

Proven Practice: Implementing ITIL v3 Capacity Management in a VMware environment

Introduction

For those familiar with the ITIL framework you will of course know that by its very nature it tries to remain platform agnostic. But based on our experience working with ITIL and VMware, a number of key areas have been identified as being a focus point where the specifics of using VMware will have an effect.

To provide some guidance to people not familiar with the ITIL v3 Capacity Management terminology I will provide a brief overview of each section prior to discussing where the key areas of interest are in relation to VMware.

Intended Audience

This paper aims to provide some initial guidance for people that havent implemented or are in the middle of implementing ITIL v3 Capacity Management and have an infrastructure that comprises either fully or partly of VMware technology. We will discuss the key areas of ITIL where VMware will have an impact and provide guidance based on the authors own

experience of implementing ITIL processes and working in a VMware environment.

Targeted at Capacity Management and Service Management professionals, this will also be of interested to VCPs.

Outline

1. Business Capacity Management
2. Service Capacity Management
3. Component Capacity Management
4. Summary

Author



Metron is a privately owned limited company which was founded in 1986. Metron-Athene Inc is a wholly owned subsidiary of Metron technology Ltd. The company is Europe's foremost Capacity Planning and Systems Performance Management specialist. Metron's flagship product, Athene, provides fully integrated ITIL-compliant capacity management, automatic performance analysis and reporting for UNIX, Linux, Windows and Mainframe Servers .

[Find out more about Metron](#)

Robert Ford

Resources

This document on the web @ [Proven Practice: Implementing ITIL v3 Capacity Management in a VMware environment](#)

Disclaimer

You use this proven practice at your discretion. VMware and the author do not guarantee any results from the use of this proven practice. This proven practice is provided on an as-is basis and is for demonstration purposes only.

PROVEN PRACTICE: IMPLEMENTING ITIL V3 CAPACITY MANAGEMENT IN A VMWARE ENVIRONMENT

1. Business Capacity Management

The main focus for the business capacity sub-process is to ensure that the future business requirements are understood and that we have sufficient capacity to meet those requirements. This is our bridge between the organization and the Capacity Management process. This information will be provided from a number of sources and will depend on your organizational structure. In the purely ITIL sense they should come from the following sources:

- Service Strategy
- Service Portfolio

However, in the real-world this information could be provided by a range of additional sources that may include:

- Service Level Manager
- Service Manager
- IT Director

We would be expecting information on changes to existing services, any new services, expected growth and importantly patterns of business activity. It is this latter area that is of interest; the patterns of business activity are provided by the Demand Management process.

This process will work with Capacity Management to determine a long-term demand management strategy to meet the requirements of the business. All short-term demand management is undertaken by the operational Capacity Management process and could include the restriction of user access, or by balancing workloads. By using the dynamic nature of the VMware products we have been able to more effectively fit the technology around the needs of the business.

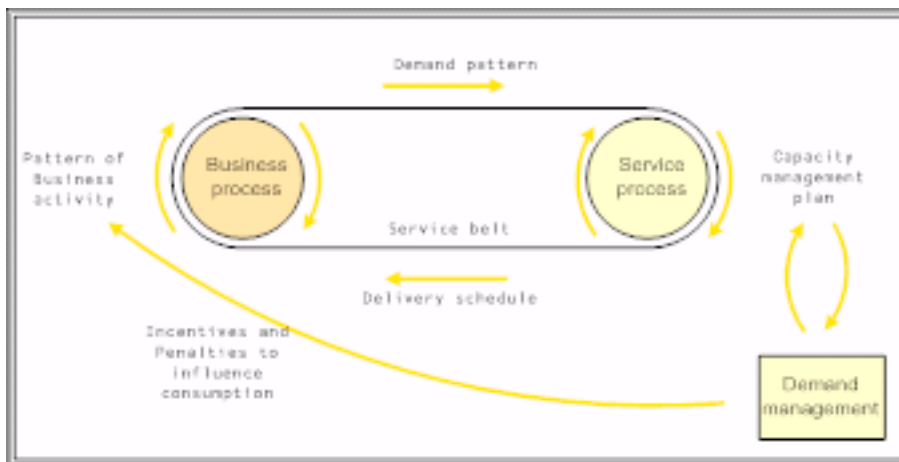


Figure 1 – ITILv3 Demand Management

By analyzing the business patterns (using whichever method you currently use, perhaps monitoring key transaction rates and combining those with business forecasts) you can determine any potential short term usage peaks and respond with additional capacity. This can be done via a number of options:

- The creation of Virtual Machines, if its appropriate for the system and you have a suitable license structure
- Throttle back the existing resource allocation on some of the key VMs to allow for the increase in demand
- The other option and one unique to VMware is to utilize Distributed Resource Scheduler (DRS) combined with VMotion. When configured correctly i.e. selecting the correct level of automation and migration level it is possible for any short-term demand management activities to be managed automatically.

Prior to implementing any of the above options it is essential from the capacity planning perspective to undertake a modeling initiative to determine what resource limits can be applied to both the Virtual Machines and the underlying hardware. From the ITIL perspective there also needs to be work conducted with the Configuration and Change Process to determine a standard process for tracking any dynamic reallocation of Virtual Machines.

A key goal for Capacity management is to provide value to the business; this can be done to great effect using VMware and consolidation. Using consolidation it is possible to reduce the number of physical servers and to utilize those reduced physical servers at a more optimal rate. We have completed a number of consultancy projects for our clients to assist in the consolidation of their infrastructure and certainly the key to this is in the planning stage. In order to undertake this exercise a modeling tool is used. Within this context there are two types of modeling tool.

For a consolidation exercise we would look to use a tool such as VMware Capacity Planner that will analyze the current utilization of your server estate and determine your optimum consolidation requirements

Once the consolidation exercise is complete the requirements of your modelling tool may change; obviously having our own modeling tool we are biased as to our choice, but whichever tool you choose it is important that it contains the following functionality:

- The ability not only to model the resource consumption and the relative response time as well
- If you are consolidating servers of differing OSs and platforms these need to be supported
- To eliminate any issues with clock drift, etc, it is recommended that the data be captured via VCenter.

Going through any virtualization/consolidation exercise should generate some additional process metrics that can be used to demonstrate the effectiveness of the process and the utilization of VMware. Within Business Capacity Management we are looking to provide metrics to demonstrate our value to the business and whether we are meeting their requirements. The sort of metrics we could look at producing are:

- % reduction in physical estate
- % reduction in power consumption (useful as documented evidence in obtaining any environmental standard e.g. ISO14001)
- % cost reduction in maintenance contracts

2. Service Capacity Management

This sub-process is concerned with resource consumption, activity patterns, peaks and troughs of the live operational services. Whilst it is the responsibility of the Service Level Manager (SLM) to police the response time targets set within the Service Level Agreements (SLA), it is the Capacity Management process that provides the data that is used.

As we know, the response time of a service/application is a key performance metric and it is used to determine the end user experience of using a service that conventional metrics like CPU utilization, memory consumption, etc wont necessarily show.

Data relating to the response time of a service can be difficult to obtain, but is critical to determining how a service is performing. This sort of data will also become more important when we look at consolidating servers into a VMware environment as the resource limits that are applied when viewing the hardware as a whole will have a greater impact.

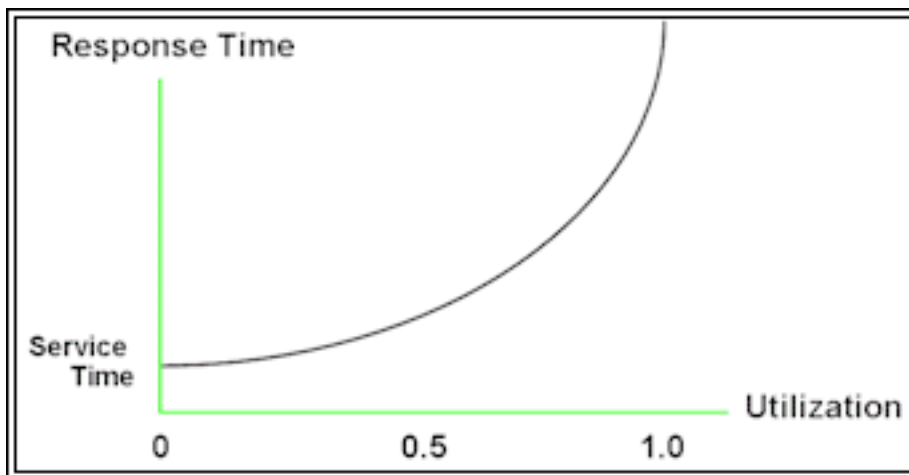


Figure 2 - Response Time vs Utilization

Figure 2 shows the relationship between utilization and the response time. As you can see the relationship isn't a linear one; as the utilization starts to get closer to 100% the response time starts to increase greatly. As discussed previously this issue will become more apparent as the number of Virtual Machines hosted increases and will need perhaps more management than if the servers were physical. The key to managing this is to ensure accurate collection of the response time. Again this may need to be changed when

implementing VMware, but it will largely depend on your current setup. If you are using a network sniffer style tool to capture response time data at each stage of a user transaction this may have to be amended to reflect a change in structure and will certainly need to be changed if you have used VMware to consolidate servers at the service level i.e. all physical components that culminate to produce service xyz are now on one physical VMware host. The same will of course be true if you have consolidated all physical hosts from a distributed environment to a centralized data center.

If you currently use a terminal emulation tool to graphically recreate a number of standard user steps to calculate the response time, then you may not need to alter the capture at all, but remember that the data captured prior to implementation will then provide a useful comparison from the SLA perspective. It will also provide a valuable business metric as to the benefits of moving to a VMware environment i. e. *We have consolidated Service A to VMware and it has reduced the user response time by 25%.*

The collection of data relating to a service can be further complicated when capturing data from servers/Virtual Machines running differing operating systems. Weve already discussed how response time metrics can give an indication of the overall performance of a service; but to determine the performance of the underlying components we need to fall back on the traditional metrics such as CPU utilization, memory consumption etc. This wouldnt be an issue in a non-virtual environment (assuming your toolset can collect from multiple platforms) but when you move to VMware you need to consider this data collection carefully.

Ultimately you either need a tool that is capable of collecting data from all of the platforms that you have virtualized or you need to be able to collect it from a central source that has visibility of all the Virtual Machines.

In our experience this is where the data provided by the VCenter can prove invaluable. From this central point you can access key metrics on CPU, memory consumption, across all platforms and Virtual Machines.

When this data is combined with response time data it will provide a valuable capacity planning feed for your operational services.

In addition to the VMware estate you may also need to capture data from additional non-virtualized platforms i.e. UNIX, mainframes etc. If a service is hosted across multiple platforms it is important that the toolset used to collect the data either supports those platforms or is capable of merging/aggregating that data into a service view. This view can prove invaluable when used by enterprise scale operations that have large numbers of services and need an efficient way of viewing their performance. Usually this data aggregation is provided as a benefit of using a central Capacity Management Information Store (CMIS) to store all business, service and component data. We will discuss both reporting and CMIS in greater detail in a later instalment.

We will discuss the types of data and their collection in the proceeding section.

3. Component Capacity Management (CCM)

This final sub-process is concerned with the performance and capacity of the underlying components that support the IT services. Its the CCM process that will be looking at the processors, memory, disks etc and trying to determine whether they are running at optimal capacity and whether based on the current and projected needs of the business, they will support the required service levels.

As weve discussed this sub-process is ultimately concerned with monitoring the components. For this final section we will discuss the following

- The sort of data that we need to collect
- How often do we need to collect this data
- How we will collect this data

Traditionally we are used to collecting metrics relating to CPU and memory utilization and maybe some additional relating to paging/swapping etc to determine how a server is performing.

So what changes when you have virtualized your servers with VMware?

In some respects the metrics remain the same, but its our interpretation of them that alters. We still look at CPU and memory when we talk about the physical server(s) that host our Virtual Machines but we have an added layer of virtual complexity when were talking about their resource utilization as we have to factor in any resource management that is in place. A key part in VMware and most virtualization technologies is the management of the physical hardware resource.

In a VMware environment we talk about the following terminology:

- Limit - A logical cap on CPU consumption (in MHz) and Memory consumption (in MB)
- Reservation - The reserved amount of CPU cycles and physical Memory (in MB) for a VM.
- Shares - CPU and memory resource is split into shares (MHz and MB); the more shares that are assigned to a VM the more likely it will be to win contention competitions
- Available Memory - Fairly straightforward, this is the available Memory assigned to a VM

Prior to establishing these initial limits and values an important part of Component Capacity is to monitor the utilization of the components over time and this applies in this scenario as well. If feasible, prior to virtualizing your servers, you should build up a picture of resource consumption which you can then reference when setting up the initial Virtual Machine settings. It is important that this time period includes as many periods of operational variability to give the best picture of how the resource consumption alters.

This information can also be fed into the demand modeling that maybe required when determining which physical servers will host which VM; although this may be negated if VMotion or a service based provisioning model is used.

Another key resource area to be aware of is the assignment of virtual CPUs. In a VMware environment a Virtual Machine will only start to process once it can access the number of vCPUs that it has been assigned. This is fine if the machine is quiescent or the VM has a sufficiently high shares rating, but if there is a contention for resource then the key metric to look for is the **%CPUReadyTime** this is effectively CPU queuing, if this is high then you need to be monitoring closely as to whether you have the resource distributed correctly.

Weve talked about CPU, but another key area is memory utilization. This can really be concentrated into two areas, the amount of memory being used and the amount of memory being shared. The amount of memory being used by the VM is fairly straightforward to understand, but the shared memory requires a little more explanation.

VMware employs **Transparent Page Sharing** to share pages of memory between VMs; so if one or more VMs have the same page of memory, VMware will store one copy and share it out to all the appropriate VMs.

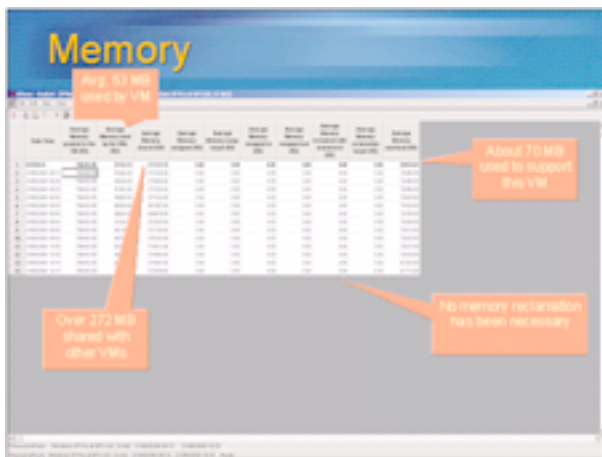


Figure 3 - VM Memory Metrics

Other key memory areas to be aware of are the **swapping and reclamation metrics** ; if these are non-zero values then it suggests that there has been a memory shortage

for that VM and more memory may need to be assigned. Figure 3 shows the short of memory utilization data that is available and the benefits of using things like Transparent Page Sharing, as you can see the memory shared is approx 272Mb, so without using this technology each VM would consume an additional 272Mb of memory.

So we've talked about the metrics and what we should be looking for, but how often do we need to collect these metrics and from where? Monitoring tends to be split into two kinds real-time and performance or planning. From our perspective being a company focused on the performance and Capacity management areas we tend to look for at the latter with real-time monitoring being more in the area of alerting or incident management.

If you are gathering data for the purposes of performance monitoring for an application sizing project or to analyze a particular performance issue then we would usually suggest that the sample time is between 2 and 5 minutes. If you are looking more at capacity planning then we would suggest around the 10 minute capture period for dynamic metrics such as CPU, memory, etc and maybe 60 minutes for largely static metrics such as disk utilization.

As for the source of data we have already discussed the merits of the central source provided by the VCenter versus a VM based monitoring solution, but either way you should consider the following:

- If your monitoring tool captures via the VM you/it needs to be aware of the resource limits that have been applied
- The VCenter provides an excellent overview of the resource utilization both at the host and VM level, but you may need to deploy a specific agent if you need a greater degree of granularity for instance a per process usage etc
- A slice of resource will be utilized by any monitoring agent, but if you have one per VM then ultimately the amount of resource consumed will be greater than collecting via the VCenter.

The majority of the logical resource thresholds are built up over time with careful monitoring and trending, but if it relates to the priority of the VM then you will need to link back to the Business and Service Capacity Management processes to work with the Business and Service Level Manager to determine this.

4 . Summary

Hopefully you have found this Proven Practice useful and it has at least provided some guidance to a few of the areas you should be aware of when managing the ITIL Capacity process in a VMware environment.

Ultimately whilst there are some technologies and terms that we need to be aware of as Capacity Managers, the techniques we have always used just need to be updated rather than replaced.

As discussed careful consideration should be given to both the Capacity Management toolset you currently use (if you have one) and whether that will provide the necessary information for managing the performance of your VMware environment.

From the ITIL perspective there is certainly scope to expand your KPIs to include metrics that will reflect the new environment and you may also need to consider how you interact with both the business and the other processes.

This document is the first in a series, with the next installments based on a similar theme but concentrating on the following topics:

- Capacity Reporting Structure and the Capacity Management Information System
- Capacity Modeling
- Implementation Guidelines

Should you have any questions regarding this document, any of the broader topics discussed or the services Metron provide please feel free to contact me on

Email: rob.ford@metron.co.uk

http: www.metron.co.uk

ITIL is a Registered Trade Mark, and a Registered Community Trade Mark of the Office of Government Commerce, and is Registered in the U.S. Patent and Trademark Office