

Realtime
publishers

New Best Practices in Virtual and Cloud Management: Performance, Capacity, Compliance, and Workload Automation

Greg Shields

sponsored by

vmware[®]

Chapter 2: New Best Practices in Capacity Management.....	15
Private Clouds, Resource Pools, and the Supply and Demand of Resources	15
Abstracting and Simplifying Capacity Management.....	19
Reintroducing the Jeejaw	19
Simplifying by Abstracting	20
Wither the Trendline?	21
Capacity Management, Converged Infrastructure, and Hardware “Designed with Private Cloud in Mind”	22
White Boxing, Generation I.....	23
White Boxing, Generation II.....	24
Virtualizing I/O.....	25
Perhaps Not Chargeback But “Showback”	25
Show Me the Money.....	27
Everything (Virtual) Is Economics	28

Copyright Statement

© 2012 Realtime Publishers. All rights reserved. This site contains materials that have been created, developed, or commissioned by, and published with the permission of, Realtime Publishers (the “Materials”) and this site and any such Materials are protected by international copyright and trademark laws.

THE MATERIALS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. The Materials are subject to change without notice and do not represent a commitment on the part of Realtime Publishers its web site sponsors. In no event shall Realtime Publishers or its web site sponsors be held liable for technical or editorial errors or omissions contained in the Materials, including without limitation, for any direct, indirect, incidental, special, exemplary or consequential damages whatsoever resulting from the use of any information contained in the Materials.

The Materials (including but not limited to the text, images, audio, and/or video) may not be copied, reproduced, republished, uploaded, posted, transmitted, or distributed in any way, in whole or in part, except that one copy may be downloaded for your personal, non-commercial use on a single computer. In connection with such use, you may not modify or obscure any copyright or other proprietary notice.

The Materials may contain trademarks, services marks and logos that are the property of third parties. You are not permitted to use these trademarks, services marks or logos without prior written consent of such third parties.

Realtime Publishers and the Realtime Publishers logo are registered in the US Patent & Trademark Office. All other product or service names are the property of their respective owners.

If you have any questions about these terms, or if you would like information about licensing materials from Realtime Publishers, please contact us via e-mail at info@realtimepublishers.com.

Chapter 2: New Best Practices in Capacity Management

There exists an economy of resources in a virtual environment. Hardware contributes to resource supply, while resources are demanded by needy virtual machines.

It was years ago that I obtained my bachelor's degree, an achievement that required no small amount of study in the field of economics. I counted myself among the few who loved the topic, and for a very specific reason: I had the privilege of learning from one Professor Fred M. Gottheil at the University of Illinois Department of Economics. Dr. Gottheil was a consummate presenter, consistently delivering memorable lectures to his massive auditorium of students.

It wasn't just his presentation prowess that's kept Doc Gottheil in the back of my mind; it was his grasp of the "real world" of economics. He professed that although economics at face value concerns itself with dollars, dig shallowly beneath the surface and *you'll find economics in everything*. The doc believed, as I do now, that there's supply and demand everywhere: How the number of lanes on a highway impacts traffic flow. How the price of a relationship is governed by emotion. Heck, how even the Golden Rule¹ can arguably be described within the lens of economic theory. Little did I know that these rules would later apply to IT as well, and most specifically to the approaches we use in managing our virtual and cloud environments.

Private Clouds, Resource Pools, and the Supply and Demand of Resources

I begin with this story because IT data centers perhaps unknowingly follow the rules of supply and demand, or at least the good ones do. Today's ever-increasing embrace of virtualization and cloud computing creates the situation where direct measurement of supply and demand has become the new best practice.

At the center of it all is the resource pool.

Explaining this assertion requires first stepping back from the capacity management activity and focusing instead on what we're asked to manage. In recent years, we've been told that the most efficient approach to delivering virtual environment resources involves thinking like a Private Cloud.

¹ The Golden Rule: Do unto others as you would have them do unto you.

To think like a Private Cloud, one can assume you first need to have one. Yet the question seemingly on everyone's minds is, "What makes a Private Cloud?" Study the IT press and the vendors' marketing glossies and you'll learn that a Private Cloud today enables

- Availability for individual IT services
- Flexibility in managing services as well as in deploying new services
- Scalability when physical resources run out
- Hardware resource optimization, to ensure that you're getting the most out of your investment
- Resiliency to protect against large-scale incidents
- Globalization capacity, enabling the IT infrastructure to be distributed wherever it is needed

At first blush, this list makes sense in terms of IT needs². We need high availability for our IT services. We want flexibility in managing and deploying new services, as well as scalability to ensure existing ones can expand when necessary. As the first chapter argues, our costs also demand optimization on that investment to ensure we're squeezing out every dollar of benefit. Our businesses lay resiliency and geo-location requirements on us to protect and distribute assets wherever they're needed.

However, although these marketing-friendly terms define what a Private Cloud *strives to accomplish*, they don't say much about what it *really is*. In fact, our industry's disagreement on a commonly-accepted Private Cloud definition might just be at the center of our confusion on how to construct one. Maybe we need to simplify. Here's a definition I've used before³:

Although virtual machines are the mechanism through which IT services are provided, the Private Cloud infrastructure is the platform that enables those virtual machines to be created and managed at the speed of business.

This definition argues that the hypervisor constitutes a layer of data center resource abstraction. That hypervisor abstracts physical resources to virtual machines, enabling virtualization—and, thus, virtual machines—to perform their tasks. Supporting that hypervisor, then, is an entire infrastructure of hardware and processes, the collection of which enables the hypervisor to do its job. In this description the Private Cloud is that infrastructure.

² At least as I see it. I originally wrote this list for Chapter 2 of *Private Clouds: Selecting the Right Hardware for a Scalable Virtual Infrastructure* (Realtime Publishers, <http://nexus.realtimepublishers.com/pcsrh.phpH>).

³ Ibid.

Yet although it sounds good, this definition doesn't well describe *in IT terms* what makes that Private Cloud. It defines the infrastructure but not the resources that make up that infrastructure. As a result, I've found myself using a second definition:

A Private Cloud at its core is little more than a virtualization technology, some really good management tools, and those tools' integration with business processes.

This second definition suggests that a Private Cloud is perhaps something far simpler than what we assume, and acknowledges that a Private Cloud requires a hypervisor as well as a set of tools to manage that hypervisor; those tools' functions align with business needs. This second definition argues for simplicity. Even so, I've found it doesn't resonate well with IT practitioners.

I talk about these two definitions first because they introduce my third—and so far best—definition. To me, the following definition best explains the Private Cloud concept in IT terms. It does so by focusing on the resources we're responsible for managing:

A Private Cloud is a host cluster with high availability and disaster recovery services turned on, plus a little bit.

And that's it. A Private Cloud, at its most fundamental, is a resource pool (see Figure 2.1). It is a logical collection of all the little data center resources that IT services—and virtual machines—require to accomplish their mission. Facilitated by the scheduling function of a distributed hypervisor, resources such as processor cores and threads get aggregated to fill the pool. The same holds true with memory, disk, and network resources as well.

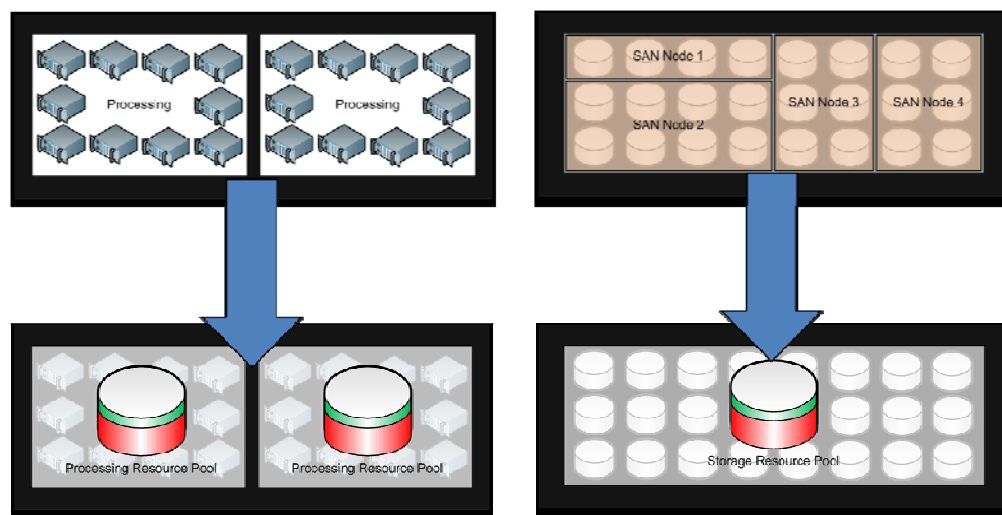


Figure 2.1: A Private Cloud is a resource pool.

It's the screen in Figure 2.2 that helped me realize this third definition. This interface is borrowed from one of the major virtual platform players, but the vendors have solutions that are similar.

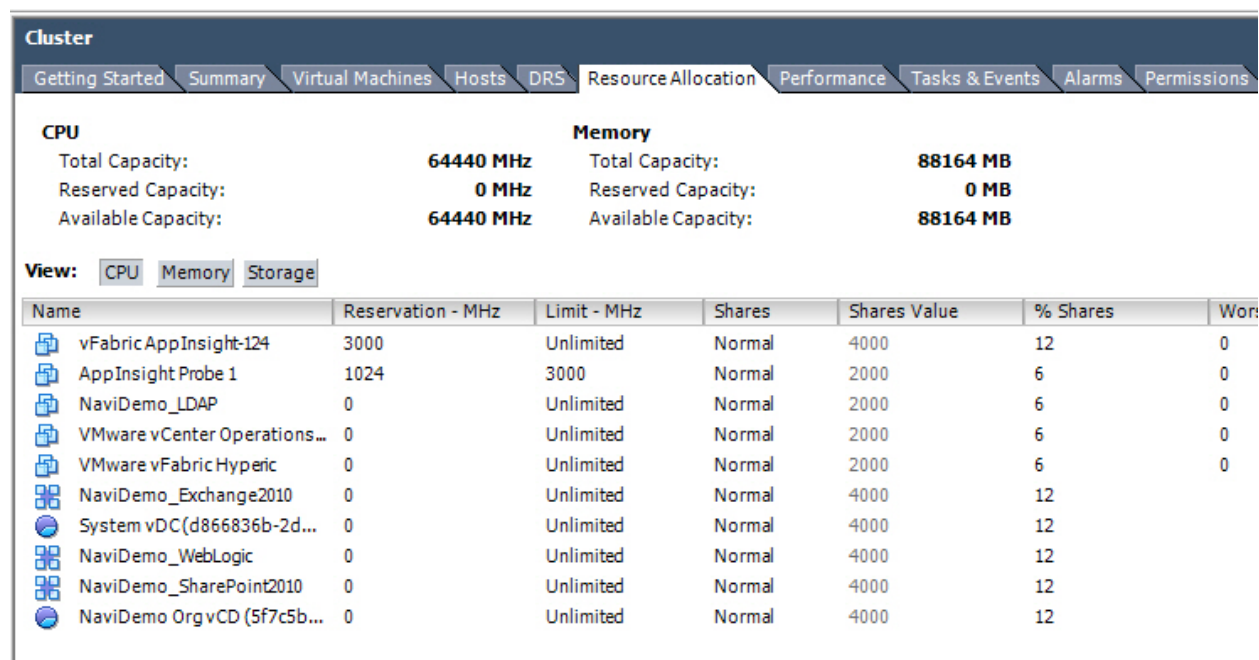


Figure 2.2: Resource allocation in a host cluster.

In this screen, you see the resource allocation for a host cluster as well as the total CPU capacity for the cluster in whole. That capacity is represented as:

$$\left(\text{Number of Logical Processors/Host} \right) \times \left(\text{Number of Hosts} \right) \times \left(\frac{\text{MHz}}{\text{Local Processor}} \right)$$

It further shows the memory capacity of the cluster, represented as:

$$\left(\frac{\text{MB of RAM}}{\text{Host}} \right) \times \left(\text{Number of Hosts} \right)$$

These formulas might seem excessively simple when considered within the greater framework of a dynamic virtual environment. But they're valid, and they constitute the foundation of capacity management: *How many resources do you have?*

Recognize also that those values are but half of the equation. Paired up with them are the values for reserved capacity and overhead utilization as well the quantity of resources that are currently being consumed. These, at a very high level, constitute capacity management's other half: *How many resources are you using?*

Or, as Dr. Gottheil might say, these two values identify your Private Cloud's *supply of resources in relation to its virtual machines' demand*. As you should quickly see, there truly is economics in everything.

Abstracting and Simplifying Capacity Management

Chapter 1 discusses how virtual and cloud environments are complex interconnections of hardware, software, and services. Their functionality requires a careful orchestration of components, even as each is managed independently. In the prototypical virtual and/or cloud environment, networks tend to be managed by one set of tools and administrators, storage by a second set, and servers and the virtual environment by a third. Doing so facilitates the separation of duties as well as the separation of administrative domains.

Chapter 1 further argues that measuring performance in such an environment can only happen by watching metrics at every layer, all at once. I reintroduce a figure from Chapter 1 as Figure 2.3 to reinforce those areas where monitors might get targeted.

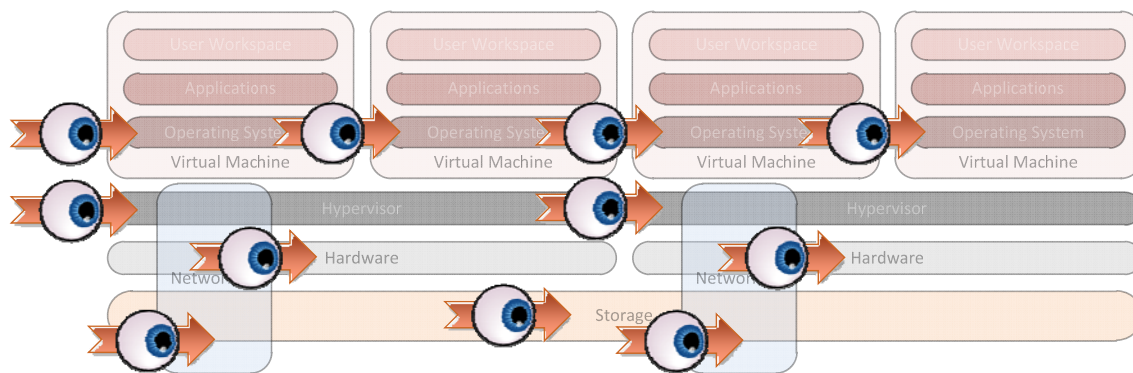


Figure 2.3: Multiple components require multiple monitors.

Important to recognize at this point is that performance and capacity management are very different activities. The former concerns itself with the experience in using the system, “Is it fast enough today? If not, why?” The latter focuses on what amounts to a single equation, “How many resources do I have, and how many will my workloads need?”

Where the two activities get commonly confused is in the analytics used to answer their questions. You can use the same kinds of monitors to measure capacity as can you for measuring performance. *The difference is in the questions you’re attempting to answer.*

Reintroducing the Jeejaw

To illustrate this activity, let’s bring back the nonsense metric first introduced in the first chapter: *the Jeejaw*. Just like before, the Jeejaw measures some aspect of the various components that constitute our Private Cloud environment⁴. Different here are the questions we’ll use the Jeejaw data to answer.

⁴ I’m assuming that you’ve bought into my earlier definition that a fully-realized virtual environment (with all the failover and load-balancing features enabled, plus a little bit) is also essentially a Private Cloud.

As any virtual administrator knows, virtualization primarily concerns itself with The Core Four: processing, memory, storage, and networking. It's the hypervisor's job to abstract these core four resources and make them available for co-located virtual machines. Each virtual machine demands a specific quantity of each resource at every point in time: A heavily-taxed database might need more, while a lightly-used IT file server might need much less, and so on.

Also important is the recognition that virtual environment resources are shared resources. These resources are highly dynamic, which makes them difficult to measure unaided.

Capacity management (see Figure 2.4) at its most elemental is concerned with ensuring enough resources are available (the supply) to meet the current and future needs of these workloads (the demand). This activity is made challenging by the "messiness" that's intrinsic to virtualization: resources are used dynamically, virtual machines can be relocated anywhere, stuff is constantly being powered on and off. These combine to make the capacity management activity just as difficult as performance management when one has no tools to assist.

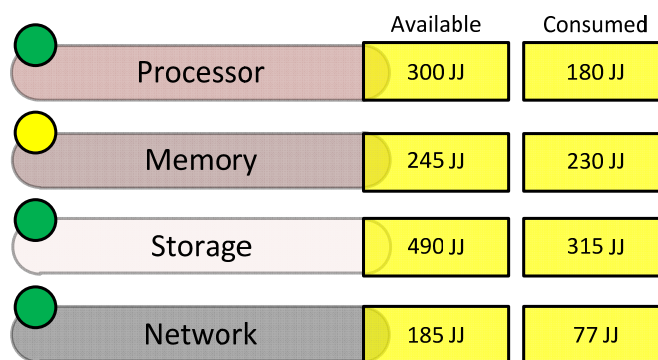


Figure 2.4: Abstracting Core Four resources into integer values.

Simplifying by Abstracting

An important new best practice focuses on tools that simplify capacity management *by abstracting the data*. Figure 2.4 shows a representation of how this might look. In it, a virtual environment's innumerable metrics have been replaced by representative values for each of the Core Four resources. For each value, there is an assertion of the capacity of that resource in contrast with how much is currently being used.

Armed with this information, a virtual administrator can, with a passing glance, get an immediate feel for where resources are getting low. In the example in Figure 2.4, processor, memory, and network resources are sufficient to meet virtual machine demands. Memory resources, however, appear to be running out.

Admittedly, these numbers should be of only limited value in their absolute form. One can argue that a well-managed Private Cloud will never see a green capacity light go yellow or red. Its proper planning involves acting before resources get low by ensuring more will arrive before they run out.

The problem here is again virtualization’s “messiness”—its incredible complexity that goes beyond the analytical limits of the unaided human brain. As a consequence, it is exactly this kind of planning that is incredibly difficult to accomplish without tools to assist. Figure 2.5 shows one such visualization that shows a virtual environment’s memory consumption over time.

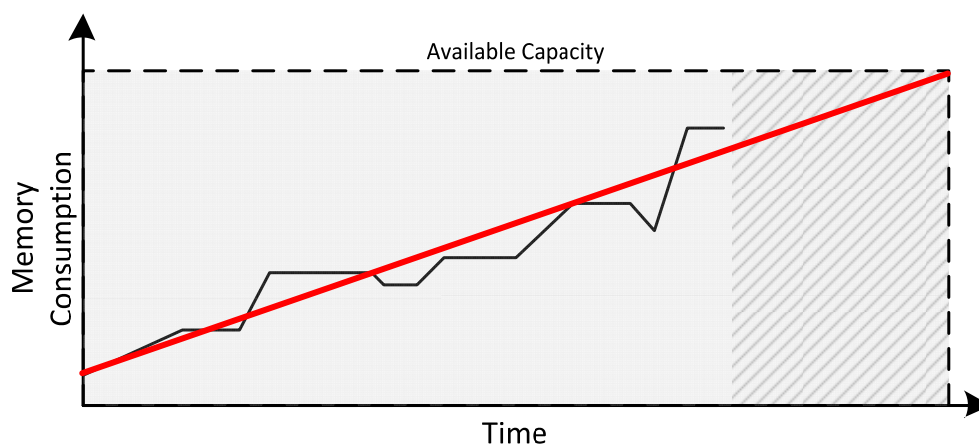


Figure 2.5: Trendlining memory consumption over time.

Graphs like these are necessary to show consumption trends over time. More important than the actual values is the graph’s red trendline. That trendline points to some future date when memory consumption can be expected to exceed available capacity. *Your job in the capacity management activity is all about ensuring resources are always available before that day comes.*

Wither the Trendline?

That stated, one must be careful with simplicity alone. Trendlines can be insidious, and any IT professional with a copy of Microsoft Excel and a passing grade in statistics can generate them. Like statistics, poorly-constructed trendlines can lie. A couple of clicks inside Excel, and Figure 2.5’s first-order trendline can be easily converted into a second-, third-, or greater-order slope, any of which can greatly shift the graph’s end date forwards or backwards.

The data feeding into Figure 2.5’s memory consumption prediction can also lie. Getting good data in a virtual environment can be very challenging. Tools or manual methods that don’t include abnormal workloads, consumption peaks and valleys, availability reserve, seasonal and cyclical trends, and virtualization overhead in their equations will generate less-than-trustworthy results.

A final challenge is in divining actionable intelligence out of the data you've collected. Consider as an example how the aforementioned Excel spreadsheet might fail when its metrics can't be forecasted to meet the range of possible future situations. Figure 2.6 shows a representation of this, where three possible future states are predicted: The first shows anticipated memory consumption when future business is not expected to change, the second predicts consumption should the market drop, while the third predicts consumption for the case where the business gets acquired.

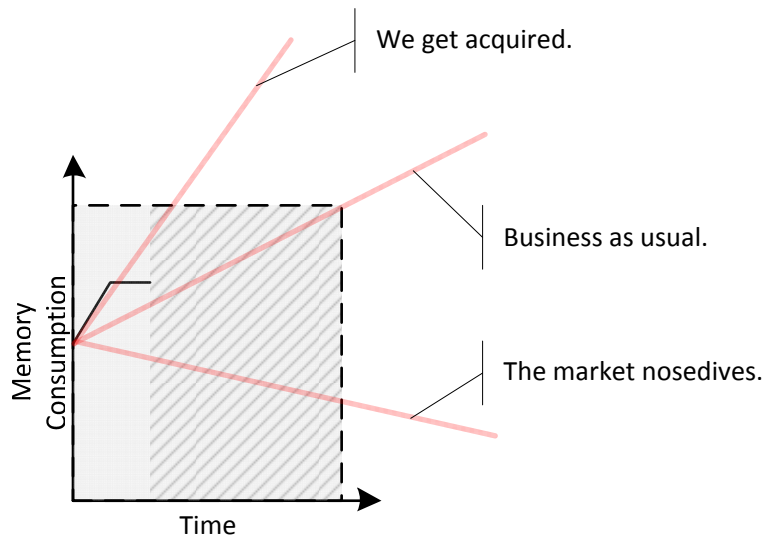


Figure 2.6: Trendlining across situations.

These kinds of *What-If Modeling* tools are useful for predicting resource demand and supply and are invaluable in the highly-dynamic virtual environments that follow Private Cloud thinking. The result of their modeling directly impacts the relationship between IT and a business' supply chain, and goes far in ensuring that IT resources can always meet business demands.

Capacity Management, Converged Infrastructure, and Hardware “Designed with Private Cloud in Mind”

Admittedly, the activities in capacity management are all academic without real-world actions that are the result of their effort. At the end of the day, capacity management is very much a function of *purchasing*. One performs its activities to ensure that just enough resources are always available—never too much, and definitely never too little. Buying hardware is often the real-world action that results.

Capacity Management: Not Necessarily Always More

Actually, that last statement isn't entirely true. It might be true in today's completely on-premise environments. In those, there's generally always a need to "keep up with the demands of business." More business ventures usually mean more resources that need to be brought to bear. Such internal services are notorious for rarely being decommissioned, even with IT's usual due diligence in seeing legacy equipment out the door.

Everything changes, however, when businesses begin to extend their Private Cloud environments into the public cloud. There, services are priced not as hardware but on a consumption model. That model naturally incentivizes decommissioning services the moment they're no longer useful, or alternatively, bringing those services in-house when they've been deemed a long-term asset worth managing.

Although much of the talk in this chapter discusses how capacity management tends to directly impact new purchases, it is important to recognize that that needn't necessarily be the case in certain Public and Hybrid Cloud situations.

In fact, that recognition might just be one good reason to consider extending into Public Cloud services. Doing so leverages your capacity management tools and experience for cost containment instead of merely always buying more equipment.

There is, interestingly enough, a new best practice associated with the kinds of hardware one associates with Private Clouds. The current buzzword for this class of hardware is *Converged Infrastructure*, although it might better be described as "hardware that's designed with Private Cloud in mind."

Describing this new somewhat-specialized hardware to experienced IT professionals I've found to be something of a challenge. The prototypical IT professional has gotten used to focusing on hardware and its management as a primary function of their job. Even in today's virtual world, we still think of servers as "servers"—even when they're virtual hosts. A server, and the hardware chassis that encapsulates that concept, is in many ways the unit of management for the average IT practitioner.

White Boxing, Generation I

The hardware that defines Converged Infrastructure strives to evolve that preconceived notion of "server." It does so, at least as I like to describe it, *by eliminating our industry's second generation of white boxing.*

Bear with me now, because the story makes a lot of sense: Many years ago in the time before vendor-engineered server hardware, one of IT's tasks involved constructing "servers." In those days, server hardware followed the same approach as does some types of consumer hardware today: You buy a motherboard from one vendor, a case from another, and memory and hard drives perhaps from a third. The process of "building a server" required actually constructing that equipment out of whole cloth, assembling all its individual pieces to create *a white box*.

This white boxing practice worked well enough in the days before vendor engineering created today's notion of a server. The early practice created downstream problems, notably, in that every white box was a little bit different than the one before. Eventually, the practice was abandoned as we realized the configuration control and stability benefits in buying vendor-engineered servers over assembling our own.

White Boxing, Generation II

At some point, virtualization became the new best practice, and virtual servers began to outnumber physical servers in data centers everywhere. What many IT professionals don't recognize, however, is that the embrace of virtualization inadvertently created a second generation of white boxing. But this time, we're white boxing our entire data centers.

As a consequence, succeeding generations of virtual environment hardware—whether purchased ad hoc or as the result of capacity management activities—began to accumulate (see Figure 2.7). You see evidence of these activities in data centers everywhere: Virtual hosts that were purchased in groups inadvertently create islands of compatibility for our selected hypervisor. This situation creates big pains: Virtual machines can't migrate across different hosts, and so a single, unified Private Cloud is forced to become a collection of smaller ones. Just like before, our actions have inadvertently created more work for ourselves.

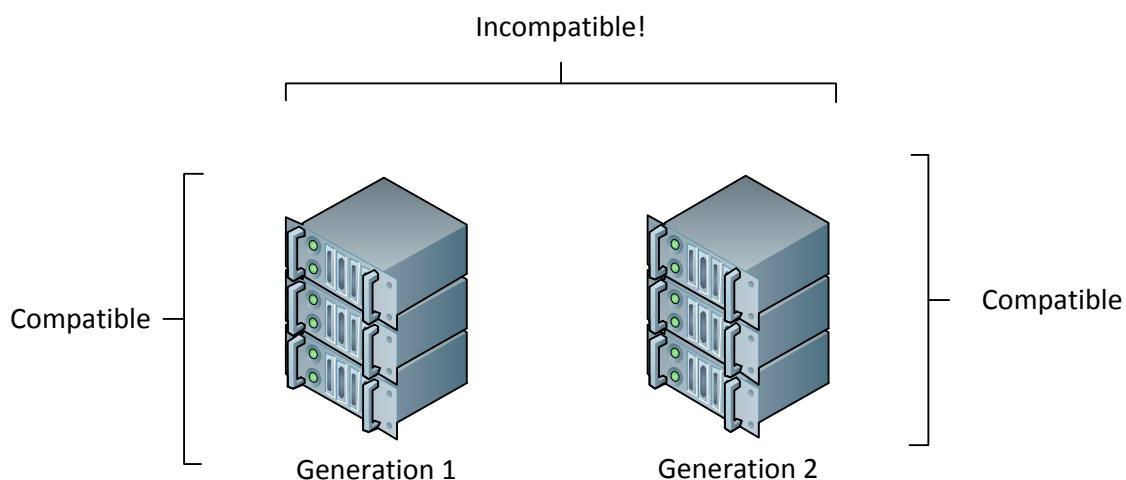


Figure 2.7: Compatibility within purchasing generations, but incompatible between generations.

Virtualizing I/O

Overcoming incompatibility is only one of Converged Infrastructure's goals. Improving the activities in capacity management has also become a new best practice. Converged Infrastructure's hardware aims to accomplish this goal by increasing the level of commonality among physical hardware, while adding Private Cloud-aware enhancements to the hardware itself.

Many of those enhancements center on reducing the complexity of interconnections between components. Leaning on ever-faster technologies in networking, these interconnections evolve from being predominantly physical to *almost completely logical*. As logical connections, their behavioral patterns are far more easily monitored and profiled by a performance and/or capacity management solution.

That's a good thing because it is the interconnections between components that touch everything in a Private Cloud environment. As Figure 2.8 shows, the interconnections exist at the hardware layer, the hypervisor, and even into virtual machines' interaction with storage. Keeping a very robust eye on that network's behaviors goes far towards delivering the kind of data that a capacity management solution requires.

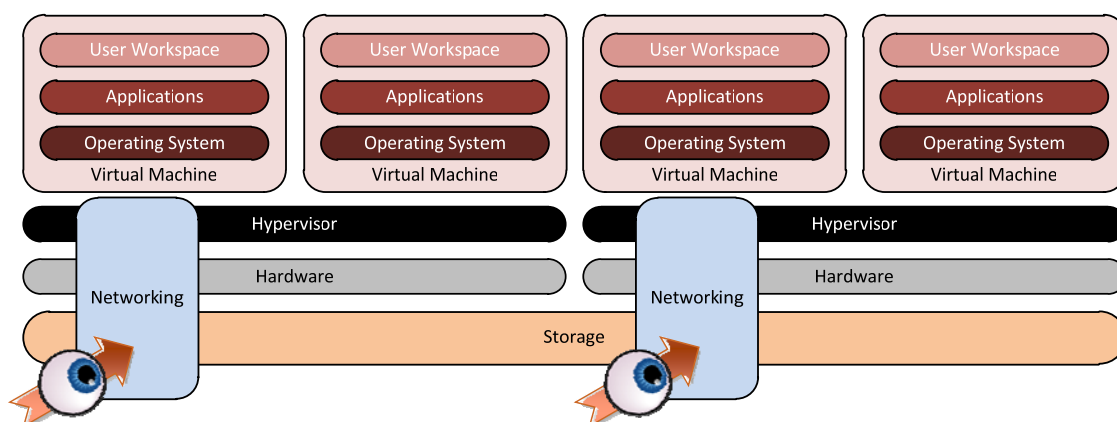


Figure 2.8: Networking touches everything.

Perhaps Not Chargeback But “Showback”

The final topic worth discussing in this capacity management exploration deals with the business' desire for IT alignment. “Aligning IT with the business” has over the years incurred an almost-humorous series of missteps:

- There were the “make IT a profit center” campaigns a number of years ago. Many weren't entirely successful.
- Others attempted an “IT as a business-within-a-business” approach, whereby services were charged back to those requesting them.
- Even others embraced the outsourcing model, which traded reductions in unexpected costs for a rigidly inflexible service delivery model.

One can argue that all of these missteps attempted to accomplish one thing: Bring business relevance to IT costs.

The tools that facilitate capacity management add another new spin on IT-business alignment. One such approach eschews the challenging-to-implement chargeback approach for a simpler-but-no-less-effective model called *showback*.

Recall my earlier (“third”) definition of Private Cloud: “A Private Cloud is a host cluster with high availability and disaster recovery services turned on, plus a little bit.” I’ve purposely held off discussing the “little bit” until this point in the story.

Many IT pundits believe that one facet of that little bit centers on *self-service*. In a self-service environment, entitled users are given the ability to generate their own IT services at will. Generally, as long as they meet certain requirements—resource use, business need, and so on—such users are free to create, manage, and decommission whatever services they need. It becomes IT’s job to manage the templates, ensure everyone plays nicely with each other, and maintain appropriate resource reserves.

As you can imagine, implementing self-service without capacity management is a recipe for chaos. Lacking capacity management, self-serving users tend to consume resources until those resources are exhausted. That’s not proactive management.

Conversely, a resource pool that is capacity managed has the ability to set hard limits on how many resources each consumer gets to work with. Figure 2.9 shows how the total resource pool in a Private Cloud can be broken down into sub-pools by project, with business rules defining the percentage each receives.

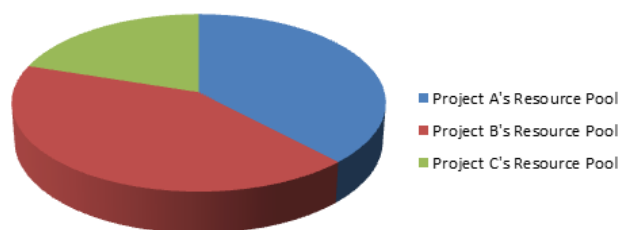


Figure 2.9: Dividing the resource pool by consumer.

Show Me the Money

Creating resource pools and subdivided resource pools is a feature commonly found in hypervisor platforms. Such pools work for some organizations but not all and not in every situation. What they're missing is that important linkage between IT and the business: *the dollars*.

The showback model facilitates assigning a dollar value for IT services, then showing that value back to the consumer. Different than chargeback, where costs are actually charged back to the person requesting the service, the showback model brings real-world dollars-and-cents valuations to IT services (see Figure 2.10): *Need more disk space? That'll cost you ten bucks. Another processor? Fifteen. New server? That'll be a hundred and fifty.*

IT Services Dashboard

Please Select a Service...

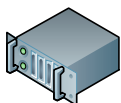
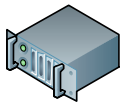
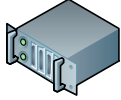
		<u>Processors</u>	<u>Memory</u>	<u>Disk</u>	<u>Cost</u>
<input type="checkbox"/>		1	512M	20G	\$3.75
<input type="checkbox"/>		2	1024M	40G	\$5.50
<input type="checkbox"/>		4	2048M	80G	\$9.33

Figure 2.10: Assigning costs to services.

It is in this space where a significant amount of IT leadership is being seen in today's businesses. An IT organization that knows the costs of its services greatly aids business decision making. Budgeting for new projects becomes a science rather than an art. Forecasting becomes as much a budgetary activity as a technology activity. Although no one directly gets penalized when they over-consume, as is the case in the chargeback model, service consumers are given the business-relevant data they need to be more successful with their decisions.

Everything (Virtual) Is Economics

Dr. Gottheil was indeed right. Economics are to be found in everything. You can find the principles of economics in the ways IT delivers services to its users. Many of us for a long time haven't applied those principles consciously, perhaps limited by the resource challenges in our early physical environments. Make those environments virtual and begin applying Private Cloud thinking, and suddenly that invisible hand becomes far more recognizable.

Coming up in the next chapter, I'll be changing course to focus on another of the new best practices in virtual and cloud computing. Chapter 3 will leverage the same kinds of data that are collected for performance and capacity management. This time, however, that data gets used for managing compliance and configuration control. You'll find that these three activities—performance, capacity, and configuration management—are more tied together than you'd think.