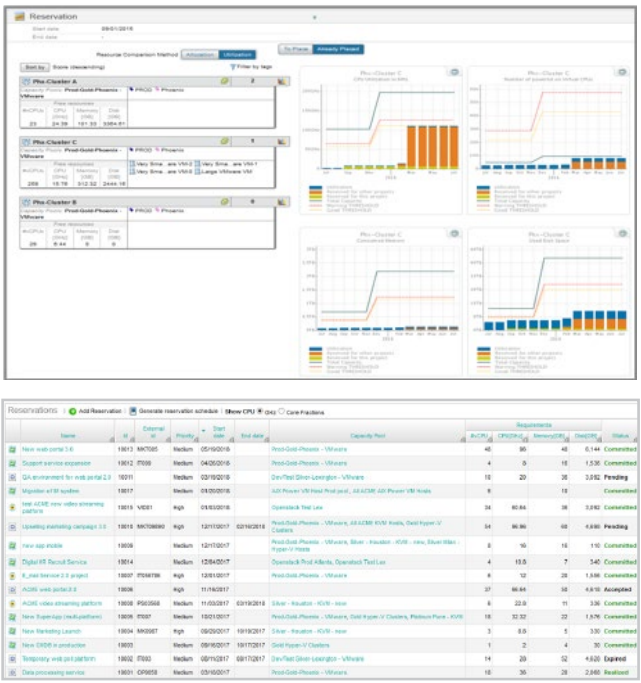
Now that you understand projected organic growth on existing systems, applications, and business services, you're ready for Step 4, which centers on planning for new projects, applications, and business services. This is often referred to as demand management or reservation-aware capacity management.

In this step, organizations focus on two questions:

1. Do I have enough capacity for these new projects?
2. How will these new projects affect the other applications and business services currently running?

Capacity management data can be fed into a resource reservation dashboard to provide answers to questions including:

- What do you have and how are you using it?
- What is being on-boarded and when?
- Do you have existing resources or are you adding new?
- What are you off-boarding and when?
- How much capacity does this free up?
- When will resources be reclaimed and added back to available resource pool?



A reservation dashboard provides clear visibility into the resources required by each service, when they'll be needed, and whether they have been committed.

STEP 5: PREDICT CHANGES AND RECLAIM CAPACITY

The next step in the evolution of a capacity management practice evolution is predicting the impact of changes in service demand on existing systems, applications, and business services. This is often referred to as queueing network modeling at the business service level, or as optimization of IT infrastructure resources (both compute and storage).

In this step, capacity managers simulate system changes made necessary by specific business scenarios. For example:

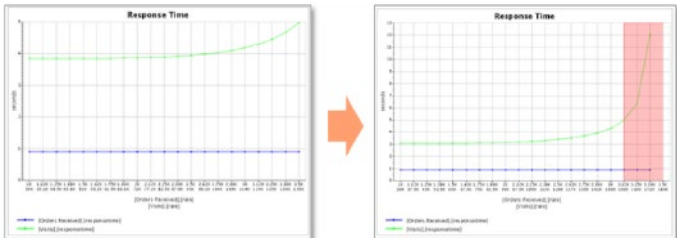
- Simulating the impact of IT infrastructure changes in tandem with business growth on calculated response time and resource utilization constraints.
- Simulating consolidation and virtualization scenarios to identify how potential changes would postpone or eliminate saturation.
- Simulating service impacts resulting from a disaster recovery scenario or the decommissioning of resources as part of a cloud migration initiative.

Capacity management makes it possible to predict the future behavior of system resources in scenarios such as these and many others, as well as the resulting impact on business KPIs. This helps IT correlate business needs to capacity demand and align resources as needed to support them. If an upcoming event could change application resource needs, you can model and plan accordingly. For example, an insurance company may need additional resources to support an open enrollment period. Universities need more resources at the beginning of the school year to support student enrollment. Retailers need resources to support Black Friday, Cyber Monday, campaign and product launches, and other sales events. Every business has events that dramatically change workloads for selected applications. These applications are typically customer-facing, critical to business, and delivered using resources spanning multi-tiered and shared environments.

The Black Friday example clearly illustrates the value of this level of capacity management maturity. In some cases, more than half of an organization's annual revenue is generated during the year-end holiday season, which now also includes Cyber Monday and Cyber Week. If a retailer's website goes down or lags during the all-important year-end holiday shopping season, consumers defer quickly to a competitor's website and organizations not only lose a sale but also a customer. Capacity modeling can prevent these resource shortages from happening.

In the screens below, we see that roughly 5,000 people visit an organization's web page every hour, generating 1,000 orders. We want to know whether the current shared, multi-tiered environment can handle an increase in web traffic because our organization is running a promotion and the business expects 5x the usual traffic and 3x the usual order volume. If there is a constraint, where will it be? Will there be only one? How can we correct the constraint in order to support the change in service demand? This would be difficult if not impossible to determine without an effective capacity management practice, likely a matter of rough estimation or sheer guesswork.

As the anticipated surge in demand approaches, we need to make sure that we have enough capacity to handle it without jeopardizing our calculated response times. In addition, we need to know where in our shared, multi-tiered infrastructure the constraint(s) will be, and what we need to do to change our environment in order to support the increase in service demand.



Capacity managers can ensure adequate capacity for anticipated surges by simulating the impact of service demand changes as well as various configuration or capacity modifications made to address them.

By modeling the impact of these service demand changes, we can estimate saturation and capacity constraints as well as understand what configuration and or capacity modifications are needed to alleviate the constraint. We know exactly what we need to do and when we need to do it in order to support the business.

Other questions that can be addressed at this step include:

- How many additional VMs can we still deploy?
- Which is the best cluster to allocate them?
- Are our availability zones close to saturation?
- 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?

For cloud resources, capacity management can clarify the following:

- 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 for a given application?
- How much will these changes cost?
- Would it be cheaper to move to a different cloud vendor?

STEP 6: REPORT WITH DATA

After data import, visibility, analysis, forecasting, planning, and modeling, the next step toward capacity management maturity is the ability to automatically generate reports and dashboards that can be distributed to stakeholders. These stakeholders can include personnel responsible only for individual technology silos, the health of the business, specific applications, or a cross-section of all of the above. As a result, it is important to automatically generate a variety reports and views with different content for each stakeholder on a periodic basis. This can also include generating exception-based reports or showback reports.

CAPACITY MANAGEMENT USE CASES: WHY AND HOW

MANAGING CLOUD CAPACITY

Preventing cloud 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. To get the most for your cloud spend while ensuring a good experience for customers and business users, you can use your capacity management tool to:

- Scan your current environment usage for configuration corrections that can be applied to improve the performance of cloud-based services.
- Identify possible configuration remediations to enable performance improvements.
- Scan cloud resources for additional opportunities for efficiency or performance enhancement, such as identifying resources not properly decommissioned or resources still available but not in use.
- Create new polices based on data-driven recommendations; for instance, sizing, unused or overprovisioned capacity, and so on.

MIGRATING ON-PREMISES RESOURCES AND APPLICATIONS TO CLOUD

Before any cloud migration initiative, you need to understand how you have your on-premises infrastructure provisioned and the usage patterns of those workloads so that you can make informed decisions about the cloud resources you'll need, including type, size, and configuration. For example:

- Before moving any service or application to cloud, it's crucial to clean up on-premises usage in terms of efficient resource utilization so that you don't end up overprovisioning and overpaying.
- Many on-premises applications aren't engineered to take advantage of modern technologies in cloud platforms. This can lead to less resource-efficient or process-efficient services that in turn cause overprovisioning and high operating costs.
- As you see how usage is changing as you move to the cloud, you can look at before-and-after metrics to guide decisions about which on-premises resources to decommission while also looking ahead to determine how to make the best use of your available cloud infrastructure.

OPTIMIZING AND CONTROLLING COST

Effective cost control and optimization enable IT to deliver greater return on investment for the business while ensuring that funds remain available for innovation. To this end, you can use your capacity management tool to:

- Monitor metrics such as total daily spend; daily spend per resource type (VM, database, storage, and so on); month-to-date spend, usage, and compute hours per service; and monthly and annual spending forecasts. This data can be used to track budget spending and identify any significant changes in spend.
- Perform a cost comparison for an active workload to see how its costs would differ between instance types and platforms.
- Create a policy to pause or terminate an instance based on use and spending thresholds.

TYING THE CAPACITY OF COSTS OF IT SERVICES TO BUSINESS SERVICES

A business-centric view of capacity and utilization can help you ensure that IT spend is aligned with business priorities. This process includes:

- Defining business services in terms of the IT services that power them.
- Understanding on-premises and cloud capacity by IT service.
- Tracking resource utilization and cost by business service.

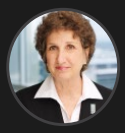
GENERATE COST MANAGEMENT REPORTS

To optimize spend, avoid waste, and align investment with evolving needs and priorities, it's important to gain clear visibility into the true cost of each service. A capacity management tool can capture the data needed for both IT and lines of business to make cost-effective decisions based on the relative costs of various technology options, and the relative profitability of the business services they power. This can include:

- Associating IT costs to applications, business services, departments, or customers.
- Simulating the IT cost impact of infrastructure or cost model changes.
- Generating cost breakdown reports to share with stakeholders to help them understand how the IT budget is being spent.
- Using service cost data to charge internal groups, business units, partners, or external customers according to their resource utilization to recapture spend and/or drive organizational changes by incentivizing people and business units toward more optimal decisions and behaviors.

CONCLUSION

To deliver the greatest value for the business, IT needs to strike the optimal balance of resource capacity, cost, and utilization across both on-premises and cloud environments. By ensuring adequate capacity for high-quality service delivery while avoiding waste, you can get maximum return on spend while avoiding the downtime and disruptions that can send customers elsewhere. A holistic approach to capacity and cost management provides the visibility, insight, and control you need to keep your business running at its best.



Written by Robin Reddick, Sr. Manager, BMC
Robin Reddick is Sr. Manager, Solutions Marketing for Digital Services Operations products at BMC. Her areas of expertise are capacity management and cloud expense management.

DZone, Inc.
150 Preston Executive Dr. Cary, NC 27513
888.678.0399 919.678.0300