THE VIRTUAL TWIST

Capacity Planning in a virtualized world

sumerian®
Forward Thinking

White Paper

Although capacity planning should be viewed as one of the most important IT Operations Management disciplines to get right, many organizations continue to find it challenging to implement a successful approach. Done well, capacity planning can significantly reduce risk, improve IT performance and save costs. So why do so many IT organizations find it difficult?

Part of the answer is that today's IT environments are incredibly complex, with high rates of change. Of course, the upside of more modular environments and new technologies, such as virtualization and cloud, is that organizations have a new level of flexibility within their IT environments — VMs can be added and changed on a daily basis. The downside is that in these dynamic and heavily virtualized environments it is increasingly difficult to maintain any accurate picture of infrastructure capacity — let alone keep it aligned and optimized to changing business needs. Achieving visibility of resource demand and capacity headroom is a challenge — and without that, it's impossible to accurately predict how much more or less infrastructure is going to be required for the future.

Traditional capacity planning approaches are also one of the limiting factors as they have a tendency to be silo-ed. For example, different teams across the IT organization often manage capacity discretely within their own particular domain or application field. This means it is challenging to gain a joined up view of infrastructure resource utilization and headroom for growth, and inefficiency in the use of infrastructure resources is common.

What is capacity planning?

There are many definitions of IT capacity management and planning — frameworks such as ITIL, CoBIT and others each have their own — but fundamentally they're all about the same thing: ensuring you have enough resources deployed to meet the needs of your customers and the applications they use. In an IT Operations environment the resources you have to manage are typically compute, storage and network capacity, and so the challenge is in determining how much of each is required for the workloads your customers are generating.

It's worthwhile pausing for a moment to consider this from your customers' perspective. They (generally speaking) never know how much resource you have deployed, and in many cases care less. Their interest is much more in the performance of the applications they're using, as opposed to capacity. The two, however, are generally related — to improve performance it's often necessary to provision more compute, storage or network capacity; the challenge is figuring out where.

Why taking a statistical approach makes sense

In many ways, capacity planning is a game of chance: you can never know exactly what behavior tomorrow is going to throw at you — particularly in those complex and heavily virtualized environments we talked about earlier. Like other games of chance, it is therefore best served by a statistical approach — observe what has happened in the past, quantify the probabilities of various events, and then use these to predict the likelihood that certain occurrences will happen in the future.

From the perspective of managing IT infrastructure, this suggests the following approach:

- Capture historical data on the usage of various resources (compute, network, storage, mainframe MIPS and so on);
- Combine this with inventory metadata about resource allocation (which servers sit in which datacenters, which storage LUNs are used by which servers, and so on);
- Use these two datasets to create a statistical baseline of utilization, which is capable of showing the "big picture" of resource usage across your entire estate, but also enables drilling down into finer detail such as usage by particular applications or customer groups;
- Use the statistical nature of the baseline to estimate the likelihood of certain thresholds being breached – for example, decide on a CPU usage threshold you're comfortable with (80%?) and use the baseline to determine

how often you exceed this; if you're over your threshold 50% of the time, that's probably not good, but 5% or 1% might be acceptable depending on the criticality of the environment.

It's this statistical assessment that is key to successful capacity planning. You can't guarantee that you'll never exceed a specific threshold, but you can start to manage the probability of that happening.

The virtual twist

One of the key impacts that virtualization (and its new cousin, cloud computing) has had in the past few years is a demand for (and expectation of) speed and agility in deploying and changing infrastructure. Customers expect that new virtual servers can be deployed, or existing VMs reconfigured, in minutes or hours. While the management and orchestration tools available from various hypervisor vendors support this flexibility, very little thought has been given to the impact on capacity planning.

In addition, virtual environments require significantly more resource sharing than older physical environments — indeed, this is one of their big attractions. But this means that the potential for one workload to interfere with another is significantly higher, and capacity planning is the key mechanism needed to understand and control that risk.

Collecting the right data

As described above, capacity planning is the art of using data about the past, and assumptions about possible changes, to predict the future. The starting point here is to understand what data you have available to tell you what's been happening in the past.

Inventory data

Asset lists, configuration management DBs (CMDBs), server inventories, assorted spreadsheets — the likelihood is you have numerous sources of information about servers in various places in your environment. Bringing these together into a single set of inventory data is crucial for effective capacity planning.

Virtualized environments tend to make this job a bit easier, as it's relatively easy to generate a list of physical and virtual servers and their configuration metadata.

The information that you will have available may vary, but you will typically want at least the following for each server:

- Physical or virtual server;
- Associated host (if virtual server);
- CPU configuration (type, cores, speed);
- Memory allocated;
- Disk configuration (including data stores used for virtual servers);
- Physical location (datacenter, geography)

In addition, the following metadata can be useful in providing different views of your estate:

- Hypervisor (e.g. VMware, Hyper-V) if relevant;
- Operating system and version;
- Application(s) using this server;
- Application function (e.g. web server, DB server);
- Environment (e.g. Production, Dev-test, etc.).

This inventory data allows you to create hierarchical views of the allocation of your server and storage resources to various datacenters, applications, environments and so on. Crucially, you can then use these hierarchies to determine aggregate resource allocation and usage for higher-level entities; for example, how much CPU power is allocated to (and used) in your development environment versus your production environment.

Usage data

The inventory data provides information about how resources are allocated to various servers and environments. Usage data provides information about how those resources are actually consumed. When servers are initially deployed, it is common for resources to be allocated based on a best guess as to what is actually required. By proper assessment of the usage data, you can refine this allocation over time to ensure resources are right-sized according to what is actually required.

Physical servers

Usage data for physical servers will typically come from one or more monitoring tools. These are the tools that your operations team uses to generate

real-time alerts when there is a problem with servers. The specific measures you need to capture vary with operating system and your monitoring tools, but they specifically need to address to key resources that can limit the infrastructure capacity: primarily CPU, memory, disk I/O and storage space.

Virtual servers

The things you want to measure for virtual servers are pretty much the same resource usage metrics as for your physical server. One crucial point, however, is that whatever tool you use to extract those measurements has to be aware of the fact that these are virtual machines. Generally speaking, any tools that rely on agents on the Guest OS (the VM itself) will not be aware of this and will provide inaccurate information. The reason for this is that agents running on the Guest OS will not be able to make any measurements when the Guest OS itself has been suspended by the hypervisor (and will not even be aware of this suspension).

As a result, the most common source of usage metrics for virtual servers tends to be the hypervisor management system: vCenter for VMware, or SCOM for HyperV.

Understanding what's being measured

When using measures gathered by any tool it is important that you understand exactly what is being measured. For example, if you have a measure of CPU usage every 5 minutes, is this the average level of usage in that 5-minute period? Or a sample of what the usage was at some point in that 5-minute period? Or something different again?

As a more concrete example, VMware provides a number of different measures of memory usage by VMs. Two of these are active and consumed memory. The latter provides a measure of how much memory the hypervisor has allocated to the VM, while the former is an estimate of how much memory the VM is using based on activity monitored by the hypervisor. Using consumed memory for capacity planning will likely result in underestimating the capacity of a host, because a VM's consumed memory will only decrease if the hypervisor forces it to give up memory, which will only occur when the host is running out of memory. By contrast, because the active measure is a statistical estimate of the memory actually being used by the VM, it can be more usefully used in capacity planning to determine the real capacity of a host; however, care must be taken to ensure that allowances are made for the statistical nature of the estimate.

Establishing a baseline

Once you have gathered the right dataset the next step is to establish a baseline. The goal of a capacity planning baseline is twofold: firstly, it summarizes the historical data to provide a 'big picture' view of your estate; and secondly, it provides you with a view of headroom for growth, to enable you to quickly see which areas have capacity to absorb change, and which don't.

Taking the long view

For capacity planning you ideally want to be using the same data as used by your real-time monitoring tools, but over a longer time scale, so that you can build a statistical picture of resource consumption. So, while in a real-time environment you might want to track resource usage at 10–minute intervals (or even more frequently for critical servers) and always be able to see that last few hours in detail, for capacity planning you will want to roll these up over the course of 60–90 days to create a statistical profile.

It's also important not to rely solely on average values of metrics — ideally you need to calculate at least averages and variances, and in some cases you might need to create an empirical distribution of usage values (percentage of time spent in various utilization ranges). The goal is to be able to create a statistical profile, so that you know the likelihood that you reach or exceed given levels of usage (e.g. CPU usage on server abc123 is below 60% for 95% of the time).

Applying rules for headroom

A necessary starting point for capacity planning in a virtualized environment is understanding how much headroom exists in the environment at any point in time. This answers the question of "how much more work can the environment do over what it's currently doing?" The statistical approach described above provides a good mechanism for determining this. The baseline is, by definition, telling you how close

you are to your self-imposed usage thresholds. You can also use it to determine the resource requirements of an "average" VM (and again, you can make the decision about what you take as an average — across the estate as a whole, or for a set of critical applications). Combining these two then enables you to determine how many more average VMs you can accommodate in various parts of your estate.

An example might make this clearer. Let's assume you have a hundred VMs across four physical hosts. The hosts each have 8–core CPUs running at 2.5GHz, for a total of 20GHz CPU capacity (we're ignoring memory for the time being). On creating your baseline, you determine that hosts A & B the combined CPU demand from VMs is less than 12GHz for 95% of the time; and on hosts C & D, the combined demand is less than 15GHz for 95% of the time. So hosts A & B are running at 60% utilization; with C & D at 75%.

In order to determine your headroom, you need two more things: the usage threshold that you're willing to accept and the average VM demand. Let's assume that you are willing to let CPU usage reach 80%, and the average VM demand is 0.5GHz at the 95th percentile (i.e. the average VM demand is less that 0.5GHz for 95% of the time). A usage threshold of 80% equates to a demand of 16GHz on each of your hosts. So hosts A & B each have headroom for 8 more average VMs; hosts C & D each have headroom for only 2 more average VMs. Across your estate you have headroom for 20 more average VMs.

Predicting the future

In the previous section we discussed how a statistical baseline can be used to quickly give insight into headroom for growth, based on the number of average VMs that can be accommodated. Of course, the limitation of this approach is that many, perhaps most, of your VMs will be different from the average.

In order to address this issue, you need to create a statistical model that enables you to work through specific change scenarios and understand their impact. This will then give you the insight you need to pro-actively and accurately plan any adjustments you need to make to your future infrastructure capacity.

Modeling future scenarios

With the baseline as a starting point, you can now create a scenario model covering one or more related transformations to your environment that you wish to consider — such as changes to workloads (e.g. new workloads, scaling workloads, migrating workloads) and/or changes to resources (e.g. new servers, changes to server configs, rightsizing VMs).

Any scenario model built should also incorporate trends that you have discovered from your baseline data. For example, trends in usage of resources, and/or in number of VMs. These can be mapped on to the model as additional 'transformations'.

Each 'transformation' you add into the model will result in a re-calculation of demand for resources on a host, allowing you to determine predicted utilization. And because you're using a statistical model, you will also be able to determine confidence limits (e.g. 95% level). By using these confidence limits, along with the thresholds you have set, you will then be able to determine if any particular transformation, or combination of transformations, can be accommodated without a change in infrastructure capacity. For example, you may be able to determine with a 95% level of confidence that your threshold won't be exceeded. Whereas, if the model shows that a transformation can't be accommodated, you will be able to understand which resource is the limiting factor (is it CPU, memory, I/O) and how much more of that resource you need.

Three steps to successful capacity planning

Taking this statistical approach to capacity planning:

- collecting granular data and measures from your current infrastructure resources
- using that data to build an accurate baseline data model of your current environment
- then creating a scenario modeling capability that takes that baseline data and lets you try out any number of 'what if' change scenarios and see the impact

is the best way to ensure the future performance and efficiency of your critical IT infrastructure resources.

More information

To find out more about Sumerian Forward Thinking® predictive capacity planning and using Sumerian's capacity planner, just give us a call on 0131 226 9300, drop an email to sarah@sumerian.com or visit our website at www.sumerian.com

sumerian®
Forward Thinking