IT Optimization through Predictive Capacity Management

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) White Paper
Prepared for CA Technologies

March 2012

EMA™
IT & DATA MANAGEMENT RESEARCH,
INDUSTRY ANALYSIS & CONSULTING

IT Optimization through Predictive Capacity Management

Table of Contents

IT Optimization through Predictive Capacity Management

Abstract

Virtualization, cloud, and the resulting IT-as-a-Service paradigm constitute the core elements of the new role of IT within the enterprise. No longer is IT regarded as a cost center, but today is seen as a strategic differentiator, supporting and enhancing key business processes. Allowing business stakeholders to take advantage of more and more IT infrastructure elements in a self-service manner has led to a proliferation of virtual and physical infrastructure. Each additional infrastructure item comes with increased CAPEX and OPEX attached. Overprovisioning is often used as insurance against performance problems and downtime, as conventional methods of capacity planning are too static and costly to calculate the optimal balance between load balancing and adding new infrastructure components. To resolve the issue of overprovisioning, capacity management has to replace capacity planning.

Capacity management is based upon three core elements: empirical analytics, policies, and planning. If all three aspects are taken into account, the enterprise benefits from a dynamic capacity management solution that provides specific data center sizing instructions that are founded on future business and compliance requirements. The capacity management solution also exchanges information with other enterprise systems, such as service assurance, workload automation, data center orchestration, and composite application development. Due to this central importance, capacity management should be regarded as part of the backbone of corporate IT, helping the enterprise resolve its most important IT challenge of optimizing resource utilization, while at the same time guaranteeing application performance.

From Capacity Planning to Capacity Management: A Brief History

Capacity Planning

Since the beginning of enterprise computing about fifty years ago, the ultimate goal of capacity planning has been to minimize the cost of IT infrastructure resources, while at the same time ensure application performance and availability. Uptime, application performance, and batch job windows have been the central health metrics for IT since the days of the mainframe. And since that time, the IT department had to address the conundrum of finding the right compromise between overprovisioning as an insurance against SLA violations and achieving high enough utilization rates to keep OPEX and CAPEX low. Overprovisioned data center environments not only require a significantly higher CAPEX, but also lead to an increased OPEX caused by costs for ongoing maintenance, upgrades, power, heat, floor and rack space, air conditioning, and software licensing (see Figure 1).

Overprovisioned data center environments not only require a significantly higher CAPEX, but also lead to an increased OPEX caused by costs for ongoing maintenance, upgrades, power, heat, floor and rack space, air-conditioning, and software licensing.

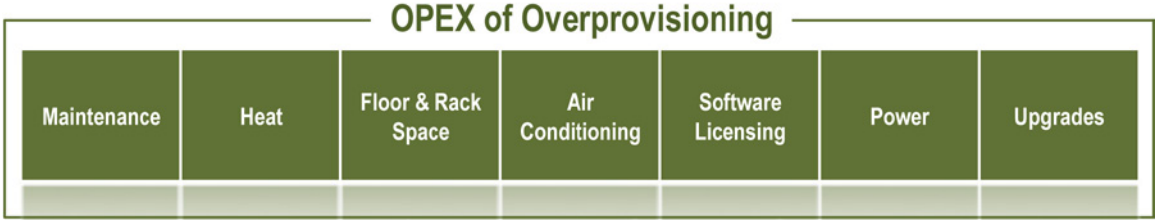

Figure 1: Overprovisioning leads to significant OPEX

While the OPEX and CAPEX attached to overprovisioning has always been a significant pain point for the enterprise, the core instruments for rebalancing utilization rates across the data center were limited. Workloads were dependent on specific operating systems and reallocating existing hardware resources was a major effort involving downtime and risk.

On the other hand, the data center infrastructure was much less complex, consisting of mainframes and few distributed systems. There were only a small number of different hardware types, few IT projects, no self-service portals, and very little involvement of business stakeholders. IT was regarded as a cost center, necessary to keep the organization running. In the best case, line of business staff did not notice the presence of IT. In case of major failures, IT was regarded as the culprit, disrupting operations.

To prevent these disruptions, the IT department engaged capacity planners to create mathematical models to determine future infrastructure requirements. Despite their complexity, these models were often based on incomplete information. The empirical data for capacity planning calculations frequently came from static Excel or Word documents that were dated. Gaps were filled in by rule-of-thumb estimates and assumptions. The entire planning process, including these estimates and assumptions, was more of an art than a science, and mostly not made transparent to business stakeholders who had to approve the budget. These business stakeholders are not interested in details about performance metrics and hardware specifications, but require accurate estimates of the short-, mid-, and long-term investments necessary to ensure SLA compliance and provide the business with a competitive advantage.

Business stakeholders are not interested in details about performance metrics and hardware specifications, but require accurate estimates of the short, mid, and long term investments necessary to ensure SLA compliance and provide the business with a competitive advantage.

Due to the labor intensity of the manual capacity planning process, only large business initiatives would be included and the modeling effort only happened once every year. In reality, we often saw utilization rates of around 20%, despite this resource-intensive planning procedure.

The New Paradigm: IT-as-a-Service

Through the rise of data center virtualization over the past five years, and ultimately through the dawn of the cloud, IT is no longer viewed as a cost center, but as a vital source for providing the organization with strategic differentiators. Virtualization enables the organization to break down traditional silo solutions, by allowing individual workloads to almost freely move to the place where they can be processed in the most efficient manner. Certain workloads can even be "bursted" to the cloud, for additional scalability. At the same time, resources allocated to specific workloads can be dynamically scaled and reconfigured, without complex migration procedures or downtime.

This radical emphasis on business requirements eventually resulted in the IT-as-a-Service paradigm. IT-as-a-Service is focused on the core concept of providing end users with a self-service catalog of resources and services.

Business users can now quickly create complex proof-of-concept environments, based on building blocks taken from the corporate service catalog (see Figure 2), to accommodate the deployment of new applications, setting the organization apart from the competition. New technologies are constantly

added to the mix, often with minimal involvement of IT. In many cases, the IT department is not aware of what workloads are processed and where. As resources are now more easily available to the end user, IT has become even more business critical. The 24x7 availability is mandatory and application performance plays an increasingly important role.

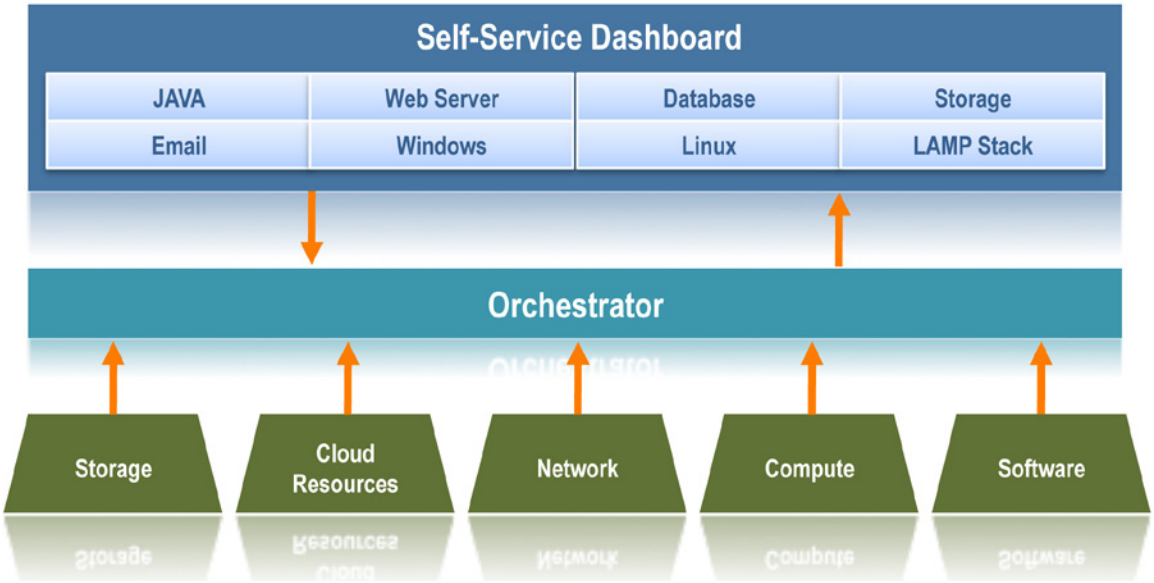

Figure 2: IT-as-a-Service

Virtualization and the self-service approach to IT often lead to end users requesting more resources than they really need. To be on the "safe side", many end users liberally assign RAM and CPU cores to their virtual machines, without worrying about IT infrastructure considerations. A large share of these overprovisioned servers ends up unused, but still requiring maintenance, power, cooling, etc. This intentional overprovisioning through the service catalog, in combination with the problem of many of these resources not being used at all, has a sobering effect on the overall utilization rate of most virtualized data centers. Often this utilization rate is similar or the same as in pre-virtualization days.

To be on the "safe-side", many end users liberally assign RAM and CPU cores to their virtual machines, without worrying about IT infrastructure considerations.

The Emergence of Capacity Management

The static and costly capacity planning process from pre-virtualization days is not able to support the constantly changing requirements of the IT-as-a-service age. Today, file, email, application, and Web servers are virtualized and moved within the data center and to the cloud, as necessary. To anticipate short-, mid-, and long-term capacity requirements, IT must find a way of utilizing monitoring data to obtain a baseline and analyzing historic trends. At the same time, IT has to be aware of policies and plans to create accurate what-if scenarios based on success and growth of the various business units. Modeling these scenarios founded on future application workload requirements, hardware and software replacement or upgrade decisions, or general business growth, provides IT and business stakeholders with the information necessary to make optimal purchasing decisions. Instead of spending weeks on

manually calculating capacity models, the goal of capacity management is to provide a real-time data-driven platform that allows IT to model scenarios and pass on the corresponding reports to business decision makers. Capacity planning has turned into capacity management (see Figure 3). Capacity management systems constantly and automatically ingest configuration and performance data from other enterprise systems. This empirical basis allows IT to provide line of business stakeholders with accurate dollar amounts required for IT to support different strategic and growth scenarios.

Capacity Planning	Capacity Management
Static	Ongoing
Expensive & slow	Software-driven & dynamic
Inflexible	Easy "what if" scenarios
Manual	Automatic
Word & Excel based	Enterprise data driven
Targeted towards engineers	Targeted towards business stakeholders
Requires mathematician	Easy-to-use software
Fragmented	Single pane of glass

Figure 3: Capacity Planning vs. Capacity Management

For the ultimate degree of resource optimization, all IT infrastructure elements – mainframe, physical, virtual, and cloud – have to be viewed through one single pane of glass. Capacity management solutions deliver this single pane of glass, providing IT with all the information needed to match current and future workloads with the most efficient processing resources. This matching process considers possible interactions between workloads competing for the same resources.

The Three Pillars of Capacity Management: Analytics – Policies – Planning

Continuous and dynamic capacity management relies upon three main pillars: analytics, policies, and planning. Only when all of these three factors are included in enterprise IT capacity calculations, can we be confident in the accuracy of our planning scenarios – **for example time-to-live calculations.**

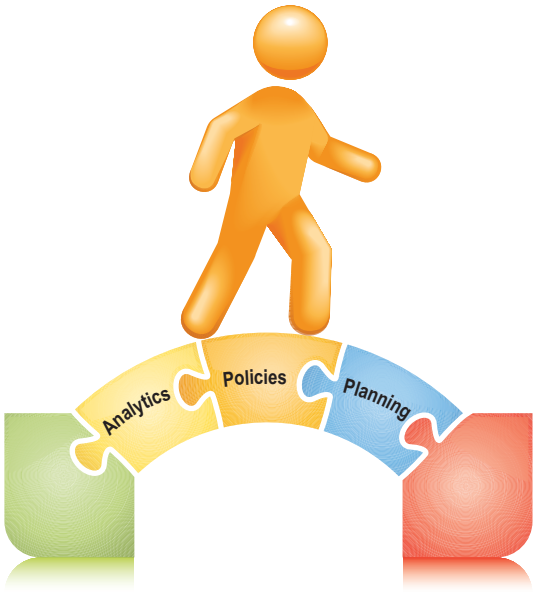

Figure 4: Three Pillars of Capacity Management

Analytics

Real-time configuration, performance, and utilization data constitutes the baseline for any accurate capacity calculation. This data often is available from multiple enterprise applications sources, such as CMDB/CMS, APM, and systems management tools. When analyzing data, it is essential to not only take a look at peak and average activity, but also include mid and long-term resource usage patterns. This information helps the IT department understand how workloads are currently growing

and how they might contend for resources in the future. To ensure accuracy, capacity management software must include algorithms to validate the quality and completeness of the imported data.

Policies

Policies define infrastructure rules and requirements that are based on business considerations for the safe operation and efficient placement of workloads. Availability and disaster recovery requirements, regulatory compliance, resource reclamation guidelines, VM density and performance, and any type of preprogrammed automated actions, can be included in the organization's rulebook. To prevent resource overprovisioning, different policies can be applied to different types of data center elements. For example, a development environment has fewer performance and availability demands, compared to production setups.

Planning

Without including predictions regarding future business user demands, the entire capacity planning effort would occur in a vacuum and only produce marginally useful results. Therefore, the impact of IT resources on central business processes has to be made transparent. Then, assumptions regarding planning parameters such as number of users, operational behavior, software upgrades and implementations, growth of business transactions, retirement of specific infrastructure elements must be derived and considered. Finally, a scenario analysis has to be performed, in order to examine the impact of different business scenarios on the IT infrastructure. For example, the organization may want to look at how different growth scenarios affect individual business processes and the applications supporting these processes. The planning process has to be made transparent to business stakeholders, as technology decisions often have a long-term strategic impact on the entire organization.

Capacity Management: A Proactive Approach

The key challenge of capacity management is to anticipate the IT impact of future business requirements. Capacity management aims at preventing the occurrence of IT bottlenecks, by leveraging existing capacities through strategic load placement, and through the acquisition of additional infrastructure capacities. The more solid the capacity management setup, the more accurate its predictions, and the higher the utilization rate for the entire data center. If the organization fully trusts the predictive analytics capabilities of its capacity management system, there is no need for overcapacities as an insurance against performance and availability issues. These utilization gains can easily be translated into concrete dollar figures and presented to senior management as a business case for the acquisition of new data center infrastructure.

If the organization fully trusts the predictive analytics capabilities of its capacity management system, there is no need for overcapacities as an insurance against performance and availability issues.

Today, capacity management systems have evolved into business and IT planning tools, facilitating the calculation of CAPEX and OPEX of new business projects with IT impact and the transparent presentation of these calculations to business stakeholders. Instead of burdening the IT department with OPEX and CAPEX that is caused by business projects, business stakeholders can now be charged for the costs caused by their projects, in a precise manner.

Capacity Management as Part of Automation

Capacity management software constitutes a central hub where actionable IT metrics are received, stored, and processed. Therefore, capacity management is located at the heart of enterprise IT automation, with direct impact on the following four disciplines (see Figure 5):

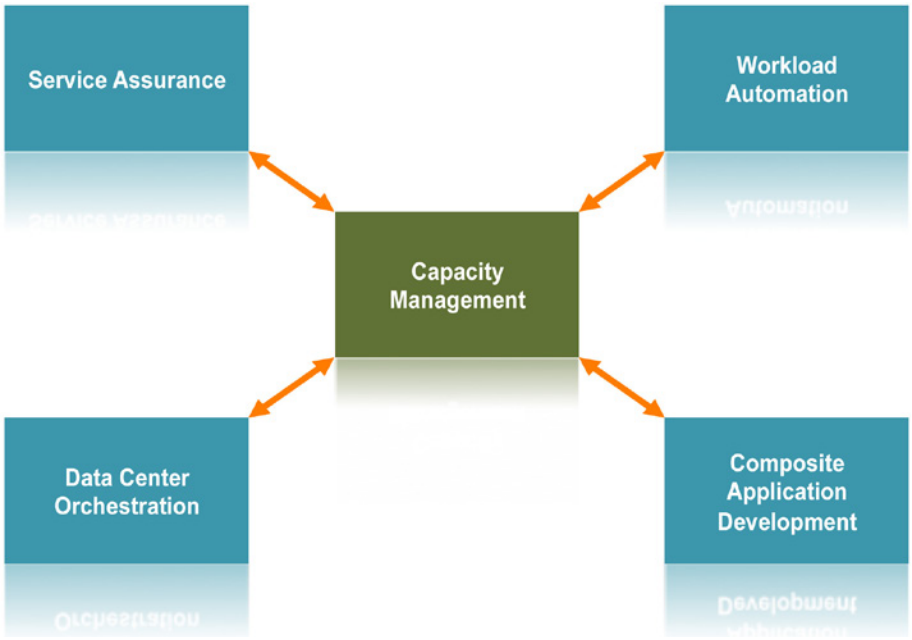

Figure 5: Capacity Management, Located at the Heart of Enterprise IT

Service Assurance

Service assurance software monitors and manages applications, infrastructure, and service performance within the enterprise. Metrics collected by service assurance solutions can be used as input for capacity management software. Capacity management software is then able to include the end-user-experience into its calculations. This is essential, as the true goal of IT infrastructure planning is to calculate the type and amount of resources needed to make the end user as productive as possible.

Data Center Orchestration

Data center orchestration suites are designed to help the enterprise leverage its current IT infrastructure in the most effective manner, by automating the deployment, configuration, and scaling of resources and processes. The orchestration software can receive instructions from the capacity planning system to proactively re-balance workloads and automatically provision physical, virtual, and cloud resources.

Workload Automation

Modern workload automation solutions are business process and resource aware. Dynamic workload placement capabilities ensure that workloads are load-balanced, based on business requirements, and resource availability. Critical path monitoring reports provide an instant overview of the root cause of potential or actual SLA violations. This monitoring data is passed on to the capacity management solution, so that potential SLA-relevant bottlenecks can be proactively addressed.

Composite Application Development

Integrating capacity management with composite application development solutions such as ITKO LISA (acquired by CA in 2011) adds another puzzle piece to the empirical data foundation for capacity management. Receiving information on the future production requirements of components currently under development enables the capacity management solution to more quickly react to potential resource contention and to include infrastructure requirements into future hardware and software acquisition plans.

Automatically exchanging data with service assurance, data center orchestration, workload automation, and application development frameworks, provides a solid empirical basis for capacity management software to provide accurate scenario analyses and instigate corrective actions as necessary.

How CA Capacity Management Can Help

CA Capacity Management guides and facilitates the above described integrated and proactive approach to capacity management. CA offers so-called Solution Kits that automate the data collection from industry standard systems, such as BMC Software, HP Operations, Oracle, SAP, IBM WebSphere, and VMware Virtual Center. Once the data is imported, CA Capacity Management performs a data normalization process, in case of multiple data formats, and evaluates the accuracy and completeness of the imported data.

Based on predictive analytics, CA Capacity Management conducts scenario analyses, identifying suitable hardware options to accommodate the outcome of each business case. These options can consist of shifting loads between existing data center hardware elements and recommendations for acquiring additional hardware. CA Capacity Management offers a regularly updated catalog of server, network, and storage hardware by almost any vendor, so that the customer can leverage existing hardware vendor relationships.

CA Capacity Management offers easy-to-read reports for different audiences, such as business stakeholder, IT executives, IT managers, etc. These reports provide evaluations of the current infrastructure efficiency and compare the outcomes – response time, monthly cost, batch windows – of various consolidation scenarios, such as the virtualization of multiple physical hosts using different hypervisor versions. CA Capacity Management can provide these predictions for all of the individual stages of this migration scenario, so that business users can be made aware of an anticipated temporary degradation of service.

CA Capacity Management provides instructions regarding how to optimally place workloads to ensure SLA compliance in terms of availability and performance, as well as a high degree of infrastructure utilization of physical, virtual, and cloud resources. CA Capacity Management constitutes a design tool that can be used by all IT and business stakeholders to obtain a clear representation of how the entire data center could be improved.

Case Study: Major Pharmaceuticals Distributor

A large pharmaceuticals distributor was planning to migrate off of its legacy mainframe system without risking performance issues. The company considered creating a full-scale test lab that would mostly, but not completely, match the production environment. Creating this lab would have cost $5 million and taken several months.

Instead, the pharmaceuticals distributor chose to adopt CA Capacity Management software to evaluate "what-if" scenarios to determine the service level impact of retiring the mainframe. This performance analysis also identified opportunities for load balancing to increase hardware utilization.

At a fraction of the initially estimated $5 million, by adopting CA Capacity Management software, the customer now has a tool in place for any further IT modeling effort.

EMA Perspective

Virtualization and cloud computing have changed the rules of planning data center capacity and architecture. Today, IT is offered more and more as a service, and capacity management must replace capacity planning, as constantly changing business needs require IT to adjust continuously and dynamically.

Even within drastically overprovisioned environments, VM sprawl and VM stall often lead to performance bottlenecks. The IT department must obtain a comprehensive and real-time overview of the entire data center, including mainframe, physical, virtual, and cloud infrastructure elements. It is essential for this type of overview report to be automatically fed by existing IT monitoring and automation solutions. This real-time overview can then be used to derive concrete measures to achieve better resource utilization.

Combining and processing policies, business plans, and empirical data in a real-time fashion constitutes the only viable way to achieve resource optimization. Resource optimization is the holy grail of systems management, as it allows the organization to finally reap the benefits of its virtualization projects.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help its clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals and IT vendors at www.enterprisemanagement.com or blogs.enterprisemanagement.com. You can also follow EMA on Twitter or Facebook.

Corporate Headquarters:
5777 Central Avenue, Suite 105
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

2437.032112