Capacity Planning Fundamentals

Support Business Growth with a Better Approach to Scaling Your Data Center
Executive Summary

As organizations scale, planning for greater application workload demand is critical. IT cannot afford to be seen as a bottleneck to successful business growth. Moreover, the larger the organization and the faster the growth the higher the stakes.

Planning ahead is key, but as this ebook will discuss, equally important is the foundation you use to make projections. The reality is that smarter capacity planning starts with a better understanding of the virtual environments and the interdependencies of data center entities.

Today’s virtual and cloud environments are increasingly complex. That complexity demands a new holistic approach to growing your infrastructure as quickly as your business needs.
Capacity Planning Distilled

Infrastructure Capacity Planning is fundamentally about ensuring adequate compute, storage and network resources to deliver on business and application service levels, based on current and projected growth of existing and possibly newly introduced applications.

This can often be distilled to “buying the right amount of infrastructure at the right time.”

Are you buying the right amount of infrastructure at the right time?
The capacity management challenge can be distilled to “buying the right amount of infrastructure at the right time.”

Buying too much hardware results in the financial cost that comes with over provisioning. Moreover, given the ruthless cost curve of technology you’ve paid too much for hardware that, when finally used, the technology is often obsolete. Buy too late and risk hosted application performance problems, resulting in end user issues.

In the ITIL (“Information Technology Infrastructure Library”) v3 framework, the “Capacity Management” process is broken into 3 sub-processes:

- **Business Capacity Management**: Understand the future needs of users and customers.

- **Service Capacity Management**: Ensure specific IT services delivered can meet agreed upon service levels.

- **Resource Capacity Management**: Make sure underlying IT resources can support these service levels while being used efficiently.
What’s in a Plan?

In a virtualized environment, physical compute, network and storage resources are shared between multiple application workloads and virtual machines (VMs or containers). It is therefore important that the right amount of underlying resources (for example hosts and datastores) are available for current and projected workload resource demands.
The steps to produce a plan include:

- **Understanding resource consumption demands** (peak, average) of existing workloads across the 4 “food groups” of CPU, memory, network and storage, over appropriate time frames.

- **Understanding available capacity** from underlying hosts and datastores in the virtual environment, and possibly the underlying IT infrastructure including physical storage arrays and compute fabrics / blade servers.

- **Factoring in VM placement across hosts and datastores** to maximize infrastructure utilization while meeting desired service levels.

- **Accounting for possible adjustments** in VM resource allocations and sizing (CPU, memory) enabled by virtualized environments.

- **Projecting growth in resource demands** based on growth (e.g. users/customers) of existing applications.

- **Understanding the expected resource “profile”** of net new applications that may be added and factored in to the plan – for example by estimating resource demands of the VMs or workloads making up that application.

In today’s hyper-dynamic virtual environments most capacity plans are obsolete the moment they are completed.

The environment changes constantly and too fast; the analysis cannot keep up.
• Accounting for any capacity “reservations” that have been committed or need to be granted – a “reservation” is typically given to a set of VMs comprising an application that will be deployed at some time in the future.

However, completing these steps in today’s virtual environments is a nightmare. Put simply, in today’s hyper-dynamic virtual environments most capacity plans are obsolete the moment they are completed. The environment changes constantly and too fast; the analysis cannot keep up.

Even with appropriate real-time data, current available methods are often so labor intensive that migration or consolidation plans take an inordinate amount of time to complete. Spreadsheets and constant manual data entry are inefficient. They distract from productivity more than anything else because they are obsolete the moment they are finished.

As a result, virtualization managers often over-provision to meet quality of service requirements—to the detriment of efficient IT management.
What happens when the economic rules of supply and demand are applied in the data center? In economic markets, buyers and sellers exchange goods and services at an agreed upon price.

Consider virtual machines and physical machines in the data center. They are essentially buyers and sellers of resources (CPU, MEM, Storage, etc.). In fact, every entity in the stack is a perfectly rational actor that either buys or sells resources.
When data center entities make independent resource decisions based off of supply and demand within the environment to assure application QoS, it results in a perfect market. Moreover, decisions are driven by application demand, not infrastructure supply.

What that means for the data center is that application performance is assured, while infrastructure utilization is maximized. Turbonomic’s approach puts the burden of application QoS on software. The software continuously analyzes the data center’s market conditions and applies pricing principles. The data center entities determine how to allocate scarce resources among themselves. Turbonomic is essentially an “Invisible Hand” for the data center.

Why does this matter?

With this unique approach leveraging efficient-market principles, Turbonomic has a holistic, real-time understanding of your current capacity needs. As such, it is better equipped to help you plan for the future by connecting planning with real-time operational data from your own environment. It enables you to proactively scale for future workloads based on projections based on real data – yours.
Capacity Planning with Turbonomic

Turbonomic discovers your virtual infrastructure almost instantly. It knows exactly how many physical machines, virtual machines, and storage devices you have and processes the utilization of every resource in real time.
Turbonomic allows you to create plans for a multitude of capacity “what-if” scenarios. It ensures that you always have the right amount of hardware at the right time to assure application service levels while utilizing your infrastructure assets as efficiently as possible.

Examples of “what-if” scenarios include:

- Optimal workload distribution across existing resources
- Changing hardware supply
- Impact of downsizing, or removing resources for a DR scenario
- Optimal workload distribution to meet historical peaks demands
- Projected infrastructure requirements for a new applications
- Merging clusters or data centers

To run these scenarios, Turbonomic creates a copy of your real-time market. It then uses its Economic Scheduling Engine to perform analysis on that market copy. You can modify the market copy by changing the workload, adding or removing hardware resources, or eliminating constraints such as cluster boundaries or placement policies.
As it runs a plan, Turbonomic repeatedly executes shopping cycles on the market copy until it arrives at the optimal conditions that market can achieve. At that point the Economic Scheduling Engine cannot find better prices for any of the resources demanded by the workload. The plan then stops running and displays the results, which include the resulting workload distribution across hosts and data stores, as well as the actions the plan executed to achieve the desired result.

For example, assume you run a plan that adds virtual machines to a cluster. The plan runs repeated shopping cycles, where each entity in the supply chain shops for the resources it needs, always looking for a better price — looking for those resources from less-utilized suppliers. These shopping cycles continue until all the resources are provided at the best possible price.

The results might show that you can add more workload to your environment, even if you reduce compute resources by suspending physical machines. The recommended actions would then indicate which hosts you can take offline, and how to distribute your virtual machines among the remaining hosts.
Conclusion

When evaluating capacity planning solutions for scaling the data center the complexity of virtual and cloud environments cannot be understated. These environments operate on the whims of increasing and dynamic end-user demands.

The continuously changing environment precludes traditional manual approaches to planning for the future. Only software can address this challenge.

The burden of assuring application QoS as you provision infrastructure must be placed on software.
For capacity planning to meet the business needs of proactively scaling for the future, the burden of assuring application QoS must be placed on software. Turbonomic’s economic abstractions break the complexity and dynamism of the data center—a big problem—into a series of resource exchanges between data center entities—smaller, albeit infinite, problems. The software’s algorithms simply process these small problems in perpetuity.

Turbonomic capitalizes on software’s suitability for process and calculation. It attains a better understanding of your environment and how to drive it to a healthy state, which in turn enable a more informed and scalable approach to capacity planning.
About Turbonomic

Turbonomic delivers an autonomic platform where virtual and cloud environments self-manage in real-time to assure application performance. Turbonomic’s patented decision engine dynamically analyzes application demand and allocates shared resources to maintain a continuous state of application health.

Launched in 2010, Turbonomic is one of the fastest growing technology companies in the virtualization and cloud space. Turbonomic’s autonomic platform is trusted by thousands of enterprises to accelerate their adoption of virtual, cloud, and container deployments for all mission critical applications.

© 2015 Turbonomic, Inc. All Rights Reserved. All trademark names are the property of their respective companies.