6 Steps to Successful Capacity Management for Retailers

How retailers deliver high-quality digital services while avoiding wasteful spending
Why capacity management matters—more than ever

Today, retail success depends on digital excellence. To compete, differentiate, and win, companies need to deliver the high-quality services their customers’ demand and they must maintain this reliability no matter how fast or how often things change. With IT infrastructure budget now spanning both operational and capital, it’s also essential to get the most possible out of every dollar spent.

Capacity management provides the visibility and insight to keep your IT resources closely aligned with business needs. According to Gartner, approximately 28 percent of server capacity currently goes unused, as well as 40 percent of storage. By taking a holistic approach to planning, management, and optimization for both cloud and on-premises infrastructure, you can cut that waste and recapture that spend.

That’s especially important in modern hybrid cloud environments, which are especially prone to overprovisioning, sprawl, management complexity, and unnecessary operating expense. According to ESG, 81 percent of enterprises use two or more public cloud environments, and 51 percent use three or more.

With capacity management, you can make 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 cost and avoid overspending.

In this ebook, we’ll explore the practice, value, and steps for successfully managing your IT infrastructure – regardless of whether they are on premises or in the cloud.
Step 1: Collect and visualize data

Data is power. A critical first step in capacity management is to collect metrics for performance, capacity, and configuration, as well as business KPIs, for resources including:

- Physical/virtual/cloud servers
- Databases
- Storage
- Networks
- Big data
- Facilities

There are a variety of ways to collect this data, including importing it from real-time monitoring and log tools, using industry-standard ETL extraction, or through API integration. Data collection frequency and granularity are typically determined by the criticality of the service. Most organizations collect data hourly or daily for most services and applications, and as frequently as 1 second for highly critical services and applications.

Connecting resources with services

To unlock the full value of the performance data you gather, you’ll need to maintain an up-to-date inventory of your infrastructure resources as well as business service models showing how they’re used by specific applications. You can use a discovery tool like BMC Discovery in tandem with a configuration management database (CMDB) to capture and maintain inventory data and application maps.

Tagging the configuration items (CI) in your CMDB makes it simple to filter them while building analyses, models, reports, and dashboards. Tags can be applied as resources are provisioned, but you’ll also likely need to define and apply additional tags over time.
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 they are organized into technology silos and use a variety of disconnected solutions. For accurate insights, data needs to be maintained in one location. This brings us to the second step: data analysis.

Your utilization analysis should be performed from several perspectives:

- **Visibility** – As an essential foundation for capacity management, analyze your discovered data to gain insight into what assets you have and where they are located.

- **Baselining** – To understand typical usage patterns and cyclical behaviors, use machine learning to analyze data for creating baselines and profiles of normal utilization. The longer the period of time you analyze and the more data you collect, the more accurate your baselining and profiling becomes.

- **Peak analysis** – Understanding when changes in workloads occur is essential to efficient use—especially in the cloud, where you’re paying by the day, hour, minute, or second. By understanding periodic use and busiest periods, you can make better decisions to ensure performance without wasting resources.

- **Optimization** – Fine-tune the use of resources on an ongoing basis. Automated analysis and recommendations is critical; manual efforts 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.

For example, to proactively identify storage capacity saturation:

1. **Identify** when storage pools may run out of capacity
2. **Quantify** the additional capacity required to meet allocation requirements
3. **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 you can prevent downtime. At the same time, accurate sizing helps you avoid over-purchasing and wasted storage capacity.
Planning for new projects, applications, and business services ensures you have the resources needed for both current and future projects. This is often referred to as demand management or reservation-aware capacity management.

In this step, organizations focus on two questions:

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

Capacity management data can be visualized 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?
Step 5: Predict changes and reclaim capacity

To maintain service quality, it’s essential to be able to predict the impact of changes in service demand on existing systems, applications, and business services. You can use TrueSight Capacity Optimization to simulate system changes made necessary by specific business scenarios.

For example:
- The impact of IT infrastructure changes in tandem with business growth on calculated response time and resource utilization constraints
- How various consolidation and virtualization scenarios could postpone or eliminate saturation
- Service impacts of a disaster recovery scenario or the decommissioning of resources as part of cloud migration

By predicting the behavior of system resources in scenarios like these, as well as the impact on business KPIs, organizations can ensure adequate capacity for critical high-demand periods.

For example:
- Insurance companies entering an open enrollment period.
- Universities enrolling students at the beginning of the school year.
- Retailers supporting Black Friday, Cyber Monday, campaign and product launches, and other sales events.
The scenario modeling capability of TrueSight Capacity Optimization can have a truly transformative impact on IT operations.

Imagine having answers to questions like these at your fingertips:

- 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 our most constrained resource and the most likely to impact our services?
- How much spare capacity do we have, and when will we run out?

For cloud resources, capacity management can clarify:

- Do we need to buy more VMs, increase or decrease our existing ones, or change their 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?
An example: Staying out of the red on Black Friday

If a retailer’s website goes down or lags during the all-important year-end holiday shopping season, consumers defect quickly to a competitor’s website and organizations not only lose a sale, but also a customer. Here’s how to keep that from happening.

Let’s say 5,000 people/hour visit an organization’s web page, generating 1,000 orders. Will the current shared, multi-tiered environment be able to keep up when a promotion yields 5X the usual traffic and 3X the usual order volume? Will there be any constraints that jeopardize response times? What changes can we make to alleviate them?

By using a TrueSight Capacity Optimization to model the impact of these service demand changes, we can estimate saturation and capacity constraints, and take active, informed steps to better support the business.

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.
Step 6: Report with data

Finally, to complete the capacity management lifecycle, you’ll likely want to generate reports and dashboards for distribution to stakeholders. These stakeholders can come in many roles, including those responsible for individual technology silos, the health of the business as a whole, specific applications, or a cross-section of all of the above.

TrueSight Capacity Optimization will let you automatically generate and distribute custom reports and views with different content for each stakeholder on a periodic basis. This can also include generating exception-based reports or show-back reports.
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.

To learn more about TrueSight Capacity Optimization, go to www.bmc.com/capacity
About BMC

BMC helps customers run and reinvent their businesses with open, scalable, and modular solutions to complex IT problems. Bringing both unmatched experience in optimization and limitless passion for innovation to technologies from mainframe to mobile to cloud and beyond, BMC helps more than 10,000 customers worldwide reinvent, grow, and build for the future success of their enterprises.

www.bmc.com