The Gorilla Guide to...®

What's NEXT for Enterprise IT Architectures and Solutions

Inside this book:

- Discover emerging data center architectural opportunities that can help you transform your IT organization
- Find out how composable infrastructure accelerates digital transformation efforts, reduces costs, and eliminates complexity
- Learn why the public cloud is not your only option and why private cloud environments are on the rise
- Ensure that you're current with the latest efforts in maintaining essential disaster recovery and desktop support services

Scott D. Lowe
David M. Davis
James Green

ActualTech Media
www.actualtechmedia.com

In Partnership With
anexinet

THE GORILLA GUIDE TO...

What's NEXT for Enterprise IT Architectures and Solutions

Authors
Scott D. Lowe
David M. Davis
James Green

ActualTech Media

Editor
Geordie Carswell

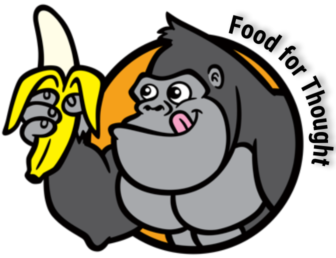
ActualTech Media

ActualTech Media
Okatie Village Ste 103-157
Bluffton, SC 29909
www.actualtechmedia.com

Table of Contents

Icons Used in This Book

Discusses items of strategic interest to business leaders.



In the these sections, readers are served tasty morsels of important information to help expand their thinking.

This is a special place where readers can learn a bit more about ancillary topics presented in the book.

Takes readers into the deep, dark depths of a particular topic.

When we have a great thought, we express them through a series of grunts in the Bright Idea section.

Chapter 1

Transforming IT with Hyperconverged Infrastructure

Although many enterprise technology trends pop up each year, the vast majority of them end up in the dust bins of history, or at least never really achieve breakout fame. It's a rarity for a brand-new technology to break out, but it's rarer still when such a technology achieves immediate hype status — and then actually lives up to the hype. Yet this is exactly the scenario that the market is seeing around hyperconverged infrastructure, a technology that, when used properly, has the potential to reshape the entire IT function by addressing challenges around data center technology, IT budgets, and IT staffing.

In this chapter, you will be introduced to the challenges inherent in the use of traditional data center infrastructure, and you will discover the potential benefits of using a hyperconverged infrastructure. You will also learn about the impact that this technology has on the IT budget and IT staffing.

Traditional Infrastructure Challenges

The traditional data center is not a panacea. Complexity abounds, and it's really expensive to operate. With a traditional data center, you're basically paying people to deploy hardware and software that often barely works together and that, in some cases, creates such rigidity that adjusting course as business needs change becomes a tremendous and expensive challenge. Further, even the technologies—such as virtualization—that were originally intended to help you ease data center pain have now turned on you and are actively working against you.

The Troubles with Virtualization

Remember when virtualization was touted as the technology that would solve all the problems in IT? Those days are long gone. As more and more organizations have approached the sought-after 100% virtualized state, cracks and even entire canyons have emerged in the infrastructure. Although virtualization has allowed the market to evolve to where it sits today, the technology has many inherent problems that need to be addressed.

Inefficient Resource Procurement, Deployment, and Management

Your traditional data center is inefficient in many ways, starting with how you actually procure your technology. If your organization is like most, you're on a three- to five-year replacement cycle for data center technology. These rigid cycles often push organizations into buying more technology than they might need. For example, sizing storage is an art, not a science. As you consider your storage needs for your full replacement cycle, there are all kinds of variables that come into play, including current raw storage needs, anticipated future raw storage needs, potential business shifts that could have an impact on storage consumption, and even the ability for data to be reduced—that is, deduplicated and compressed—to reduce overall storage capacity needs. With so many variables, people tend to over-buy storage with the idea that they'll "grow into" the investment. Sometimes they do and sometimes they don't. It just depends on how accurate those initial capacity predictions happened to be.

This is just one part of the data center procurement process. You also must worry about storage performance, servers, the hypervisor, the network, and a whole lot more. All of these pieces must be carefully constructed to best meet the needs of the business while also staying within required financial guidelines.

Even if you get the perfect ratio of data center capacity and performance, you then need to put it all together and hope that all the pieces interoperate without issue. If there are problems, you're left with vendors battling one another in their attempts to blame it on the other guy.

Management Interface Sprawl

A big part of putting together all the components of your data center is mastering the use of all of the individual interfaces that support each piece. At a minimum, you will probably have interfaces for:

- Your storage layer
- The hypervisor layer
- Backup and recovery and/or disaster recovery
- The storage area network fabric

Of course, this is just the minimum. You may—and probably do—have far more points of administration to deal with. It all really adds up. Every time you add a component, you're also adding another touch point. Each touch point requires separate skills and knowledge and contributes to overall data center complexity.

In short, although virtualization solved the problem of physical server sprawl, we've traded server sprawl for new challenges: each new component now requires specialized skills. Although the benefits of virtualization remain highly sought, pushing companies to get to full virtualization, the market needs solutions that can cover the unsustainable downside and sprawl. In fact, as you consider hyperconverged infrastructure, bear in mind that hyperconvergence is predicated on your workloads being virtualized. So, hyperconvergence is not intended to replace virtualization, but to make it sustainable and more economical.

Simplifying Console Sprawl Saves Money

Time is money and complexity is expensive in a couple of ways. First, complexity makes routine tasks take longer to complete and can increase overall staffing costs. Second, complexity can increase the error rate in those routine operations. The more touch points there are in a process, the more likely it is that someone hits the wrong button. Errors can lead to downtime, which, like complexity, costs money. With these facts in mind, it becomes apparent why enterprise IT vendors today are simplifying their offerings—in which console consolidation is often an outcome—and creating new architectures intended to drive complexity out of the IT equation.

Storage Performance

If you haven't yet noticed, storage has become *the* challenge that upstart vendors try to solve. There are myriad storage options available in the market today (some of which are discussed later in this chapter). Further, although maintaining sufficient storage capacity has often been a challenge, it's actually storage performance that really fouls this up.

Getting more capacity is generally pretty easy. You simply buy another array or shelf of disks and you're good to go. Sure, you can do the same thing to add more storage performance, but adding more disks just to get more performance seems wasteful.

Now, you may be thinking, "But isn't performance solved by flash?" and the answer is: yes, but with caveats.

First, adding flash hasn't been as simple as just throwing some SSDs into a traditional storage system. Instead, we've had to wait as vendors create solutions that can leverage flash's capabilities while at the same time solve some technical challenges inherent in flash media, such as handling garbage collection.

Second, along the way to all-flash systems, some vendors realized that there is now an opportunity to go beyond storage. They discovered that there is a chance to help address other IT challenges while solving storage performance and capacity challenges,

Read on!

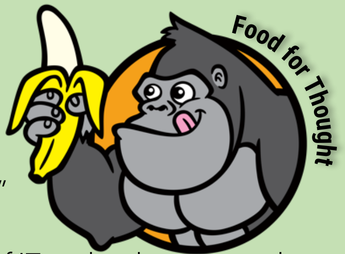
The Flattening of IT

For years, CIOs and business decision makers have wanted to create an IT function that is simpler, leaner, and far more focused on the business. To this end, we've begun to see the rise of the "infrastructure engineer" as a viable and popular IT profession. Rather than hiring many people, each with very narrow breadth but massive depth in a specific area, such as storage, the business instead hires someone with massive breadth and less depth that can manage storage, the network, the hypervisor, and the systems in the data center.

It's not just technical challenges that have created this new paradigm. All too often, in organizations that have created silos in their data center support function, there is infighting, turf battles, finger pointing, and "red tape" issues that can hold the business back. Many times, these individual groups aren't great at communicating with one another, which creates more problems.

However, to get to a point at which a more general approach to the data center is really viable, companies need data center architectures that are friendly to this approach.

And that's where hyperconverged infrastructure comes in.



The Rise of the Infrastructure Engineer

Today's "infrastructure engineer" is really just yesterday's "IT generalist" in that they have a wide breadth of knowledge about various aspects of IT or the data center, but may not have roots-deep knowledge of every single area. Infrastructure engineers are able to procure, deploy, and administer various technologies, but may not have – and may not need to have – the depth of knowledge that resource-centric specialists possess. This makes infrastructure engineers more flexible and easier to adapt to business and technology changes.

Hyperconverged Infrastructure Basics

Hyperconverged infrastructure has emerged as a popular data center architecture intended to help businesses rein in spiraling data center costs and complexity.

Defining Hyperconverged Infrastructure

At its most basic, hyperconverged infrastructure is defined as a conglomeration of storage, compute, and a hypervisor, all neatly packaged on a single server, with no separate SAN or storage array needed. All of the storage resides locally inside each server. At first glance, many people look at hyperconverged infrastructure and their initial reaction is, "Why would I want to go back to the days of direct-attached storage?" However, while it's true that hyperconvergence does bring you back to those days by fully eliminating the stand-alone storage layer, it does *not* bring back the downsides inherent in that approach to storage.

Unlike the old days in which all of that storage was trapped on individual servers, hyperconverged infrastructure binds all of that storage together with a powerful software layer connecting all of the nodes over a standard Ethernet network.

In essence, you're binding all your storage and creating a distributed, highly available storage fabric that's fully integrated with the server and hypervisor components.

Software-Defined Storage

People sometimes confuse *software-defined storage* (SDS) and *hyperconverged infrastructure* because there is a lot of similarity between these two approaches to storage. In the world of software defined storage, it's sufficient to have a system that abstracts storage hardware resources and aggregates those resources for eventual consumption by connected hosts. The aggregation process is handled by a software-based controller of some kind.

Hyperconverged infrastructure can be considered a subset of software-defined storage in that there is a software-based controller of some kind (we'll discuss this later in this chapter) which effectively abstracts the physical storage layer, and presents it to a higher order construct. In the case of hyperconverged infrastructure, that construct can be a hypervisor (for in-kernel storage management) or a controller virtual machine that is assigned the task of managing local storage on an individual node. The key difference between SDS and hyperconvergence is that this next-level construct also operates on the same hardware as the storage. This integrated storage component is one of the major selling points of hyperconvergence.

In non-hyperconverged infrastructure software-defined storage systems, the actual storage devices look a lot like traditional storage arrays. You create connections from host machines using your protocol of choice, and the storage layer places data on correct nodes behind the scenes. Meaning, there is physical separation between hosts and storage.

The Role of Hardware

Up to this point, we haven't said much about hardware in this new world order. That's intentional. With hyperconverged infrastructure, hardware is actually a secondary consideration. You still need to have hardware that can perform well, but you don't necessarily have to go out and buy from a specific brand. Instead, for the server-based nodes in your hyperconverged infrastructure cluster, most hyperconverged infrastructure vendors provide multiple server vendor options from which to choose.

Many equate this deemphasizing of hardware as the hardware being unimportant. That is simply not true. Hardware choice is critical no matter what data architectural option you select. The difference here is that you can take a commodity, off-the-shelf approach to hardware and you don't generally need to buy very specific brands.

Architectural Options and Capabilities

You've already been introduced to some of hyperconverged infrastructure's architectural options in this book. Now, let's discuss the important nuances of making your selection in the following sections.

Virtual Storage Appliance (VSA) vs. In-Kernel Storage Management

There are various ways by which local per-node storage can be managed, and it's important to understand the overall operating paradigm by which storage in these systems works. As mentioned earlier, storage is abstracted on each node, and management for that storage is assigned to a local software-based controller (Figure 1-1), which runs as either a virtual machine or as a kernel module.

Each node-based controller coordinates its activities with all of the other controllers on the other nodes and the cluster, with each contributing their local storage capacity to the cluster-based storage pool. The resulting storage then becomes available cluster-wide, while the individual controllers are simply responsible for managing their own resources.

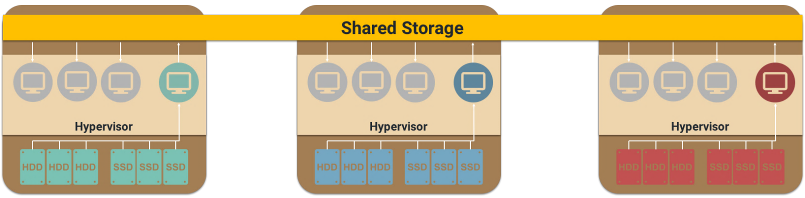


Figure 1-1: Storage management in the world of hyperconvergence

The choice of whether to choose a hyperconverged solution that uses a virtual storage appliance (VSA) versus one that manages storage using an operating system or hypervisor-based kernel module is one that many pixels have been spent discussing. There are people that feel passionately on both sides of the debate. So, here's a bit of guidance.

A VSA typically enables a bit more flexibility, particularly when it comes to hypervisor choice. After all, it's just a virtual machine that happens to be running a storage controller. Virtual machines can be easily migrated from hypervisor to hypervisor, making it easier for VSA-centric solutions to support multiple hypervisors. On the downside, is the concern that forcing all host storage I/O through a virtual machine imposes a significant performance challenge. While it's true to a point that a VSA will have a bit more latency than in-kernel options, the performance results are far from poor. Also bear in mind that there is a VSA on every host, so the overall I/O load is distributed.

In-kernel storage management constructs may perform a very little bit better than VSAs, but even that's debatable as there are a lot of variables that come into play. In arguments favoring in-kernel over VSA storage management, many point to the fact that a VSA uses significant host resources — RAM and CPU — to accomplish its management goals. However, in-kernel management systems also have similar resource requirements, so this argument becomes a moot point.

Frankly, most customers don't care about VSA vs. in-kernel, but it's one that hyperconverged infrastructure vendors use in their marketing, so it's important to understand the differences. The key item to remember is that an in-kernel system effectively locks a customer in to that hypervisor. However, while "lock in" is generally considered as something to avoid in IT, you're really locked in only for the duration of your replacement cycle.

Software vs. Appliance

Another significant choice you will need to make in your hyperconvergence journey is between a software-only solution or a bundled appliance, which includes server hardware and the hyperconvergence software. This choice is one that you need to make carefully.

With a software-only solution, you purchase the hardware and software that will comprise your hyperconverged infrastructure cluster separately. With a software-only solution, you get to choose the server vendor you want to use as well as specify the exact hardware you intend to deploy. You truly get the full range of choice when it comes to hardware. However, in some cases, you also have to personally install the software on the hardware, which can add complexity and extend time-to-value. However, there are software-only vendor solutions that take what's known as a "meet in the channel" approach in which you're able to get an appliance from any server vendor you choose while still leveraging the software-only tool. By taking this approach, you get to avoid the need to manually integrate the hardware and the software and you still get to choose your own hardware.

In an appliance model, you buy hardware and software as a bundle. The appliance vendor decides what hardware you can use, by generally offering a range of hardware models, each with distinctive hardware characteristics and features. When it comes to hardware choice, though, more and more, appliance-centric hyperconverged infrastructure vendors are forging partnerships with multiple server vendors, so you can often continue to work with your server hardware of choice with only some resource restrictions imposed by the vendor.

Data Efficiency

With most hyperconverged infrastructure solutions on the market, you will discover that there are some data efficiency capabilities built right into the platforms. Such features include data deduplication and/or data compression. Data compression works by making files smaller while data deduplication works by removing duplicate blocks of data from a storage system.

Used individually, these technologies are powerful. Together, and with the right workloads, you can often reduce the overall capacity necessary to store your data by huge factors. The key phrase though, is "with the right workloads." If you're running workloads that already have highly compressed data, such as a system that stores video files, compression won't do much more. On the flip side, if you're running a workload in which there is a ton of commonality, such as virtual desktops, the deduplication levels can be significant.

Figure 1-2 gives you a quick overview of various HCI options.

	VSA	**In-Kernel**
VSA vs. In-Kernel storage management	**+** Provides support for multiple hypervisors	**+** Marginally improved performance (very small margin) **-** More easily locked in to the vendor since an in-kernel solution is tied to the hypervisor (but… not really that critical)
	Hardware appliance	**Software**
Software vs. Hardware appliance	**+** Ready to roll immediately; no integration needed **+** Simple to purchase **+** Single vendor support **-** Less hardware flexibility than software alone **-** Can be more expensive than software alone	**+** Eminent hardware flexibility **+** Supports servers from any vendor you choose and that meet compatibility standards **+** Easier to "try and buy" **+** May allow you to leverage existing hardware assets to save money **-** May require manual software/hardware integration

Figure 1-2: The pros and cons of various approaches to hyperconverged infrastructure

As you consider your hyperconverged infrastructure journey, it's critical that you understand how your workloads operate so that you can buy enough nodes to meet all of your storage capacity and performance needs.

The Staffing Dilemma

At this point, you understand how hyperconverged infrastructure works and, if you're like most, you're wondering what impact it has on IT staffing. After all, if you go all-in on this technology, you're fully eliminating the storage silo in the data center, a shift that has obvious ramifications, especially for your storage staff.

That's why it's important to understand that IT is undergoing what amounts to a seismic shift in how IT needs to operate. The days in which IT infrastructure support groups are broken into individual groups are coming to an end as businesses seek to streamline and make more efficient the overall operation of the department. Instead of hiring subject matter experts with narrow but deep expertise, businesses are beginning to take a more general approach by hiring people with wider experience, but not quite as much depth in expertise. Therefore, as the infrastructure gets simpler to deploy and to manage, it becomes far easier to staff infrastructure positions this way. In short, the market is seeing the rise of the Infrastructure Engineer.

This provides several major benefits to an organization. First, over time, the data center can be managed with fewer people. You no longer need to have individual teams, with redundancy on each team. Now, it's possible to have one or two teams — a network team and an everything-else team — that handles it all. Second, this new paradigm opens budget lines that IT can leverage to focus more on business-facing activities. This is the holy grail for the business because business analysts can help drive top line revenue rather than consume expense budget.

The challenge though is getting from today's structure to tomorrow's. There will be obvious disruption to the department and to people's careers. However, in the world of IT, it's critically important that staff take charge of their own career and stay current on their skills. Your organization should do everything possible to retrain people and to ensure that they have modern skills that best meet the needs of your business.

Retraining Retrofit

Part of the reason that some people raise concerns around hyperconverged infrastructure is because of the potential impact on their livelihood. If there's no more storage to manage, then what do these people do? Fortunately, there is no end to the demand that is on IT departments today, but it will take concerted effort on behalf of both individuals and companies to retrain people for a different set of responsibilities. Storage people tend to be very technical in nature, and are well-suited to other technical needs in an organization. Areas of increasing focus in the enterprise include security and cloud integration, two areas that demand a high degree of technical skills.

Chapter Summary

It's clear that hyperconverged infrastructure carries with it the potential for significant IT and business transformation. On the technology front, hyperconvergence is simpler to deploy and manage and can reduce the time and staff resources that you must dedicate to the data center. Through the strategic deployment of simpler technology, hyperconverged infrastructure allows IT to begin to refocus their efforts more on the business and less on technology.

Chapter 2

An Introduction to Composable Infrastructure

Every year, it seems like marketing pros invent new words and phrases. This time around, the term is “composable infrastructure,” and it’s gaining ground with more and more tech companies that are working overtime to differentiate their data center infrastructure offerings from other solutions on the market. However, unlike the technology that underlies so many other buzzwords that have come and gone, composable infrastructure actually has some potential benefits that are important to understand as you continue to survey the modern data center infrastructure landscape.

The New Way of IT

If you’re here, you may suspect that your data center is changing, but, at the same time, so is your business. That change demands an IT function that can keep pace with the business and that doesn’t result in adding complexity or delay to the change. Traditional IT infrastructure can often create complexity that inhibits change. Too many legacy IT environments are overly complex and require complicated change management processes to adjust as the business adjusts. Further, as the legacy environment does expand, such expansion can be somewhat impactful and may require downtime and other disruptors.

At the same time, other data center architectures—such as hyperconverged infrastructure—are emerging that, while intended to provide a boost to IT and the business, may sometimes not be fully suitable for the workloads that are deployed.

An emerging data center architectural opportunity is arising in the form of what has become known as composable infrastructure, which will be described shortly. Before we jump to that, let's take a quick look at three key characteristics that define how IT must function now and in the future.

Any Time

As you contemplate new data center infrastructure, you need to consider the fact that you need to be able to add more infrastructure on demand, when the business needs it, not weeks or even months later. Sure, it may take a bit of time to procure additional hardware, but once it's on your dock, it should be deployed in hours or days. Additionally, once you have physical infrastructure deployed, as new workloads need to be brought into operation, making that activity happen should be seamless and, where appropriate, automated. This provides the business with almost cloud-like flexibility and agility.

Anywhere

Bare metal, traditional virtualization, and containers are all viable workload execution platforms today (and we'll talk about those later), but some infrastructure solutions don't always make it easy to mix and match workloads across these platforms. For example, even with the greatness that you may get from hyperconverged infrastructure alone, if you have workloads that can't be virtualized or containerized, you can't operate those in a hyperconverged world, which requires that all workloads be non-physical.

For many businesses, a data center environment that can handily support workloads that can simultaneously run on bare metal, atop a hypervisor, or nested cozily inside a container is critically important. It is this kind of flexibility that ensures that the business can rely on IT be have the necessary agility to support a constantly changing business landscape.

Any Workload

Finally, you need to make sure that whatever data center architecture you're operating can support the wide variety of workloads that are available in the market. You should have the capability to support traditional tier 1 applications such as SharePoint, Oracle, and other services, but you should also have the ability to deploy virtual appliances and microservices-focused container-based applications. Your infrastructure should not limit you to running, for example, only virtualized applications. For most organizations, that would mean having to deploy silos of infrastructure in order to support workloads that have different hardware requirements.

Introducing Composable Infrastructure

You may be wondering how you manage to get to this seeming panacea of data center goodness. The answer lies in the use of *composable infrastructure*, an emerging data center paradigm that goes beyond what is possible with legacy infrastructure, converged infrastructure, and even hyperconverged infrastructure.

There is an increasing need for IT to be able to support both traditional applications and well as emerging ones. This phenomenon is sometimes referred to as "bimodal IT" which is just a way of saying that you need to be able to support those business critical legacy applications while also making it really easy to build new, innovative systems. Whereas the public cloud is often considered to satisfy the second part of the bimodal equation, what if you could bring the necessary cloud-like characteristics to your local data center instead?

That's part of what you get with composable infrastructure but you also get so much more. The primary characteristics and benefits that define composable infrastructure are defined in the following sections.

Fluid Resource Pools

The real secret sauce behind composable infrastructure is its support for what is often referred to as *disaggregated resource control*. This term basically means that you have the ability to independently scale resources at will. If you need more storage capacity, you simply add more disks to the system. The system is architected in such as way to as to make it dead simple to scale in this way without having to incur downtime or complex expansion procedures. Sure, you've been able to scale resources independently in the traditional world forever, but those resources were always part of the same management plane. When considering composable infrastructure, you're also looking at ways to far more seamlessly deploy and manage all of the various resource areas.

This fluidity makes it very easy to granularly target resource deployment where you need it most. You're not forced to live within predefined limits and resource ratios established by a vendor as you are with hyperconverged infrastructure, for example. The world is yours!

Workload Support Environments

Workload infrastructure is a pretty interesting place and there is far more variation than people may realize. Although many organizations have made it their mission to eliminate physical workloads in favor of virtualized ones, that's not always the right answer. Further, there are emerging workload services that need to be considered, particularly since they align so well with new software development methodologies.

Physical

If you thought the physical server was dead, think again! Even today, well over a decade since the inception of VMware, physical workloads remain popular for a number of reasons. We still have applications that need the kinds of performance that only a physical server can provide and there remain applications that are less than friendly toward virtualization from a licensing perspective. Although supremely frustrating, there are even companies out there that still refuse to support their applications if they're installed on a virtual machine. With those thoughts in mind, it's clear that physical servers will need to enjoy continued support for the foreseeable future, but that doesn't mean that they should be considered an exception. In fact, new workload dynamics may mean that you should make sure that these systems are just as well supported as everything else in your environment.

Virtual

Today, standard mainstream workloads are generally virtualized. Over time, the percentage of systems that have made the jump from physical to virtual has steadily increased, particularly as hypervisor vendors have continued their efforts to make virtual machines first-class citizens. They have accomplished this by enabling what have become known as "monster VMs" and by continually adding more and more capability to their products.

It's clear that, today, virtual machines rule the roost and get most of the mindshare, but that doesn't mean that they should get special treatment. After all, (almost) all your workloads are probably very important to you, no matter where they run. Your infrastructure should reflect that. And, over time, you should expect to see newer workloads begin to operate in newer environments.

Container-based

Which brings us to containers: the new kid on the block in some ways. Although containers have actually been around forever, in recent years, they have gained a new lease on life as companies like Docker have found ways to make them easy to use, which has brought them to the masses. Whereas virtual machines force you to install a separate operating system instance for every single virtual machine, each of which requires a certain baseline of resources to run the operating system, containers work inside a single operating system instance (Figure 2-1). So, rather than have sixty copies of a Linux server running, with each one supporting a discrete task, you can have those sixty services running inside containers on a single operating system instance.

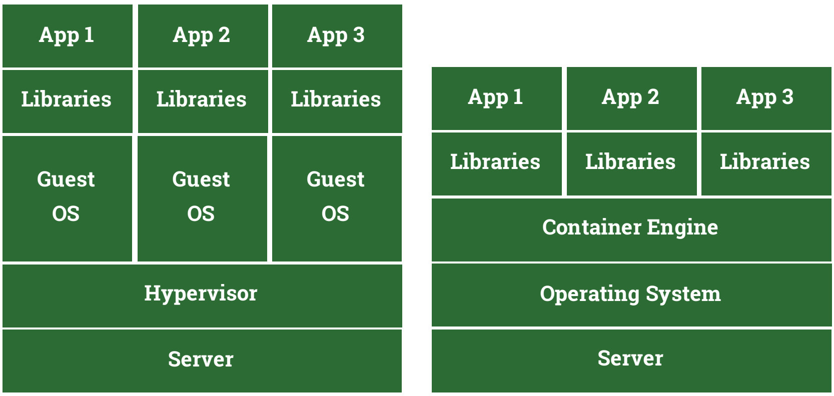


Figure 2-1. Virtual Machines vs. Containers

Containerized applications are packaged with their individual dependencies and configurations, making them eminently portable across systems. Moreover, due to the fact that containerized workloads run atop operating system provided binaries and libraries, containers can operate in both physical and virtual environments.

As an aside, as you review the world of containers, you will probably see people asking if you should be doing containers or if you should be doing virtualization. It's really kind of a misleading question since you can do containers on top of virtualization. Further, they're both simply workload foundations. The choice you make around workload architecture is really dependent on the workload you're running. If it's a traditional application, you run it in a virtual machine. If it's a microservices-like app or containerized, you run it in a container (which, again, may be hosted on a virtual machine).

Figure 2-2 compares the management touch points among several data center architectural options.

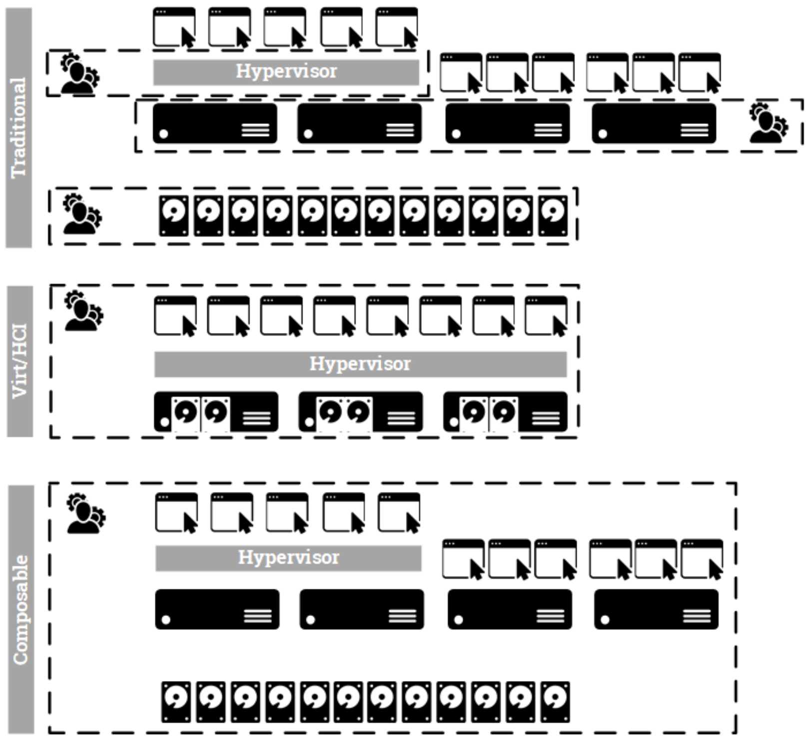


Figure 2-2. Traditional infrastructure vs. hyperconvergence vs. composable infrastructure

Scaling

If you're like every other business on the planet, you want to grow over time. That means that your IT systems have to be able to scale in lockstep with your business. However, that doesn't mean that today's scaling options are always particularly good. At times, it can be downright difficult to grow resources the way you need and linear scaling available in hyperconverged infrastructure isn't always the most desirable choice for businesses.

With composable infrastructure, scaling is both simple and granular. You can scale any resource you need individually by simply adding more of that resource and the management system seamlessly assimilates those resources into the collective whole and makes them available for use.

API-driven IT

You may have heard and, if you haven't, will certainly start hearing the term "infrastructure as code" bandied about. What this really means is that there is a need to more tightly couple applications and the infrastructure on which those applications reside. There are a number of outcomes that can be realized from treating infrastructure as code, such as:

- Enables developers to programmatically deploy new virtual machines and other structures so that they can more quickly test their code
- Allows an application to automatically instantiate new infrastructure based on performance conditions at the present time
- Enables automation-based user self-service, which is a prerequisite to be considered a private cloud

You should expect to see deeper and deeper integration between applications and infrastructure over time. There is no reason that applications can't be far more self-managing than they are today. With the right tools and the right infrastructure, admins can be brought to bear on an exception-only basis.

How It Works

At this point, you might be wondering how composable infrastructure actually works and what makes it tick. Composable infrastructure is a combination of hardware and software that you deploy into your local data center. On the hardware front, you buy equipment that is composable ready. There are an increasing number of solution vendors selling such equipment, although the original composable player is HPE, who sells an entire line of composable-specific hardware as a part of their Synergy lineup. With this lineup, HPE makes available servers, storage, and networking that all exists inside the confines of a series of blade chasses.

Of course, hardware alone is not sufficient. You also need software tools that leverage this hardware and that provide the actual services upon which your workloads operate. Again, all vendors that sell composable-branded solutions have a software layer, but HPE has the most well-known. HPE's Composer tool allows you to get the fluid resource management that is at the heart of composable infrastructure. Taking things one step further is HPE OneView, which adds APIs atop your Synergy environment, which unlocks the full potential of composable infrastructure.

The Business Benefits of Composable Infrastructure

There are a number of clear business benefits that can be enjoyed by those that choose to deploy composable infrastructure solutions in their data centers.

Increased Agility and Flexibility

Most importantly, a business with a data center environment that leverages composable infrastructure to its fullest extent can enjoy significant levels of flexibility and agility. The flexibility comes in the form of ease of scaling individual resources and agility comes from being able to deploy workloads without having to worry about what the underlying infrastructure needs to look like.

This is especially important as the pace of change in the IT landscape continues to increase. More than ever, IT pros need to deploy infrastructure that can keep up with the constantly changing technology market and integrate new innovations as they become available.

Business Accelerator

With a workload operating environment that is based on completely fluid API-driven resources, it becomes possible for IT to become far more cloud-like in the services that it offers. You can more easily automate your processes and orchestrate multiple processes in to streamline overall data center and workload management. Taken to the right level, or with the right tools, you can get to a point in which users in business units can provision their own systems and application developers can build systems that manage their own resource usage, expansion, and contraction.

Workload-Centricity

The whole point of the data center environment is to run workloads, but, for far too long, infrastructure has been in the driver's seat and has really dictated how workloads will operate. As we enter the era in which infrastructure is invisible to the user, programmable, and granularly and fluidly scalable, the workloads take center stage, which is as it should be.

Chapter Summary

Composable infrastructure represents another significant opportunity for data center architects. It provides businesses with the ability to build on-demand data center services with fluid resources that can support workloads regardless of underlying workload needs. It's API-driven nature makes it a must-have for DevOps-oriented environments and it carries with it significant benefits that can help to accelerate the business.

Chapter 3

New Business Possibilities Abound with Private Cloud

Although the public cloud gets a lot of attention these days, most organizations still operate local data centers. Even so, decision makers in these organizations often lament these local installations as they've heard about all of the goodness that can come from the public cloud. However, that kind of thinking reveals the real reasons that decision makers often consider the public cloud to be a panacea. It's not the fact that something is running in the public cloud, but revolves more around the *outcomes* that these organizations enjoy, such as modified economics, easier scale, and improved efficiency. In this chapter, you will explore the private cloud and gain an understanding of what this term means and how private cloud can benefit your organization.

Private Cloud 101

The term private cloud was coined in order to describe an on-premises infrastructure that looks and acts like a public cloud environment, but serves the needs of just a single organization. There are a lot of myths and misconceptions about private cloud. In this section, we'll walk through some of those misconceptions to help you better understand the term and what it means.

Virtualization vs. Private Cloud

You may have heard people say things like, "We're 100% virtualized, so, therefore, we have a private cloud." Not so fast! If you're 100% virtualized, you are just that... 100% virtualized. Without some additional characteristics, you're not operating a private cloud. These characteristics include things like easy and granular scale, user-self-service, and cloud-like economics. We'll chat about these characteristics in more depth later in this chapter, but the key takeaway for this section involves terminology. "Virtualization" does not equate to private cloud without the characteristics you'll learn about later on in this chapter.

Public Cloud vs. Private Cloud

Simply put, public cloud means that you're renting time on someone else's servers while private cloud means you're operating your own. There is, however, some nuance here. In general, private cloud is often considered as you owning all aspects of the environment, including the data center physical space. There is absolutely an in-between state in which you may lease data center space—for example, in a colocation scenario. In general, if you own the hardware, which includes servers, networking equipment, and storage, and you've added to that environment the magic dust that transforms it into a private cloud, you can carry the private cloud label with honor. If you are simply renting capacity on someone else's servers or you have not added that magic dust, which we'll talk about shortly, then you're not a private cloud.

Hybrid Cloud vs. Private Cloud

A lot of people have asked, "Do I have a private cloud or do I have a hybrid cloud?" In general, the term hybrid cloud means that you have some combination of public and private cloud operating your workloads. However, there's a bit more to the story. If you have completely disconnected public and private cloud environments that don't talk to one another whatsoever, there is some thought that these aren't really hybrid cloud environments. Once you begin bringing these environments together, they merge to form a hybrid cloud workload operating environment.

Private Cloud Characteristics

What takes an environment beyond the realm of simply "virtualized" and into the realm of private cloud are a series of characteristics that are modeled on the successes that cloud providers that realized in the public cloud. These characteristics come both in the form of the business model (the way the data center business unit operates) and in terms of technical infrastructure features. In the following sections, you'll learn about some of the ways that an organization could potential upgrade a virtualized data center into a private cloud.

Economies of Scale

Although some would say it's *technically* possible to operate a private cloud on a small scale, the same doesn't always hold true from a business perspective. Part of what makes the consumption of public cloud resources so attractive is that public cloud operators like Amazon Web Services and Microsoft are able to purchase infrastructure in massive quantities due to the scope of their operation. These huge orders allow them to negotiate great bulk pricing from their suppliers and the savings are passed on to the public cloud consumers.

While not all business that operate their own data center are able to take advantage of this principle, all data center operators can seek to find ways to structure their purchases in a way so as to reduce costs. Some large private data center operations, in fact, *do* have the scale to support a more cloud-like purchasing model.

In either case, running a true private cloud means structuring costs in a way that takes advantage of every bit of scale the business is working with. As an example, if multiple business units within an organization currently foot their own data center bills and do their own data center purchasing, moving data center purchasing to an organization level instead of a business unit level could allow some organizations to experience the economies of scale that public cloud operators enjoy.

OpEx-based Economics

The hallmark of a public cloud model is usage-based billing. Rather than a large, up-front financial outlay which is then amortized over many years, public cloud consumers pay a frequent fee (usually monthly) for only the resources they consumed during that billing period. This means that rather than purchasing hardware today to support the needs of the business for the next five years (the estimation of which can be quite tricky), a business needs only to foot the bill for the infrastructure their current application and data needs consume.

In the private cloud world, there are two major ways to take advantage of this concept. The first way is to leverage a third party that leases data center hardware. In this model, rather than purchasing three to five years' worth of infrastructure in one swing as a capital expense, data center operators can lease equipment as they need it and pay the recurring leasing invoice out of the monthly operations budget instead. The second way to take advantage of this principle is to be sure that the data center scales quickly, easily, and in small chunks.

Easy and Granular Scale

By scaling a little bit at a time, a business is able to purchase only data center infrastructure that is needed today and buy the rest later. Historically, when scaling was cumbersome and expensive, it made more sense to buy for five years out in order to avoid the upgrade process. With the ability to scale granularly and without all the headache of data centers past, buying for the immediate future instead of the distant future can potentially turn what used to be a multi-year capital expense into a quarterly OpEx expense. From a budgeting standpoint, this is often much easier to manage and preferable to businesses today. While the success of this model was proven in the public cloud, private infrastructure exists today to allow private data center owners to bring this model home with them.

Operational Efficiency

If a public cloud provider operated with all the operational efficiency of most private, enterprise data centers from five years ago, most of those public cloud providers would be out of business or in a downward spiral. *Critical* to the success of the public cloud model is the insistence on operational efficiency. This comes by way of ruthlessly removing cruft and unnecessary clutter. Even more than that, it happens when data center architects focus on removing the human element from data center operations as much as possible. We've known for a long time now that computers are much faster and more accurate than humans. Now that technology has matured to the point where we can outsource many of the data center operations tasks to the data center itself, it makes more sense to remove the humans. Increased levels of automation and orchestration in the data center decreases the likelihood of human error, increased the speed of provisioning, remediation, and upgrades, and allows data center operators to focus on the creative tasks that humans are really more suited to. Public cloud providers have been leveraging automation from the very inception of the model; there's no reason it can't be leveraged in a private data center too, to turn that data center into a private cloud.

User Self-Service

One of the reasons that IT departments over the last decade have spent so much time dealing with the phenomenon of "shadow IT" - where users turn to external providers for their IT services instead of the internal IT organization – is because of the self-service nature of the public cloud. Consider this example: within the first couple years of the existence of Dropbox, a user's needs to store and share data with his or her colleagues could play out in one of two ways:

1. **Request a new file share from corporate IT**. Wait a week for a call back to clarify what they wanted. Wait another week for provisioning to happen and the ticket to be closed. Call back and re-open the ticket because the quote on the file share was too small. Finally share files with their colleagues. **Time to value: 2.5 weeks.**

2. **Create a new Dropbox account**. Share the files. **Time to value: 2.5 minutes.**

We're becoming increasingly impatient as a society, and IT infrastructure is no exception to this trend. Therefore, to operate a private cloud, it's mandatory that some level of self-service capability exists.

This could be end-user-facing as in the Dropbox example, but at the very least, it should be administrator-facing. For example, a database administrator might be able to request a new database server instance from the self-service portal. Behind the scenes, and new virtual machine is created (on appropriate storage resources), networking information is obtained and applied, the database engine is installed and configured, the operating system and database application are patched, and security settings are applied. At the end of the process, the requesting DBA is sent access information, and no data center operator has lifted a finger. This vision for self-service has been perfected in the public cloud, and it's essential to running a proper private cloud.

Chargeback/Showback

Finally, billing in the public cloud world is simple. If a number of departments manage their own IT budgets, they can just have their own accounts with the public cloud providers and pay their own bill. Once of the challenges to overcome with turning a traditional data center into a private cloud is the need to replicate some of the multi-tenancy features of a public cloud. Even if separate billing from the IT business unit isn't actually being done back out to the separate business units ("chargeback"), most organizations today at *least* want the ability to break down IT spending by department (this is known as "showback"). The key to bringing this public cloud characteristic in-house is to implement some sort of higher-level data center management tool that treats different parts of the organization as different entities. In the sense that different organizational units become "tenants" in this multi-tenant environment, a private cloud really becomes a miniature, self-contained scale model of a public cloud.

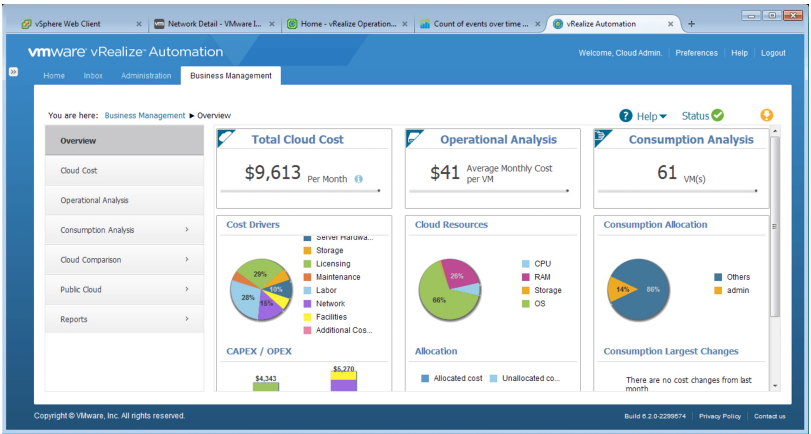


Figure 3-1. VMware vRealize Automation's Chargeback and Showback Capabilities

Migrating from Traditional to Hybrid/Private Cloud

"Private cloud sounds great!" one might say, "But how does a traditional data center really begin to adopt cloud-like practices?" It's important to understand that there are some applications that were developed with a cloud-like infrastructure in mind; you may hear these referred to as "cloud native" applications. On opposite end of the spectrum, many enterprises today are running legacy software and hardware with a history measured in decades.

The culmination of these two facts is that the future of enterprise IT is a hybrid cloud scenario for most businesses. Some workloads will be best suited for a cloud – be it public or private. Some workloads will simply never be migrated from where they sit today, either for technical, cultural, or financial reasons. So rather than look just at how to move to the private cloud, consider the possibilities across the spectrum of the hybrid cloud and the improvement that landing anywhere new on that spectrum might make in terms of capturing the cloud-y data center characteristics described above.

Workload Placement Decisions

In the hybrid cloud would, besides public and private, there are a few other buckets that workloads can be placed in to. The era of virtualization is very much upon us, both by way of machine-based virtualization (vSphere, Hyper-V, KVM) and in the way of OS-based containerization (LXC, Docker, rkt). However, despite what cloud pundits may say, physical servers aren't going anywhere any time soon. Physical servers still have a significant footprint in the data center, especially when it comes to certain workload types, like databases and messaging.

Virtual

If an application is suitable for virtualization, then the question becomes which cloud environment should the workload reside in? Given the unique expense structure of each cloud, there are a variety of reasons a data center architect might choose to run a workload in a public cloud or a private cloud. Some considerations include:

- Compute requirements (CPU & Memory)
- Storage capacity requirements
- I/O performance requirements
- Data protection requirements
- Network latency and throughput requirements
- Workload longevity/lifespan
- Integration and/or dependency on other applications
- Regulatory/compliance requirements
- Security requirements
- Staff skill set in one cloud or the other

And that's the short list! Suffice it to say that even once an application has been defined as 'compatible with virtualization,' there's still a pile of details to sort through before settling on public or private cloud residence. This list will be a good starting point.

Physical

As mentioned earlier, physical servers still make up a remarkably sizable portion of total data center space. And as much as your first thought might be that these are simply legacy workloads that can't be moved the to cloud – and that would be a reasonable thought – you'd be wrong. In fact, that sentiment couldn't be further from the truth. Some of the most cutting-edge architectures with a focus on technologies like container hosting and big data rely heavily what physical servers. Is a post-virtualization world, you'll commonly hear this design referred to as "bare metal." It sounds new, hip and cool that way.

SaaS (Public Cloud)

There's a third option for workload placement that takes away all of the debate about how the specific architecture for a workload should be handled. SaaS, which is an abbreviation for Software-as-a-Service is the notion of outsourcing an entire software application to a third party. That third party is responsible for every part of the infrastructure below the application and all the consumer needs to worry about is the application itself. Perhaps the most iconic piece of software in this space, and undoubtedly a pioneer in the model, is Salesforce.com. This massive, S&P 500-listed company will take the infrastructure needs for your CRM entirely upon themselves. All a consumer needs is an Internet connection, and they can leverage the full breadth and depth of the Salesforce.com solution.

This is a very appealing model to many organizations today because it does more than just leverage the aforementioned cloud characteristics to improve economy and experience, but it also outsources the actual maintenance of the application as well. In some large enterprises, there are full-time employees whose sole job it is to upgrade a messaging platform. All day, every day. And when they finally finish, they'll already be behind on getting upgraded to the next version. In a SaaS model, the service provider is responsible for maintenance upgrades, security (to an extent), data protection (again, to an extent), and the full burden of managing the underlying data center falls on the as well.

Hyperconverged Infrastructure and the Private Cloud

When it comes to the practical application of cloud-like characteristics in the data center, it's one thing to talk about it, but it's another thing entirely to actually do it. For example, to say that the data center should scale more easily and granularly is a simple statement, and one with which most people who've spent any time around a data center would agree. However, without a significant paradigm shift in the way hardware and software are provisioned and married, not a whole lot of real change will be made.

Hyperconverged infrastructure is one answer to this problem. The entire purpose of hyperconverged infrastructure is to reduce friction in daily operations, to allow scaling in predictable units, and to make the implementation of that scaling process as effortless as possible.

Hyperconvergence is a total overhaul of the methodology with which data centers have been built and managed over the last decade. Rather than making the focus of the design the server or the hypervisor as it has been in the past, hyperconvergence makes the focus the virtual machine and in this way dramatically simplifies the design and management paradigm. All the complicated architectural bits are taken care of behind the scenes.

Composable Infrastructure and the Private Cloud

The composable infrastructure paradigm is cut from the same cloth as the idea of Hyperconvergence – that is, the idea that the unit of management in the data center needs to change. While the focus of Hyperconvergence is on eliminating focus on the server, hypervisor, and storage layer of the infrastructure, the composable infrastructure paradigm addresses the infrastructure one level below this in the infrastructure stack.

The idea of composable infrastructure is to disaggregate physical infrastructure components in such a way that they can then be provisioned (and re-provisioned) logically and at will. Rather than purchasing a physical server from Supermicro that has two quad-core CPUs at a certain speed, a fixed amount of memory, and certain storage and networking characteristics, a composable infrastructure model takes a different approach. An administrator can provision a logical server which consists of any combination of compute, storage, and network fabric resources the administrator desires. As the needs of the workload change, the logical container is flexible and any one of those configuration variables can easily be changed.

Public or Private Cloud: The Difference Matters

In a time when the cloud seems to consume all new initiatives, there is a raging debate about what should and should not move to the public cloud. Let's step back for a second, though, and look at why you may *not* want to move to the public cloud and may want to adopt a private cloud instead. First, for most companies, no matter how much you place into the public cloud, you will still need a local data center presence anyway, so you need to plan for that presence. Second, some workloads simply aren't a fit for the cloud, and, as such, need to remain local. Finally, there remain real and valid concerns around security, control, and even cost. Stories abound of companies pulling back on their cloud efforts as costs spiraled out of control. Do your homework and research and, regardless of your public cloud plans, don't forget to address your local data center, too.

Composable infrastructure has private cloud written all over it. The composable infrastructure model was developed to get at the heart of what administrators really want from cloud: to be able to provision and use only what is needed at any point in time; to provide a fungible pool of resources that can be provisioned and re-provisioned as requested from a self-service portal; and to have high operational efficiency such that it can flex as the needs of the business flex.

Chapter Summary

It should be clear after your reading of this chapter that the outcomes associated with public cloud infrastructure are not necessarily out of reach for those organizations where public cloud consumption is not a good fit for some reason. A private cloud can have most of the same qualities that public clouds have and remain under the control of a sole business. But indeed, the future of IT is neither public nor private clouds alone. Instead, the future is a hybrid model in which public and private clouds collide in a whirlwind of virtual, physical, and SaaS-based workloads. Remember, the only objective at the end of the day is to satisfy the application and data needs of the business as efficiently, reliably, and cost-effectively as possible.

Chapter 4

Discover the Next Phase of Disaster Recovery

Every company knows that you need tested backups of your data, and those backups must be kept offsite. What most companies don't know, however, is that new data-protection and cloud technologies take disaster recovery to the next level. In this chapter, you'll learn about those new technologies, how they work, and how they will modernize your disaster recovery systems to ensure that your enterprise is prepared when unforeseen disasters, small or large, hit the data center.

Challenges with Traditional Disaster Recovery

While most companies have a disaster recovery (DR) plan of some type, there are still many companies who don’t have any plan at all. The sad thing about this fact is that, according to the National Archives and Records Administration (NARA) in Washington, D.C., 93% of companies that lost their data center for ten or more days due to a disaster filed for bankruptcy within one year of the disaster. Additionally, 60% of companies that lose their data shut down within six months of the disaster.

Hopefully you aren’t employed by one of the 48% of companies surveyed by ActualTech Media that don’t have a plan and, if you are, it’s time to make DR a priority. After all, disasters **do** happen and many of those disasters aren't large-scale catastrophes like fire, flood, tornado, or hurricane. Most disasters are much smaller in scope, such as human error, unexpected data corruption (e.g., due to a firmware upgrade on the SAN), or ransomware attacks.

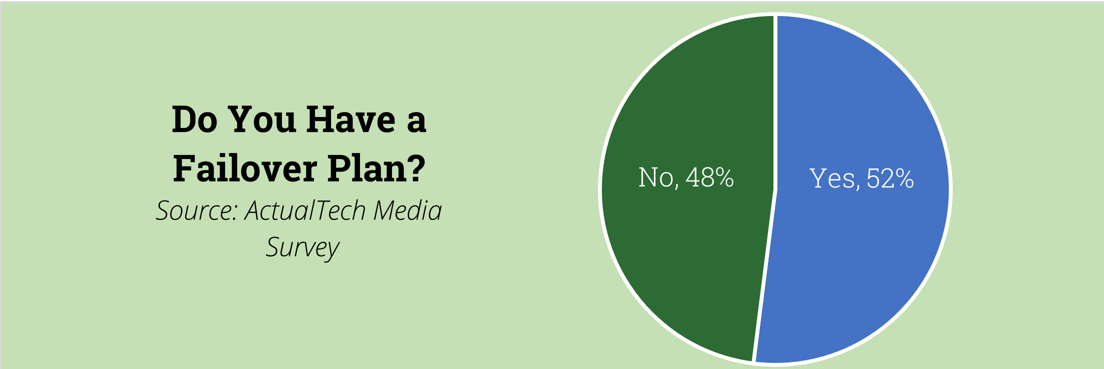


Figure 4-1. A look at who does and does not have failover plans

For companies that **do** have a DR plan, so many of those plans are out of date, using outdated technologies, **or both.** Let's review some of the challenges that enterprise companies face with traditional disaster recovery.

Tape and Offsite Storage

The tried-and-true tape backup has been around forever. The great thing about tape storage is that it's very reliable. The not-so-great thing about tape storage is that tapes are difficult to inventory, they can be easily lost or stolen, and they are time-consuming to test. Additionally, if your company needs fast restoration of applications and data after a disaster, tape storage isn't going to be able to provide that; tapes must be recalled and restored—all of which takes a significant amount of time.

While offsite tape storage is still a reliable and affordable option for long term offsite archival and data protection, it's not the best option available today for disaster recovery, because businesses need better RPOs and RTOs than what tape can provide.

Speaking of RPO and RTO, what are they and what are the challenges around them? Let's find out.

Meeting RPO / RTO

Recovery point objective (RPO) is the minimum amount of data that is acceptable to lose in the event of a disaster. Recovery time objective (RTO) is the amount of time that is acceptable to recover applications in the event of a disaster.

Traditional DR Challenges

There are numerous items that can complicate your traditional disaster recovery efforts. These include:

- Tape-based recovery is slow and tapes are easily lost
- Recovery point and recovery time objectives are difficult to meet
- High overall costs for DR, putting it out of reach for many
- Keeping the DR runbook up to date with the ever-changing modern data center

For most companies using tape backup, their RPO is 24 hours (because they do a backup each night) and the RTO might be 48 hours, because that's how long it would take them to recall the tapes, recover the data, and bring the applications back up.

While that timeframe and amount of data loss might be fine for a small business it's not going to be acceptable at medium and large enterprises. With thousands of employees working every day, the thought of losing a day's worth of data (24-hour RPO), and the cost to try to re-create that, is unacceptable. With a 48-hour RTO, the company could be down as much as two days before all applications are restored. Again, for most companies, that amount of downtime is going to be unacceptable.

New and innovative DR solutions help you to reduce the RPO and RTO while keeping DR affordable.

High Costs

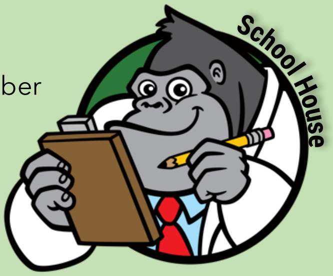
In the past, it's been widely known that "the shorter the RPO and faster the RTO, the greater the cost of the disaster recovery solution" (in terms of hardware, software, and data transmission); however, that is changing with new DR solutions (which we'll talk about later in the book).

With traditional DR solutions, to obtain short RPOs and fast RTOs, you had to use a SAN with synchronous replication and have dedicated wide area network (WAN) circuits between sites. In most DR designs this required you to have your own secondary data center to send the replicated data to and run your secondary servers and storage in the event of the disaster. Many DR replication solutions were designed just for specific application. When you wanted replication for another application, you had to purchase another replication solution for that app. All this semi-custom disaster recovery technology, plus the monthly bandwidth to support the movement of the data, resulted in a very high cost for a high-quality DR solution. Unfortunately, this put disaster recovery out of the financial reach for many companies.

Maintaining the DR Plan / Runbook

Another challenge associated with traditional disaster recovery solutions is that of maintaining the DR plan itself. This plan is the actual runbook, as it's called, as to the steps that the actual administrators would take in the event of a real disaster. That runbook must include a plan for every application, its associated data, and user connectivity, and the sequential recovery steps for the application. With applications changing and moving at a constant rate in the modern data center, the task of maintaining the disaster recovery runbook has become overwhelming for most companies. The result is that their DR runbook is out of date, and, should a real disaster occur, they would be unable to meet the recovery time objective and maybe be unable to recover at all. This is because their runbook doesn't provide the necessary information required to get the applications back up and running.

Disaster Recovery Innovations



Over the years, there have been a number of innovations that have helped to improve organizations' disaster recovery efforts, some of which are described in the following sections:

- Intelligent and efficient data replication
- Real-time runbook updates
- Cloud storage
- Automated testing
- Consumption-based model
- Disaster recovery as a service (DRaaS)

Innovations for the Next Phase of Disaster Recovery

Enough about the challenges of the past, now let's move on and talk about the latest innovations in disaster recovery technology. These innovations make disaster recovery faster, more efficient, more automated, and more affordable.

Intelligent and Efficient Data Replication

Great advances have been made in the area of data replication. Modern data replication solutions are able to replicate only to change blocks of a virtual machine disk in order to make the replication process much more efficient. What that means is that the amount of bandwidth that you'll need to replicate your data from the primary data center to the secondary data center or public cloud is far less than what was needed in the past. With that, it also means you won't have to spend as much on bandwidth between the data centers or between your primary data center and the cloud. The flipside is that you may be able to replicate many more votes machines are much more data than you were able to in the past, while spending the same amount of money on bandwidth.

More advancements around intelligent data replication include the ability to create groups of virtual machines and policies around individual or groups of virtual machines. The benefit of creating groups of virtual machines is to be able to easily replicate and recover all the working machines that make up a single application at one time. You'll also be able to prioritize your applications for replication and recovery such that you can recover the most critical applications first. With the same policies, you can create policies that replicate more critical applications much more frequently, giving them a shorter RPO while giving less critical applications longer RPOs.

Real-Time Runbook Updates

One of the great things about the latest DR solutions is that they have full insight into the virtual machine inventory, and as such, they are aware of virtual machines and application changes (to some degree). With insight into the environment, the DR solution can automatically update the disaster recovery plan, or runbook, in real time as applications or infrastructure are modified.

Cloud Storage

With the popularity and maturity of the public cloud, more enterprises are leveraging cloud storage for storing their archival data and backup data. The big benefit to using online public cloud storage as opposed to offline tape storage is that the data being protected can be constantly indexed and tested for recoverability.

With cloud storage, there's no need to recall tapes, there's no need to recover data, and you can be assured that the applications will function once you initiate the recovery. Additionally, data can be encrypted in the cloud so that you are confident of its security.

Automated Testing

Once data is in the cloud, and your disaster recovery system is aware of the recovery process, you can configure automated testing such that application groups are actually recovered in a sandbox environment and test transactions are run through the protected virtual machines. This means you can be confident that recovery would be 100% successful. This type of automated testing can be scheduled once a quarter, once a month, once a week or even every night.

Consumption-Based Model

The latest disaster recovery solutions also offer a consumption-based pricing model such that there's no large capital expenditure upfront to get started with DR. With the consumption-based pricing model, you simply pay for your DR system based on some level of consumption—this could be the number of virtual machines that you replicate or the amount of data that that's being stored in the cloud.

Should you have a disaster, you might pay a relatively small charge to immediately bring up your applications running in the public cloud so that they're accessible to your end users to allow you to continue business functions. Some disaster recovery providers don't even have any cost associated with bringing up your production infrastructure in their public cloud.

Compare that to having to purchase your own secondary data center with fully redundant server storage and network equipment, network connectivity, and having to maintain that redundant data center indefinitely.

Improved RPO/RTO

The new replication techniques that we talked about earlier allow for software-based replication between dissimilar hardware and still with approximately one-minute RPO, which allows you to meet some of the lowest RTOs that the business can ask for.

Not only are the RPO and RTO low, but, because automated testing has been done to the application level, you can be confident that critical applications will be recovered within a short RTO and that little to no data loss will occur. This is possible because the latest techniques allow you to have more frequent replications that are application-aware.

Disaster Recovery as a Service (DRaaS)

Not only can you use public cloud services to store your data offsite and bring up applications for "DIY DR" but there are many disaster recovery as a service, or DRaaS, providers who are wrapping up these disaster recovery innovations into a single service offering.

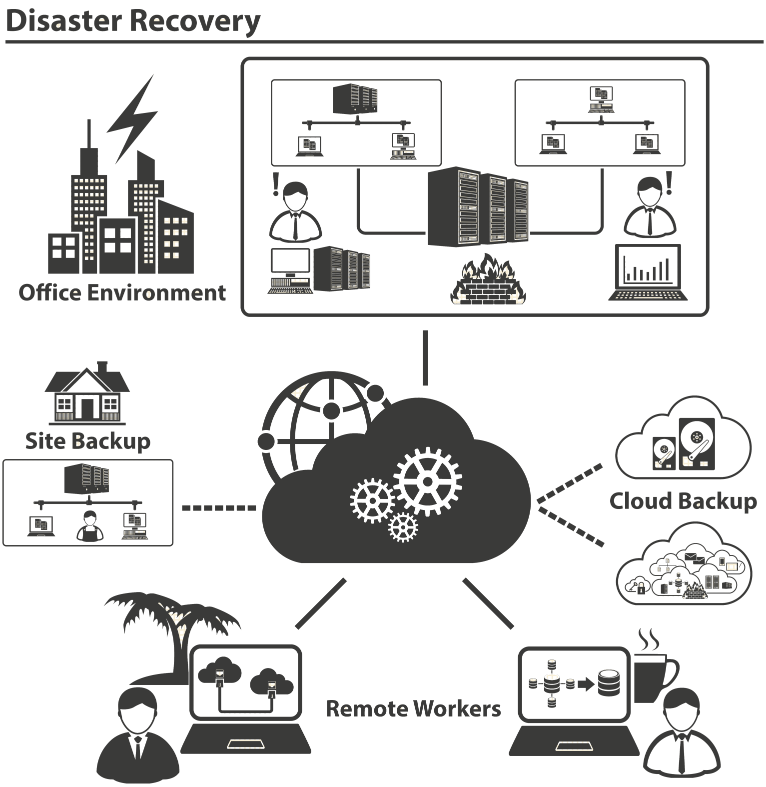
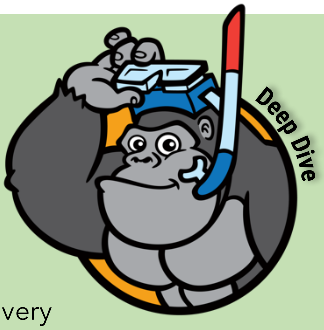


Figure 2: Disaster Recovery as a Service (DRaaS) in Action

With DRaaS you can replicate data from your on-premises infrastructure to the DRaaS providers' public cloud and use a web-based interface to manage your DR infrastructure. You could also use that interface to schedule automated testing and actually initiate recovery.

DRaaS providers typically charge using the same consumption-based model that we talked about earlier, meaning you don't need a large capital expenditure to get started. You simply pay for what you use.

Why Disaster Recovery as a Service (DRaaS)?



The public cloud has virtually infinite storage and compute scalability. That storage capacity is available in different tiers with the lowest tier being ideal for protecting a company's data, offsite, at a very low cost in a utility-based pricing model. Since compute capacity, like storage, is only charged when consumed, a company would only pay for the compute they use when they perform disaster recovery testing or have an actual disaster.

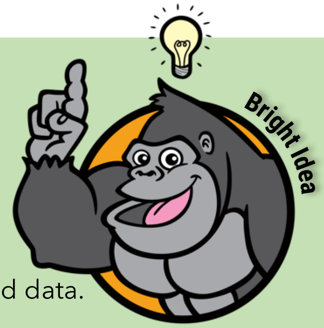
However, DRaaS is more than just storing data in the cloud and spinning up VMs. DRaaS solutions offer simple web-based interfaces for managing the entire DR process such that you may never even have to use public cloud management interfaces such as those from AWS or Azure, directly.

DR has always been a struggle to afford and manage due to the cost and complexity of building and managing a secondary DR site. Now, companies of all sizes are realizing that DRaaS is easy, affordable, and, in most cases, their most logical use case for leveraging the benefits of the public cloud.

Implementing Next Generation Disaster Recovery

At this point, you might be thinking that the next generation recovery technologies sound enticing, but how would a company go about implementing something so new? The great thing about cloud DR is that it can be easily tested and evaluated.

Let's walk through some of the most common frequently asked questions related to next generation DR solutions.



Your Next-Gen Disaster Recovery Implementation Checklist

- ✓ Identify applications and data that needs protection.
- ✓ Define RPO/RTO for applications and data.
- ✓ Evaluate DRaaS solutions.
- ✓ Replicate data to a DRaaS cloud.
- ✓ Test the DRaaS solution.

- **How difficult is it to migrate from existing DR solutions to next-generation solutions?**

  Because most of the new DR solutions are software- and cloud-based, it's very easy to evaluate and migrate to these types of solutions. For companies who are using backup tape only for DR, they would continue to use their backup tapes and offsite storage for worst-case data protection as they ramp up the replication of first their most critical applications (and later all applications) offsite using software-based replication.

  For companies who have existing and costly hardware-based or other dedicated replication solutions, these solutions could be phased out as the new software-based replication solutions are brought online and confidence is gained with them.

- **Would there be any capital expenditure needed to move to next-generation cloud-based DR solutions?**

  With cloud-based solutions that use the pay-as-you-go model, no there's no upfront capital expenditure. Over time, your monthly cost would increase proportionally to the amount of data that you replicate to the DR cloud provider's storage.

- **What are the steps that you would take?**

  The steps to take to get started with software-based and cloud-based disaster recovery begin with simple product is valuations, typically through working with a value-added reseller (VAR) or Reseller. This process typically involves obtaining a list of your applications and the size of your data set. From there, the consultant would help you to define the RTO and RPO for those applications and then make recommendations on the cloud-based disaster recovery solution that would best fit your needs.

Chapter Summary

Whether you're at a small or large company it's likely that you have struggled at some point with disaster recovery. Even for companies that have the budget for traditional high-performance DR solutions, almost everyone still struggles to keep the disaster recovery plan up to date and tested with the ever-changing modern data center. For these reasons, companies small and large welcome the latest innovations that we've discussed in this chapter on disaster recovery.

Chapter 5

Turbocharging End User Productivity Through VDI

"Is **this** the 'Year of VDI?'"

That question typically elicits a chuckle from most in enterprise technology, because analysts, pundits, and even some VDI software companies have been declaring the arrival of the year of virtual desktop infrastructure (VDI) since 2010. Thus far, the year when almost every enterprise adopts VDI hasn't arrived—and realistically, it won't.

However, enterprises have been steadily transitioning to VDI year after year, month after month. At some point, the number of end-users leveraging VDI desktops will overtake traditional desktops. To most IT pros, this move from physical desktops to virtual desktops is a logical transition considering the numerous benefits that a VDI model provides over traditional desktops.

So how does VDI work; how does it help; and what can be learned after many years of VDI implementations?

In this chapter, you'll learn exactly what VDI is, learn about the potential benefits organizations and their end-users can gain from adopting a VDI strategy, and get a bit of insight on some of the primary challenges VDI adopters face.

High-Level Business Benefits

When server virtualization took the data center by storm, the hardware consolidation provided a very nice reduction in infrastructure spending for many organizations. It's a reasonable conclusion to expect that desktop virtualization might provide the same type of economic benefits. So, for a long time, VDI vendors led their sales campaigns with the notion that VDI would save boatloads of money. While that hasn't always turned out to be the case, there are numerous potential cost savings along the way. Today, IT organizations consider the following reasons for transitioning their workforce to a virtual desktop infrastructure.

Agility

Modern IT organizations face the daunting task of keeping up with technology as it changes and matures at an exponential rate. Technological advances from manufacturers and vendors that would have taken three years in development a decade ago can be accomplished in a matter of months today. As such, IT leaders are on the lookout for strategies to make the deployment and management of their IT services more agile.

Rather than touting the cost savings as they once did, today's VDI advocates often lead with explaining how a VDI deployment can enable IT organizations to react more quickly, provision resources rapidly, and accelerate the deployment of new applications.

Security

As the era of "knowledge work" has matured, employees want more flexibility in their work setting. They want to work from home, from the train, or from anywhere else they find themselves. While technology certainly affords us much more flexibility in this area than existed a decade ago, new risks are inherent as well. One of the primary risks is that the ability to work from anywhere also means that IT must allow employees to access and manipulate data from anywhere.

For the past ten to fifteen years, this posed a significant security threat. Furnishing laptops for 500 employees created 500 opportunities for sensitive data to be lost, stolen, or otherwise compromised. Additionally, now that smartphones and tablets are ubiquitous, the bring-your-own-device (BYOD) movement has grown and, therefore, the risk continues to grow.

The possibility of sensitive data physically leaving the corporate premises becomes a non-issue, however, with a slight paradigm shift. VDI gives organizations the opportunity to use a server-based computing approach reminiscent of Terminal Services where the endpoint device is purely for visual and control purposes; all the computing power and data reside within the walls of the data center. Although there are still plenty of security risks to mitigate with VDI, data is much safer if it never leaves the data center.

Standardization of Process and Technology

Much has been said in recent years of the concept of "technical debt." Technical debt is a way of describing the impact of taking engineering or architectural shortcuts which pay off in the short term but come back to bite you in the long term. Although this metaphor for discussing the impact of deficiencies has been used for 20 years or more, the problem doesn't seem to have been solved yet.

An experience that every organization has had to one degree or another, and which contributes to the creation of technical debt, is the "one off," also known as the "snowflake." These terms denote uniqueness, and while uniqueness in itself isn't bad, every deviation from a standard (which may or may not even exist) creates an opportunity to introduce technical debt.

What does technical debt have to do with VDI? By its very nature, a sound VDI strategy encourages the reduction of uniqueness and an increasing amount of uniformity in the desktop computing environment. As a result, technical debt is reduced over time, and the environment becomes easier to manage.

Before we look in depth at both the challenges that tackling a VDI deployment can pose and the rewards that a successful implementation can bestow on end-users, let's clearly define what VDI is and what it is not.

What Is Virtual Desktop Infrastructure?

Definitions of VDI vary, but here's the most common: virtual desktop infrastructure is the presentation of desktop operating systems to end-users who leverage a variety of endpoint devices, and rather than dedicated computers, the desktop instances are virtual machines running on servers in a central data center.

VDI Versus Terminal Services

The idea of connecting endpoint devices to a central computing resource is literally as old as the idea of a data center. The Client/Server architecture has served many organizations well in various flavors over the years. Anyone with experience in remote access to corporate resources from the previous decade might say that VDI sounds awfully similar to the Terminal Services architecture that Microsoft introduced with Windows NT 4.0 and Windows 2000.

That observation would be correct in the sense that a user connects from an endpoint device (maybe managed by IT, maybe not) to central computing resources. But that's where the similarity ends.

VDI is different from Terminal Services (now known as Remote Desktop Services) in a few very important ways:

- **Focus on end-user experience.** Although it's possible to connect many different types of devices to an RDS host, the reality is that the experience isn't always a good one. Endpoint client applications designed for VDI remoting take special care to address usability for different types of devices. As an example, the method for moving the cursor can be tailored to the specific device's operating system. Therefore, using a name brand VDI client – such as one from VMware or Citrix - will often provide a better user experience than an RDS session from the same device.

- **Desktop versus server operating system.** Another challenge with RDS is application compatibility. Many IT professionals are now bald from the hair-ripping frustration of attempting to make desktop applications run on RDS Hosts (server operating systems). Many application developers don't develop with RDS in mind; as a result, the application behaves poorly when installed on a server. User permissions are another common headache with desktop applications on RDS; VDI doesn't suffer from the same limitation.

- **Personalization and control.** When multiple users connect to an RDS Host, there are concerns with security. For example, Bill shouldn't be able to access Debbie's private files, yet since they're sharing a machine, it is possible in the absence of proper security precautions. Likewise, there are personalization concerns; Debbie might not want her desktop background to be a picture of Bill's cat. With a VDI deployment, each user gets their very own desktop instance, meaning that user segregation is straightforward.

- **Offline use**. In an RDS environment, users connect their client to a server. That's the only way the resources can be used. Because the server is a shared resource, this model can never change. In a VDI environment, since each user has their own desktop, some organizations and software allow users to "check out" their desktop for offline use. Taking a desktop offline means that, for example, a traveling executive can download their corporate desktop, run it on a laptop with a Type 2 hypervisor like VMware Fusion, and use their corporate desktop on a plane that doesn't have Internet access. When they return to the office, they check the desktop back in to the data center. Checking in and checking out all happens over the corporate network, of course, so physically visiting the data center isn't necessary; this process can be done from their desk.

Desktop Virtualization Is a Horse of a Different Color

"Fair enough," you say. "VDI is not like RDS. But I already understand desktop virtualization because it's the same thing as server virtualization, which my organization has been doing for years!"

As much as that's a very reasonable conclusion, it turns out to be very far from the truth. Although virtual hardware is virtual hardware and knowing how to manage the underlying Type 1 hypervisor translates well, the challenge comes in successfully virtualizing a desktop operating system versus a server operating system. Secondarily, accounting for "user experience" in unnecessary in a server virtualization world. IT Operations staff have to focus on application response times and so on, but whether the user interface on a server is responsive is mostly unimportant.

Network Requirements

As it turns out, end-users care very much about whether the interface of their desktop is responsive. They will use words like "snappy" or "sluggish" to describe their experience, so ensuring that the user experience is stellar is crucial to achieving a successful VDI rollout. Most users will actively protest anything less.

You may have noticed that storing text files on a computer requires a small amount of space. A video file, however, requires a comparatively large amount of space. Consider that every end-user connecting to a virtual desktop is essentially streaming video (their "screen") from the data center to their client device. Whereas server applications use a small amount of network bandwidth transferring data that more closely resembles a text file, desktop users consume a significant amount of bandwidth just by transferring screen refreshes — 30 to 120 of them per second, per user, in fact.

Suffice it to say that proper network architecture is critical for a successful VDI implementation. There are two aspects that help ensure success: first, designing a network that provides ample performance for all the VDI servers and the clients; second, properly tuning the protocols (like PCoIP or ICA) that control how the clients connect to the desktops.

I/O Workload

Storage performance is a consideration in server virtualization as well as in VDI. However, what's tricky about VDI is that the I/O pattern is significantly different than server operating systems and the type of applications that typically run on servers. Those organizations with experience in server virtualization sometimes fail to account for this. It's extremely likely that if you back a VDI implementation with a storage design meant for servers, the VDI deployment will crash and burn as soon as a significant number of users are on the platform.

There are two major ways that desktop storage I/O differs from server I/O:

- **It's highly random**. Server virtualization causes storage commands to be delivered to the storage array in a random order because of an "I/O blender effect" caused when many servers communicate with a single storage target at the same time. However, server operating systems, and some of the applications that run on them, tend to issue mostly sequential I/O commands. Desktops and desktop applications, on the other hand, tend to read and write with a higher level of randomization. Therefore, not accounting for the high degree of random I/O in a VDI storage architecture can result in poor performance.
- **It's heavy on the writes**. A traditionally accepted rough estimate for I/O calculations is that servers issue read commands approximately 80% of the time, and they issue write commands approximately 20% of the time. Desktops, on the other hand, are completely opposite. The rule of thumb for desktop operating systems is 20% reads and 80% writes. This difference dramatically impacts the storage architecture required to deliver proper performance.

Licensing

Lastly, licensing in a VDI environment can be tricky. It's tempting to think that if your desktop licensing is accurate on physical desktops, then applying the licensing to the same number of virtual desktops won't make a difference. In fact, the license agreements don't transfer quite that cleanly. Hordes of IT professionals have been surprised to find that Microsoft has special rules for virtual desktops. Although Microsoft has made some improvements to their licensing model and the licensing situation is better now, many architects are still uncertain whether their VDI licensing is truly in order.

VDI Is Superior to Physical Desktops for End-Users

Despite the challenges organizations face when deploying VDI, the opportunity for increasing end-user productivity and satisfaction is tremendous. It's no secret that many IT organizations are seen as a barrier to productivity and success rather than enablers; a successful VDI implementation can go a long way toward flipping that sentiment to something more positive.

If by this point you think that VDI sounds like a promising way to increase productivity and decrease rigidity for your end-users, you're not alone. In a recent survey conducted by ActualTech Media, 55% of respondents who are planning a VDI deployment said that they're doing it because, "We want to provide our users with improved access to their applications and workloads." Additionally, 44% of those same respondents indicated that, "We want to use VDI to solve our desktop/laptop lifecycle problem."

Consider the following ways that a VDI deployment might change the game for your workforce.

Incredible Mobility

Consider this scenario: A user's desktop computer at the office is set up in just the right way to help them get their job done, but they would like to work from home for the day. Unfortunately, they're out of luck because they can't take their desktop home. With an externally accessible VDI deployment, a user can access a virtual desktop from anywhere on almost any device and have access to all their applications, just like they're used to.

This increased mobility changes the game entirely, because suddenly personal laptops, smartphones and tablets, or even a kiosk at the library, can all become an endpoint for a user to access the exact same desktop and applications they need. VDI unchains a user from their cubicle and allows them to work from Thailand if they need to, because they'll have almost exactly the same experience as they would at their desk at the office.

Desktops-as-a-Service

Traditionally, implementing virtual desktop infrastructure (VDI) in an enterprise resulted in large cost and complexity challenges. IT organizations typically purchased brand new storage and compute VDI infrastructure, VDI client licenses, and VDI broker software. There is a lot to buy.

In addition to these cost challenges, there are also significant complexity challenges to overcome, such as configuring the VDI software, creating the shared operating system images, and centralizing applications in hopes of giving the end users access to their ideal desktop, which now runs in a virtual machine in the datacenter. Even after all that, the unfortunate reality is that many IT organizations—and their users—experience performance problems with their brand-new environments.

With Desktop-as-a-Service (DaaS) solutions, the VDI environment is run off-premises, in the public cloud, and with subscription-based pricing. IT organizations simply upload their desktop image (with embedded applications), connect the already installed VDI infrastructure to their active directory, and grant end users access to their new virtual desktops, running in the cloud. Besides eliminating upfront capital expenditures, DaaS offerings are already built to scale and perform, which can eliminate many of the headaches associated with VDI and ensure the best end user experience possible.

Any Old Endpoint Will Do

In the DevOps classic *The Phoenix Project*, the protagonist Bill showcases the agony of working for a few weeks with a years-old loaner laptop that is so slow it cripples his productivity. The reality is that in a standard enterprise deployment of desktop and laptop computers, if your company-issued device breaks down, you can wait weeks for a replacement or repair. And all the while, the ability to do your job is impaired. What's even worse is that if the device is unrecoverable and critical data was stored on the device, you may lose completed work or records of important communications.

In a post-VDI world, this issue doesn't exist. The endpoint can be a device as simple as a Chromebook or a tablet. If your old device suffers damage or is lost, either a new one can be issued, or you can swing over to the local computer store and pick up a new one. Within a few hours, you can reconnect to the same virtual desktop you've been using and you're on the road to productivity once again.

Performance Boost

The shared resources in a VDI environment enable a unique economic model that can lead to better performance for everyone. For example, users might have access to an all-flash storage array in the data center, where the budget would not allow all users' laptops and desktops to be upgraded to SSDs. As a result of all I/O being served by the all-flash array, the storage performance of each user's system is improved and the overall experience is better.

Another potential performance benefit is best explained as "data locality." In an enterprise laptop and desktop computing environment, users frequently access files on network shares. Depending on geography, network performance, and the data being accessed, the user experience of accessing those network shares can vary. Unfortunately, that means that sometimes the experience isn't very good.

On the other hand, virtual desktops are often physically mere feet away from the servers hosting the file shares. Consider the previous example of the employee working from Thailand. With a standard laptop deployment, that user would likely connect a VPN client to the corporate network and transfer the file in question literally across the world before accessing it. If the application in question is CAD software like a tool from Autodesk, it may be nearly impossible for that user to get work done. In a VDI deployment, however, the visual representation of the desktop is transmitted to the user and all file access happens locally to the data center. Avoiding large file transfers across either the corporate network or the Internet provides a markedly superior experience.

Bonus: Enhanced Reliability

Finally, because a VDI deployment causes standardization within the IT organization and some uniformity across desktops, troubleshooting and issue resolution for helpdesk staff becomes easier. As a result, problems are resolved more quickly and users are back to work sooner.

Chapter Summary

When you combine the potential impact of these four benefits, it becomes clear that VDI has significant potential to positively impact the productivity of end-users. To learn more about VDI and end-user computing, it's worth visiting the websites of VMware and Citrix – two of the market leaders in VDI. Both offer a variety of resources to guide continued learning about this topic.

About

About Anexinet

Anexinet (www.anexinet.com) specializes in digital business transformation. We empower our clients to grow their customer base and improve workforce efficiency by envisioning, developing, and operating next generation technology solutions. Our core expertise is in digital applications, analytics, and hybrid IT, enabling businesses to rapidly transform. Partner with Anexinet to support the full lifecycle of your next generation digital business.

About the Authors

Scott D. Lowe, vExpert

Scott Lowe is a vExpert and a partner in ActualTech Media. Scott has been in the IT field for over twenty years and spent ten of those years filling the CIO role for various organizations. Scott has written thousands of articles and blog postings and regularly contributes to www.ActualTech.io.

David M. Davis, vExpert, CCIE, VCP, VCAP-DCA

David Davis is a partner at ActualTech Media. With over 20 years in enterprise technology, he has served as an IT Manager and has authored hundreds of papers, ebooks, and video training courses. He's a 6 x vExpert, VCP, VCAP, & CCIE# 9369.

James Green, vExpert, VCAP-DCD/DCA

James is a Partner at ActualTech Media and writes, speaks, and consults on Enterprise IT. He has worked in the IT industry as an administrator, architect, and consultant, and has also published numerous articles, whitepapers, and books. James is a 2014 - 2016 vExpert and VCAP-DCD/DCA. Follow James on Twitter at @jdgreen.