Controlling Virtual Machine Sprawl

How to better utilize your virtual infrastructure

Dell

Ganesh Padmanabhan

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Table of Contents

Overview

Virtualization has not only improved hardware efficiency using server consolidation, it has also dramatically reduced the time required to bring a new server online from weeks or months down to hours or even minutes. This ability to bring machines online quickly provides companies with improved business agility; however, as with most improvements, there are also some new issues created that if not addressed will reduce a virtual infrastructure's overall efficiency over time.

Virtual machine sprawl is one of the biggest concerns facing many companies who have deployed desktop or server virtualization. The ability to quickly create virtual machines without the discipline and control of the physical world results in machines being provisioned unnecessarily without proper justifications and approvals, machines being over-provisioned (too much CPU, memory, or disk), and machines consuming resources well after they are no longer required.

To address this concern, a number of reporting products have been created to help track and identify virtual machine sprawl. The problem with report-only products is that they provide only a partial solution. They only help identify the problem and track it if it is getting better or worse. They do nothing about preventing the problem from occurring or correcting it once it has occurred. This white paper looks at how an automated workload provisioning platform can help decrease virtual machine sprawl by automating the process of reducing, reusing, and recycling compute resources in a shared physical infrastructure.

The Roots of Sprawl

If you look back at the IT industry, virtual machines are not the first compute resource to have sprawl problems. The general rule is that the easier it is to create something the more of them you get, and the harder they are to clean up; emails, files, databases, and storage are just some examples of information technologies that also have their sprawl issues. Reporting is a key component used to identify the problem as well as the scope; however, when it comes to fixing the problem, reporting is just part of the overall solution. With any technology that is susceptible to sprawl, policies and governance along with automation to help clean up and recycle are necessary components that help achieve efficient resource usage.

Email is an example of sprawl that everyone is familiar with; most people can easily find out how many emails they have, how much space they have consumed, and even which emails are consuming the most space. Those reports, however, do little to prevent email sprawl. The only way administrators control email resource consumption is to implement policies that control the amount of capacity each user can consume, and how large and what type of attachments are permissible. Additional policies that automate the archiving of old emails or the de-duplication of common information further optimize email resource utilization.

Compute virtualization is no different than other IT resources that have caused companies sprawl concerns in the past. There are two primary reasons why virtual compute infrastructures have more sprawl issues than their physical counterparts:

Speed

Requisitioning a physical machine is a lengthy process with many justifications and approvals required along the way. Even after the machine arrives, getting it racked and operational can also be time-consuming. The long process of bringing a physical asset online assures that machines are not provisioned needlessly. On the other hand, virtual machines can typically be created in a matter of hours from pre-existing physical resources.

Governance

For many companies, the process and tools used for provisioning virtual machines are completely different than the process used to provision and manage their physical infrastructure. The tools used to manage their physical infrastructure are not appropriate for their virtual environment. In the virtual environment, the hypervisor device managers, while providing basic management controls, do not enforce the same governance and compliance that are part of the physical management process.

Improve Infrastructure Utilization Efficiency

There are many corollaries between the environmental movement trying to eliminate waste and reduce resource consumption and IT organizations trying to drive higher utilization efficiencies in their virtual compute infrastructures. Reduce, reuse, and recycle are the three basic rules that drive the overall environmental movement. These same rules directly apply to the efficient management of a virtual compute infrastructure.

Reduce unauthorized and over-provisioned machines

Reduce means using fewer resources in the first place. Of the three approaches to eliminating waste, reducing the number of machines provisioned unnecessarily or over-provisioned has the largest potential return on investment. Gaining better control of the inflow of machines is far easier than trying to weed out unnecessary or over-provisioned machines later on.

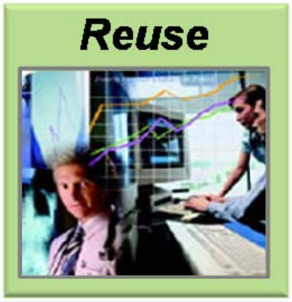
The problem with many virtualization-management tools is that they do not provide the governance and automation necessary to assure that requests for new virtual machines receive appropriate review and business justification. Once the provisioning process starts, there are few limits on what resources can be consumed and how much of those resources each virtual machine should receive. In an effort to limit the influx of new machines, many companies set up administrators who try to manually enforce governance. The problem with this model is that it does not scale as the environment grows without adversely impacting operational costs. The ideal solution is to delegate the provisioning process downstream to the resource consumers, but without appropriate controls and automation that strategy is impossible to implement without further loss of control.

What is needed is an automated workload provisioning platform that provides users with a self-service portal that allows them to initiate both provisioning requests for new virtual machines and ongoing management tasks throughout a VM's life, including its eventual decommissioning. This portal abstracts the complexities of using one or more virtualization device managers, but more importantly the software automates delivery and enforces governance relative to the resources a given user is allowed to consume. With the appropriate policies in place, administrators control not only how the machine will be built, but what resources and how much will be consumed. For additional control, policies can be established to automate the approval workflow, further assuring that machines will not be provisioned unnecessarily or over-provisioned.

The ROI Potential:

A typical company with 1,000 VMs that has five percent of its machines created without proper business justification and another five percent over-provisioned, could easily save over $100,000 - $150,000 in capital expenditures through better control of the front-end provisioning process. By delivering the “right sized” machine at the “right service level,” companies can eliminate waste, improve resource utilization, and lower costs.

Reuse resources automatically after they are no longer needed



Reuse means to use an item more than once and includes reusing a resource for the same function or a completely different function. One of the core tenants to virtualization is to improve efficiency through improved resource utilization. The degree to which resources can be reused has a high impact on the overall efficiency of a virtual infrastructure.

The problem with reuse in a virtual environment is that it is largely a manual process. There are a number of use cases where machines are only needed for short periods of time yet continue to exist well beyond when they are required. Examples of applications that require machines for short periods include data mining, calc farms, development, testing, month-end processing, etc. The same is true for virtual machine archival or snapshots.

What is needed is an automated workload provisioning platform that automatically reclaims and reuses resources when they are no longer needed. Policies control how long a machine will exist and process automation orchestrates the decommissioning and reuse upon lease expiration of that virtual machine. Machines that need to be archived may not be consuming CPU and memory, but storage capacity can be very costly. Policies that control how long a VM will be archived after it is no longer needed and automatically cleaned up after the archival period are essential to driving improved efficiencies. Snapshots are another area that, if not managed properly, has a tendency to grow and consume storage resources unnecessarily. Policies that control if a machine is allowed to have a snapshot, how many snapshots can be created, and how long they are allowed to exist before they are automatically coalesced are essential to efficient resource management.

The ROI Potential:

A typical company with 1,000 virtual machines that has five percent of its machines used for temporary applications can expect to save about $60,000 - $70,000 by automating the reuse of leased machines. Environments like development and testing that have a larger need for temporary machines, can expect even greater savings. The same sized company that archives 10 percent of its machines on a yearly basis can expect to save $30,000 - $50,000 annually by reclaiming and reusing storage resources after the archive period has ended.

Recycle inactive and abandoned machines

Recycling is the process of collecting, processing, remanufacturing, and reusing materials. As with most recycling programs, identification and collection of unused resources tends to be the most manual and time-consuming part of the process. The same is true in a virtual infrastructure,

where reclaiming inactive, abandoned, and stranded capacity can be such a labor-intensive and time-consuming process that it is performed very infrequently if at all. The result is that inactive machines tend to stay around longer and consume more resources unnecessarily than if an efficient reclamation process were in place.

The first problem with reclaiming inactive resources is the identification of inactive machines. For most companies, this is a manual process consisting of data collection scripts and spreadsheets. Even companies that have invested in reporting software to help identify these inactive machines often encounter the second problem that is filtering out the truly inactive machines from those that appear to be inactive. Some machines may meet several of the criteria that would cause them to appear in an inactive exception report, yet are known to be machines required for the business. If a large percentage of these essential machines keep showing up on an exception report, the reports become ineffective, because administrators can't see the problem machines through the clutter. This identification issue brings up the third recycling problem, which is how to validate that a virtual machine is truly inactive by confirming with the owner and then automating the reclamation process. Report-only solutions are of little help here as they were not around when the machine was created. Therefore, they do not know who the owner is and they have no process automation to help verify that the machine is no longer needed in order to initiate the process to recycle the inactive machine's resources.

What is needed is an automated workload provisioning platform that not only provides exception reports that help identify stranded, inactive, and abandoned machines, but also automates the workflow associated with reclaiming those resources. Report-only solutions solve less than half of this problem, leaving the remaining manual work to administrators, thereby decreasing the overall operational efficiency of the process.

The first step is to identify stranded capacity within the host or host clusters used to run the virtual machines. Stranded capacity is caused when one of the three critical resources (CPU, memory, and storage) has been completely consumed, yet the remaining resources are still in abundance. This capacity is effectively stranded, because it can't be used. Reports that help identify hosts with this condition allow administrators to either add capacity to the bottlenecked resource or reallocate stranded capacity to other hosts, improving overall utilization of the virtual compute infrastructure.

The second step is to identify inactive and abandoned VMs. Since there is no single criteria that can unequivocally identify that a virtual machine is no longer needed, the best that can be done is to identify machines that are potentially inactive. In order to effectively manage the reclamation process, an automated workload provisioning platform needs to be able to eliminate known exceptions from these reports, validate with the owner that the machine is still required, and decommission and reclaim the resources of the machines that are no longer needed. The problem with many of the solutions available today is that they do little to help with the validation and reclamation process.

The ROI Potential:

Even if a company automates the recycling and reuse of resources used for its temporary machines, identification and recycling of resources used for permanent machines can still be problematic. This is especially true for virtual desktops, which tend to have more churn than virtual servers. A typical company with 1,000 virtual machines and 10 percent of its resource being consumed by inactive and abandoned VMs can expect to save $80,000 - $100,000 on capital expenditures annually. Even if the automated process is run more frequently than the current manual process (e.g. quarterly versus annually), companies can expect to save another $20,000 in operational expenses associated with identification and reclamation of inactive resources.

Management: The Key to Efficient Resource Utilization

Infrastructure virtualization has an immediate and quantifiable return on investment. It is easy to compare the cost of a physical asset to its virtual counterpart. Consolidation ratios of 10:1 up to 20:1 are not uncommon for servers and are much higher for desktops. However, if a virtual infrastructure is not used efficiently, a large percentage of the savings can be wasted, resulting in actual savings being far less than what was theoretically envisioned.

Hypervisor device managers provide solid management capabilities, but do little to enforce the governance required to reduce the inflow of unnecessary and over-provisioned machines. Report-only solutions help identify inactive and abandoned resources, yet do little to automate the reclamation process. An automated workload provisioning platform that automates the service delivery and ongoing management of a virtual infrastructure can justify its additional cost through improved resource utilization alone. In addition, it provides other savings by improving operational efficiencies through process automation and helping accelerate virtualization deployments by providing the tools necessary to manage a growing virtual infrastructure more efficiently.