White Paper

Capacity Planning in a Multi-Cloud Environment

How to monitor resource consumption and availability to optimize capacity

Executive Summary

Capacity planning in a multi-cloud environment is a complex challenge affecting most (if not all) organizations. Cloud based infrastructure has added to the complexity, both in cost and optimization. Organizations must optimize the performance of their hosted applications, but also plan for future growth while justifying cost. This paper focuses on the variables to consider for an effective capacity plan, and explains how monitoring removes the guesswork by providing quantitative data organizations can use to make better decisions about their infrastructure.

Content presented by:

GROUNDWORK
Make IT Easier.

Introduction

Imagine having to explain the large invoice for cloud infrastructure actually represents a savings.

Before cloud computing, IT capacity planning was a mission-critical activity for one simple reason: infrastructure had to be provisioned for IT to function. With the advent of the cloud, infrastructure is rapidly available with the flexibility to accommodate short-term planning. It actually rewards you for not deploying in advance because you don’t have to leverage an existing infrastructure.

The downside of all this flexibility, however, is the surprise cost. The sheer ease of deploying new infrastructure makes it tempting to do so even in excess of actual need. and if you forget to turn down unused systems, they will keep adding to the cost. Add to that the charge over time rather than the sunk cost of a data center server, and you may be surprised when you get the bill. Imagine having to explain that the large invoice for cloud infrastructure actually represents a savings. With cost awareness, these bills are no longer a surprise. Capacity planning is the key to cost anticipation.

It’s still about the cost

Even if you’re not pre-buying infrastructure, you still need to make sure you’re not overpaying. Peak traffic, new releases, and marketing impact are demands that should be accounted for in scaling up or out. Accurate tracking metrics verify that you haven’t over provisioned beyond an acceptable (and hopefully low) threshold.

An optimized infrastructure with good monitoring and tracking measures in place protects organizations from surprise bills, and the total costs will drop accordingly. Costs are justified and predictable. So how do you get there?

Key questions

Creating an infrastructure with enough capacity to handle its supported applications and expected conditions relies upon answering four key questions:

- *Are applications performing according to users’ expectations?*
- *Can they handle the expected traffic?*
- *Is there enough storage available?*
- *Are partner applications and gateways performing to specifications?*

The answers to these key questions comes from data measured via monitoring. To create a capacity plan, you need to determine your applications’ requirements and performance within their capacity limitation. Testing with the same monitoring harness used in production will give you this data. With test monitoring, you can model performance and design a capacity plan for initial deployment, as well as account for additional capacity to be implemented in response to changing conditions. In many cases, scaling up or out can be automated if you have the proper monitoring data.

A typical scenario

Let's say you want to provision cloud infrastructure for the deployment of a new application. It's going to be small at first. You expect the number of users to grow later in the life cycle once you start marketing, and it will grow even larger as partners sign up and co-market the application. Until then, it has to be lightweight and cost effective. Eventually, you think the application will receive a lot of buzz and you'll need to scale up.

One of the first choices you will need to make is whether to use a platform as a service (PaaS) or infrastructure as a service (IaaS) cloud. A PaaS may be easier to deal with from the perspective of writing code, but given your capacity plans (you will not only need to scale up to a larger instance, but also out to more instances at some point), it may not make sense, since PaaS systems typically don't scale out easily. You need data to understand your application's resource-intensity and how it performs when loaded on larger instances or if breaking it into component instances will be needed. Monitoring provides the data that will allow you to make the best choice.

Capacity planning for cloud-based infrastructures is different in three important ways:

It's tiered.

A lot of cloud providers give some capacity for free. This free tier is sometimes misleading, especially when you cross the threshold to paid. When tracking costs, it's good to know when you will cross over into the paid tier. Also, many cloud services have volume tiers, or even spot-instances, in which compute power is used on an as-available basis for a much lower rate. Depending on your workloads, these tiers be more cost effective than you might imagine.

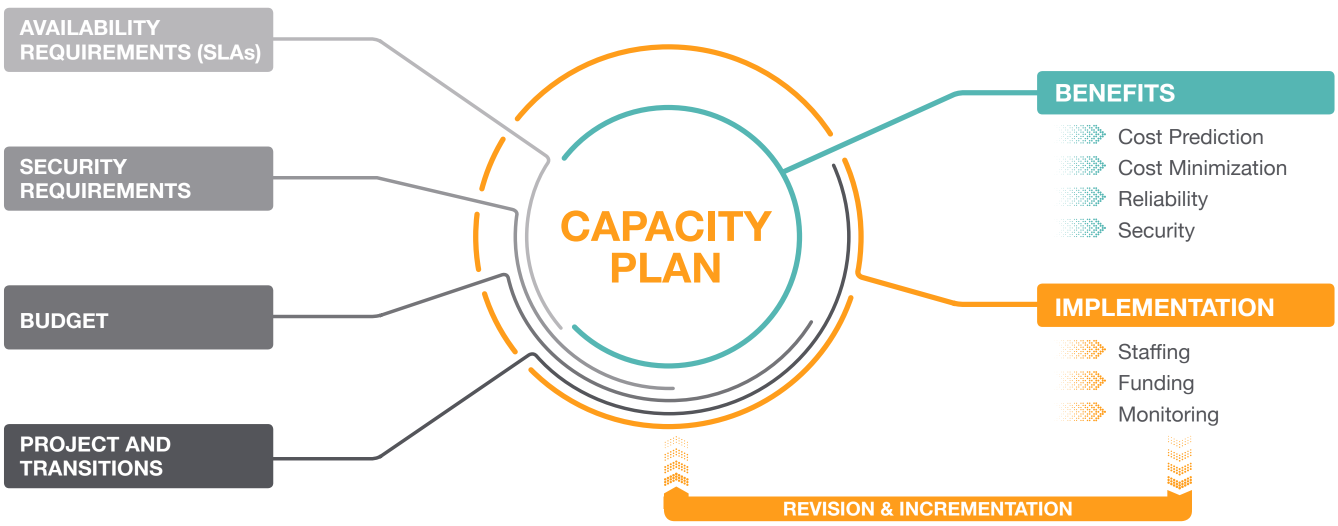
It's shared.

The infrastructure that you allocate from the VM operating system's point of view may not (actually almost never will) match the hypervisor host allocation at any one time. This over-provisioning of resources like CPU and RAM optimizes performance from the hypervisor, but complicates the calculation of actual capacity considerably. Make sure you understand the abstraction of resources your cloud provider uses, be it PaaS or IaaS or a some hybrid.

It's elastic.

Capacity calculation changes when you have an infrastructure that expands on demand. In some sense, capacity is limited by budget alone, which tends to emphasize the importance of creating a plan in the first place; care should be taken when automating the scaling of infrastructures to avoid unintended or uncontrolled up-scaling or out-scaling. Don't forget to regularly check for over-capacity as well, as many organizations experience an effect called "sprawl," in which the ease of ordering and deploying virtual infrastructure results in unutilized or underutilized virtual machines.

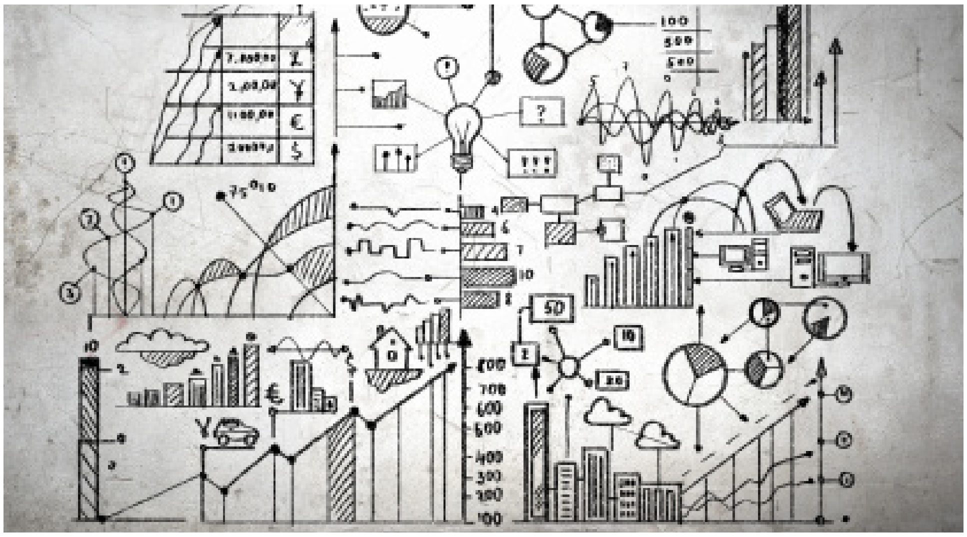
Elements of Effective Capacity Planning



Crucial elements in effective capacity planning for cloud or virtualized infrastructures

First, a good plan is paramount. Project management and planning best practices apply here. You need a timeline, resource estimate, and phases to select technology, allocate resources, and set goals, budgets, and objectives. Consider how to manage the infrastructure on an ongoing basis with processes in place for software updates and marketing events that may spike or dip requirements. An important advantage of virtual infrastructure is the ability to move from an upgrade-in-place model to a parallel upgrade model. Deploying and flipping users to a clone of running software while the old version is still intact is possible in many cases, but doubles the impact on infrastructure.

Capacity planning boils down to a simple formula: input the service level agreement (SLA) and budget, and your plan should produce the schedule. If you don't like the time horizon, lower the SLA or increase the budget. Or go back to the drawing board and select new technologies. That's what a lot of companies are doing now as they migrate from in-house legacy systems to cloud computing and elastic infrastructures.



Monitoring for success

Once you have a plan in place, you then need to track the implementation of its phases. Note that there is really no end date—you will constantly adjust the infrastructure for size, speed, reliability, and a host of other considerations unique to your use case.

While project management software can help plan the timeline and resources, you will need to measure key performance indicators (KPIs) to verify the plan's successful execution. Are you meeting your SLAs? Are you delivering adequate performance and reliability? How hard is your staff working to keep the systems up and running? Monitoring the infrastructure will answer these questions and more.

While it is important to understand the performance of systems used to deliver IT services (process counts, memory utilization, etc.), there are additional measures essential to capacity planning that create a feedback loop. That is, monitoring data showing how a well executed capacity plan meets demand. We call this "monitoring for success." This data can then inform you on how to revise the plan.

It's wise to consider external measures as key in this regard. If you have customer facing systems (internal or external), evaluate them in regards to measuring service delivery. The availability and performance of specific services provided by your application (email, order entry, forecasting, etc.) are as essential to understanding capacity as the measurement of data growth in storage systems over time.

An often overlooked set of KPIs is the impact of a particular infrastructure set on your staff. How hard are you working to maintain the infrastructure? Measures such as staffing levels, utilization rates of external services from contractors or vendors, ticket volume and resolution time, staff turnover and the rates of HR issues all indicate organizational stress levels that drive up costs. "Soft" KPIs like these are often not considered in the overall data gathered from a monitoring system, but interfacing to external systems (like ticketing for example) is generally not difficult, and such measures provide much greater insight into the true costs of a given infrastructure implementation.

Monitoring solutions

A solution like GroundWork offers the best of both worlds, by integrating several open source tools into one unified system.

Knowing what to monitor is the first step in selecting the best tool to fit your organization's needs. In all likelihood, you already have an idea of how you would like to gather data, and which tools would be best suited for your planned environment. But it's always smart to verify the tool selected will fulfill current and anticipated future needs without breaking the budget.

Many companies choose to develop their own methods of measurement. This is especially true for companies in which IT is a core service and delivery. These companies are usually aware of specific metrics that their application produces or is especially sensitive to. While building a scripted monitoring harness to capture such metrics is sometimes necessary, it is also risky. Standardization is important to easily manage and update systems, and to maintain control in the event of staff turnover.

Open source solutions like Nagiostm and Icinga2 are another avenue. These tools are available for free and offer essential measurements that work really well for some organizations. Unless coupled with other systems, however, in themselves they don't offer enough features to adequately measure the breadth of things necessary to monitor for success. Open source solutions are also unsupported, and you will need either in-house staff or a vendor to cover them.

A third option is proprietary monitoring software solutions. These solutions often lack the flexibility of open source, and constrain customization needed for arcane (and proprietary) frameworks. But they offer reporting and dashboarding features that open source solutions lack, and they are supported. The price for some proprietary solutions can be considerable.

A solution like GroundWork Monitor Enterprise offers the best of both worlds, by integrating several open source tools into one unified system. It can be customized to meet your unique requirements using well-known open source standards, and will scale as your organization grows. Its robust dashboards and integration features bring complete visibility into your infrastructure, providing pertinent data to successfully execute your capacity plan.

Most of the time we are faced with short deadlines, high expectations, and little resources. Informational articles (like this one) help frame the challenge of creating a capacity plan and monitoring its success. Remember, capacity planning should function like a formula: put in your constraints, and get back your costs and timelines. Like most things, it's an optimization problem.

GROUNDWORK
Make IT Easier.

Unified monitoring, automation and analysis. Contact us for a 15-minute demonstration today.

www.gwos.com | info@gwos.com