LEARNING MADE EASY

2nd HPE SimpliVity Special Edition

Hyperconverged Infrastructure

for dummies
A Wiley Brand

Simplify IT and reduce TCO

Solve protection and performance challenges

Achieve scalability and agility

Brought to you by

Hewlett Packard Enterprise | intel

Scott D. Lowe

About Hewlett Packard Enterprise

Hewlett Packard Enterprise is an industry-leading technology company that enables customers to go further, faster. With the industry's most comprehensive portfolio, spanning the cloud to the data center to workplace applications, our technology and services help customers around the world make IT more efficient, more productive, and more secure.

Hewlett Packard
Enterprise

Hyperconverged Infrastructure

2nd HPE SimpliVity Special Edition

by Scott D. Lowe

for dummies®
A Wiley Brand

Hyperconverged Infrastructure For Dummies®, 2nd HPE SimpliVity Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-56249-8 (pbk); ISBN 978-1-119-55853-8 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Editor: Elizabeth Kuball

Acquisitions Editor: Katie Mohr

Editorial Manager: Rev Mengle

Business Development Representative: Karen Hattan

Production Editor: Siddique Shaik

Table of Contents

Introduction

Over the past decade, a lot has happened to IT and to the technology industry. First and foremost, business leaders continue to demand more from their IT functions, including better and faster service, with seemingly endless expectations for increased efficiency.

In addition, several industry trends have had significant effects on enterprise IT:

- The proliferation of purpose-built devices
- The rise of software-defined infrastructure (SDI)
- The emergence of cloud computing — or at least cloud principles and economics — as a key driver of enterprise IT
- The emergence of converged infrastructure
- The use of flash storage in certain enterprise applications
- The increased prevalence of software-defined networking

At the center of all these trends lies hyperconverged infrastructure, which is the next generation of a series of converged opportunities that have hit the market. Hyperconverged infrastructure is a data center architecture that embraces cloud principles and economics and a software-first architecture. Hyperconverged infrastructure consolidates server compute, storage, hypervisor, data protection, data efficiency, global management, software-defined networking, and other enterprise functionality on commodity x86 building blocks to simplify IT, increase efficiency, enable seamless scalability, improve agility, and reduce costs.

About This Book

The hyperconvergence story has many chapters. In this book, I discuss the trends that are leading modern IT to hyperconverged infrastructure. I also discuss both the technical and business benefits that come from implementing a data center based on hyperconverged infrastructure.

Foolish Assumptions

In this book, I assume that you know a little something about virtualization and cloud computing trends and models. As such, this book is written primarily for IT executives and managers such as CIOs, CTOs, IT directors, and technical managers.

Icons Used in This Book

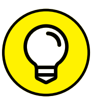
Throughout the book, you'll find several icons in the margins. Here's a rundown of what these icons mean.

REMEMBER

Anything that has a Remember icon is well worth committing to memory.

TECHNICAL STUFF

The Technical Stuff icon indicates extra-credit reading. You can skip it if you like (but I hope you won't).



TIP

The Tip icon points out helpful information.

WARNING

The Warning icon alerts you to risks of various kinds.

Beyond the Book

There's only so much I can cover in such little space. If you're eager to learn more about hyperconverged infrastructure after reading this book, visit `www.hpe.com/info/hc`.

Where to Go from Here

Like all *For Dummies* books, this book is designed to be read in any order you choose. Start with the chapter that interests you most, or read straight through. Hey, it's your book, so it's totally up to you!

IN THIS CHAPTER

- » **Understanding what hyperconverged infrastructure is**
- » **Discussing the many forms of hyperconvergence**

Chapter 1
Hyperconvergence Basics

Corporate technology undergoes a massive shift every so often as new models emerge to meet changing business needs. This chapter is about hyperconverged infrastructure, which is the culmination and conglomeration of several trends that provide specific value to the modern enterprise.

Defining Hyperconverged Infrastructure

So, what is hyperconvergence? At the highest level, it's a way to enable cloudlike economics and scale without compromising the performance, resiliency, and availability expected in your own data center. Hyperconverged infrastructure provides significant benefits:

- » **Data efficiency:** Hyperconverged infrastructure reduces storage, bandwidth, and input/output operations per second (IOPS) requirements.
- » **Elasticity:** Hyperconverged infrastructure makes it easy to scale out/in resources as required by business demands.

- **Workload-centricity:** With hyperconverged infrastructure, the focus is on the workload as the cornerstone of enterprise IT, with all supporting constructs focused on applications.
- **Data protection:** Ensuring data restoration in the event of loss or corruption is a key IT requirement, and it's made far easier by hyperconverged infrastructure.
- **VM mobility:** Hyperconverged infrastructure enables greater application/workload mobility.
- **Resiliency:** Hyperconverged infrastructure enables higher levels of data availability than are possible in legacy systems.
- **Cost efficiency:** Hyperconverged infrastructure brings to IT a sustainable step-based economic model that eliminates waste.

Hyperconvergence Constructs

Convergence comes in many forms. At its most basic, convergence simply brings together existing individual storage, compute, and network switching products into pretested, pre-validated solutions sold as a single solution. However, this level of convergence only simplifies the purchase and upgrade cycle. It fails to address ongoing operational challenges often introduced with the advent of virtualization. There are still logical unit numbers (LUNs) to create, wide area network (WAN) optimizers to acquire and configure, and third-party backup and replication products to purchase and maintain.

Hyperconverged infrastructure seamlessly combines compute, storage, networking, and data services in a single physical system. The software that enables hyperconvergence runs on industry-standard x86 systems, with the intention of running virtualized or containerized workloads. Distributed architecture lets you cluster multiple systems within and between sites, forming a shared resource pool that enables high availability, workload mobility, and efficient scaling of performance and capacity. Typically managed through a single interface, hyperconverged infrastructures let you define policy and execute activities at the virtual machine (VM) or container level.

The results are significant and include lower capital expenses (CAPEX) as a result of lower upfront infrastructure costs, lower operating expenses (OPEX) through reductions in operational costs and the need for specially trained personnel, and faster time to value for new business needs. On the technical side, IT generalists — IT staff with broad knowledge of infrastructure and business needs — can easily support hyperconverged systems. No longer do organizations need to maintain islands of resource engineers to manage each aspect of the data center.

To fully understand hyperconvergence, it's important to understand the trends that have led the industry to this point. These include post-virtualization headaches, the rise of the software-defined data center (SDDC), and cloud computing.

IN THIS CHAPTER

» Keeping up with innovation

» Using resources efficiently

» Looking at management interfaces

» Discussing touchpoints

» Setting effective policies

Chapter 2 Virtualization Challenges

Fact: Virtualization fundamentally and permanently changed IT and the data center. Today, most services are running inside virtual environments, and IT often takes a "virtualized first" approach to new application and service deployment. That is, administrators consider the virtual environment for running new applications rather than just building a new physical environment.

Although virtualization offers significant benefits, it also introduces challenges that IT must overcome to help propel the business forward. This chapter describes those challenges.

Infrastructure Innovation

Every time a start-up company releases a great new product, enterprises scramble to implement that solution. The proliferation of purpose-built devices has created unnecessary complexity — and the result has been data center chaos.

Innovation is great, and we all want it to continue, but eventually, data centers have so much stuff that they become unmanageable. It's time to clean out the closet, so to speak.

Over the past decade, IT departments have focused on solving the storage capacity problem, deploying all kinds of technologies to tame the capacity beast, such as wide area network (WAN) optimization and backup deduplication appliances. As a result, data efficiency technologies have become standard features of many different products.

But what happens when you put these products together in the data center? You end up constantly deduplicating and hydrating data as it moves among various devices. Storage deduplicates data; then you read the data to back it up, where it requires hydration (to return it to a state that the backup application understands) and often re-deduplicating it somewhere in the backup data path. The central processing unit (CPU) cost to reprocess the same data is enormous, not to mention the bandwidth cost of all that hydrated data.

REMEMBER

Deduplication is the process in which data is examined for common blocks. When identified, these common blocks are replaced with a tiny little pointer to the unique copy of the data already stored on disk — which takes up significantly less capacity when written to storage. Deduplication has a tremendous impact on storage capacity savings and, importantly, on input/output operations per second (IOPS) because fewer writes occur to disk. *Hydration* is reversing the deduplication process, such as when moving the data to a new system that doesn't support deduplicated data.

I discuss this challenge in more depth later in this chapter.

Underused Resources

Virtualization helped organizations consolidate many of their servers to run on a common platform: the hypervisor software layer. This move has helped IT departments make much better use of their server resources. Before virtualization, it was common for server utilization to average just 15 percent. Virtualization has pushed that average much higher. As a result, organizations now enjoy a much better return on their server investments. Plus, they usually don't need to buy as many physical servers as they did in the past.

Virtualization has changed the game when it comes to server resources. Unfortunately, IT departments often need to maintain separate groups of people to manage separate hardware resources. For example, one group manages storage, another group manages the server side, and a third group handles networking. When an issue arises, it's not uncommon to see a lot of finger-pointing.

Plus, emerging workloads are creating resource challenges that push IT departments to develop infrastructure environments on a per-service basis. Virtual desktop infrastructure (VDI) environments, for example, have vastly different resource usage patterns from server virtualization projects. To meet user expectations with VDI, IT professionals often implement completely separate environments, from servers on down to storage.

Aren't resource islands the very problems that virtualization is meant to solve? These islands are among the biggest culprits of underutilization. Virtualization is supposed to result in a single resource pool from which resources are carved out to meet application needs, thereby maximizing the use of those resources.

Multiple Management Interfaces

Storage devices. Optimizers. Hypervisors. Load balancers. What do they have in common? Each of these disparate components features its own management interface. If you use multiple components, each with separate management consoles (and policy engines) rather than a single, centralized, easy-to-use administrative system, you may experience the following challenges:

- Vendors blaming each other when something goes wrong
- The inability to scale your data center environment easily and linearly
- Greater complexity due to policies and management being tied to IT components versus workloads
- Endless refresh cycle — as soon as you finish updating or upgrading one component, it's time to update the next

Deployment Difficulty and Delays

Resource challenges represent the number-one reason why organizations continue to have problems deploying new applications and services. A close second is administrative overhead. Allow me to explain.

Multiple challenges exist on the resource front, including the following:

- **IO blender:** The consolidation of virtual machines (VMs) contributes to a random input/output (IO) workload — each with its own pattern for reading/writing data to storage. (I discuss the IO blender in detail later in this chapter.)
- **Capacity:** Another challenge is ensuring adequate capacity as the organization grows. With resources divvied up and islands of resources strewn about the data center, managing ongoing capacity so that there's enough to go around becomes increasingly difficult.
- **Overhead:** Even if you have enough resources to deploy a new application (see the preceding bullet), the administrative overhead involved in the process presents its own challenges:
  - A new logical unit number (LUN) must be provisioned to support the new application. If tiers of storage are involved, this process could require multiple steps.
  - One or more new VMs must be provisioned.
  - Networking for those new VMs has to be configured.
  - Load balancers and WAN optimization devices need to be managed to support the new VMs.
  - Data protection mechanisms must be implemented for the new services.

Whew! That's a lot to do. All of it is time-consuming, and all of it involves different teams of people in IT. Good luck!

CONVERTING TO FLAT IT

The legacy data center is very delicate in many ways. Any change at any level has the potential to disrupt the overall structure. With lessons and tactics learned from the big cloud vendors, hyperconvergence replaces tiered and resource-siloed data centers with a much flatter IT structure. As practically all the formerly separated data center hardware gets folded into the hyperconverged environment, the IT department can shift its focus, changing its resourcing structures and skill sets. Instead of having staff with deep subject matter knowledge in each resource area, hyperconvergence has given rise to infrastructure generalists.

Infrastructure generalists don't have deep knowledge of individual resources, but they have broad knowledge of all resource areas. They don't *need* to have deep knowledge. In a hyperconverged world, the most complex stuff is handled under the hood. In cases where hyperconvergence sits on an intelligent networking fabric, the IT generalist can manage that, too, without specialized training. Increasingly valuable resources in the IT world, infrastructure generalists need to have enough broad knowledge to meet business needs and to manage the entire environment through a single administrative interface. In many ways, these people are far more application focused than their island-based predecessors were. They just need to know how to apply infrastructure resources to meet individual application needs.

This development offers several bits of really good news for IT departments that have struggled to align IT operations with business needs:

- This new structure paves the way to eliminating the inefficient resource management islands that have emerged in IT.
- A flat data center managed by an infrastructure engineering group provides improved economies of scale compared with old resource islands.
- Infrastructure generalists are far closer to applications than specialists of old were.

FAST PROCESSORS AND ALL-FLASH ARRAYS

Knowing the underlying hardware capabilities of a hybrid infrastructure is key, and recent processor technology has gone a long way toward enabling all kinds of advances in modern infrastructure products. Intel is leading the way on this front. Hyperconvergence solutions built on Intel architecture make it possible to significantly increase per-server workload density, which can reduce infrastructure complexity and provide a lower total cost of ownership (TCO).

The Intel Xeon Scalable processor family delivers impressive workload density and incredible scale in hyperconverged infrastructure through the availability of processors featuring up to 28 cores per socket with support for up to eight sockets per server. On the memory front, Intel's latest processors feature up to 6TB of system memory. These systems represent the biggest platform advancement from Intel this decade, and the results are clearly demonstrated in the outcomes that are achieved by customers.

No longer is a server replacement program handled on a one-to-one basis like it used to be. Thanks to the density and scaling improvements offered with Intel's latest processor family, IT can replace four five-year-old servers with just one Intel Xeon Scalable processor–based server, lowering five-year TCO by up to 60 percent.

Whether choosing public cloud infrastructure or modernizing your own data center for private cloud deployments, choosing clouds that run on the Intel Xeon Scalable family delivers trusted performance, resilience, and scalability. Intel Xeon Scalable–based platforms unlock the power of scalable performance for next-generation hybrid clouds with intelligent telemetry, world-class platform resiliency, and uptime.

But processor advances aren't the only factor helping IT run more and bigger workloads in hyperconverged infrastructure environments. Lots of storage media vendors on the market today have contributed to the renaissance that has been enjoyed by storage admins over the past few years. Modern data workloads are benefitting significantly from innovations in flash storage media that have redefined the cost, capacity, and performance of storage. As storage prices have come down, new, fast, flash-based options have suddenly become affordable for organizations of almost any size. Although some of these products were originally designed to solve single-application problems — think VDI and Big Data analytics — it's become clear in recent years that many other enterprise workloads can benefit from all-flash storage in hyperconverged solutions. Let's face it: Any way you can speed up your system is welcome in a world where customers don't just *want* instant results, but *demand* them. And for applications that need to achieve hundreds of thousands or even millions of IOPS in a tiny amount of rack space, all-flash storage can't be beat.

All-flash storage is even better when it's safely ensconced inside a hyperconverged infrastructure wrapper. A stand-alone all-flash array is super-fast, but it can sometimes exacerbate the issue of infrastructure "silos." When coupled with hyperconvergence, all-flash goodness can be extended to all your workloads while also being instantly imbued with unprecedented scaling ability.

Storage

Virtualization is heavily dependent on storage, but this resource has wreaked havoc in companies that are working hard to achieve 100 percent virtualized status. Here's why.

Consider your old-school, physical-server-based workloads. When you built those application environments, you carefully tailored each server to meet the unique requirements for each individual application. Database servers were awarded two sets of disks — one for database files and one for log files — with different redundant array of independent disks (RAID) structures. File servers got RAID 5 to maximize capacity, while still providing data protection.

Now consider your virtualized environment. You've taken all these carefully constructed application environments and chucked them into a single shared-storage environment. Each application still has specific storage needs, but you've basically asked the storage to sort out everything for you, and it hasn't always done a good job.

IO blender

In the old days, storage systems were optimized around LUN management. LUNs were replicated from a controller in one storage array to a LUN attached to a controller in a second array. The storage systems took snapshots of LUNs, and LUNs could be moved from one host to another host.

Today, servers have been replaced by VMs, and many VMs are running on a single host, and many hosts are using a single LUN to store VMs. This means that the storage system has dozens or hundreds of logical servers (VMs) all stored in the same LUN. A single application, host, or VM can no longer be managed from the storage system perspective.

A VM-centric platform cuts through this *IO blender* — a term that's been coined to describe environments in which mixed IO patterns are vying for limited storage resources — and allows you to optimize individual VMs. Policies can be applied to individual VMs. Performance can be optimized for individual VMs. Backups can be managed per VM, and replication is configured per VM.

Do you see a pattern emerging here?

When all your applications attempt to work together on the same LUN, the IO blender is created. Here are some ways that common services contribute to the IO blender:

- **Databases:** Databases feature random IO patterns. The system has to jump all over the disk to find what you're looking for.
- **Database log files:** Log files are sequential in nature. Usually, you just write to log files — again, sequentially.
- **Random file storage:** File servers are very random when it comes to IO. You never know when a user will be saving a new file or opening an old one.
- **Enterprise-level applications:** Applications such as Microsoft Exchange and SharePoint are sensitive in terms of storage configuration, and each application often includes a mix of random and sequential IO.

- **VDI:** VDI is one of the primary disruptors in the storage market. VDI storage needs are varied. Sometimes, you need only 10 to 20 IOPS per user. At other times, such as when you're handling boot storms and login storms, IOPS needs can skyrocket.
- **DevOps:** In DevOps environments, the ability to quickly and easily refresh a test/dev environment from production or clone an environment is really critical, but it can be IO intensive and it's often random in nature.
- **Data protection:** Traditional data protection tools can introduce significant IO spikes and valleys and, depending on how backups are configured, may present as sequential or random IO at different times.

What the industry has done over the years is combine all these varied workloads. In other words, their very efforts to consolidate environments have created a storage monster. Many storage area network (SAN)–based storage environments suffer big-time problems due to this IO blender:

- Continued consolidation of VMs contributes to random IO workloads, each with its own pattern for reading and writing data to underlying storage.
- Highly random IO streams adversely affect overall performance as VMs contend for storage resources.

VDI boot storm

One situation that perfectly demonstrates a relatively new phenomenon in storage is the *VDI boot storm*, which occurs when many users try to boot their virtual desktops at the same time. The result: Storage devices can't keep up with the sheer number of requests.

The beginning of the day is what really kills storage. As computers boot, the operating system has to read a ton of data and move it to memory so that the system can be used. Now, imagine what happens when hundreds or thousands of users boot up their virtual desktops at the same time. Legacy storage systems crumble under the IO weight, and it can end up taking a long time for users to fully boot their systems.

TIP

The situation is largely mitigated by the use of solid-state storage as a caching layer. Adding this kind of service without considering the administrative inefficiencies that it introduces has been standard operating procedure for quite a while and is one of the main reasons why people implement resource islands when they want to do VDI.

Mitigation

Administrators can work inside their legacy environments in various ways to try to solve the serious IO issues that arise with shared workloads. They can buy

- A separate environment to support each application
- Sophisticated stand-alone storage devices that include automatic tiering features
- Multiple tiers of storage and manage them separately

What do these mitigation techniques have in common? They require administrators to overprovision storage, which requires more investment in storage hardware. They also require additional time by the administrator to configure and manage. These are not the most efficient models to begin with, and eventually they can become unsustainable.

READ/WRITE IO

In traditional data center environments with shared storage, the difference in performance between reading and writing data is incredible. Reading is generally quick and can be accomplished with just a single IO operation. Writing data is a different story; it can take up to six IO operations to accomplish.

As administrators have moved from RAID 5 to RAID 6 for better data protection, they've introduced additional overhead to the storage equation. A RAID 6 write operation requires six IOs to complete. The reason: It's not just the data that has to be written, but also the parity information that has to be written multiple times in RAID 6. RAID calculations also tend to require a lot of CPU overhead to perform the actual parity calculations needed for data protection.

Multiple Touchpoints

Touching data many times in a virtualized environment isn't so great. Consider the following scenario: A legacy, but heavily virtualized, data center has many servers connected to a SAN. The SAN has data deduplication mechanisms. The company backs up its data by using a local disk-to-disk-to-tape method; it also copies certain VMs to a remote data center each day. This way, the company maintains local backups and gains disaster recovery (DR) capabilities.

Quite a bit of redundancy is built into this scenario. Figure 2-1 examines the path the data travels as it wends its way through the various processes associated with the scenario.

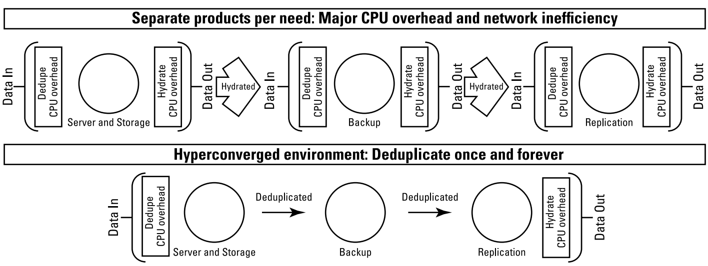


FIGURE 2-1: Hyperconverged infrastructure requires far less CPU power and network bandwidth than nonconverged systems.

Every time the data has to be hydrated and then re-deduplicated as it makes its way to different components, the CPU must be engaged. Deduplication can be an expensive operation and constantly treating data in different locations has several drawbacks:

- Constant CPU use to treat data multiple times limits the number of VMs that can be run in the environment.
- Data takes more time to traverse the network because it isn't in reduced form as it moves between services.
- WAN bandwidth costs are significant as data travels across wide-area connections in unreduced form.

It gets worse. Many storage systems — including those related to data protection — use a *post-process* deduplication method, as opposed to what is known as an *inline* deduplication process.

Post-process deduplication means that data is not deduplicated until after it's actually been written to disk. Here are the steps:

1. **Write data to disk undeduplicated.**

   This requires availability capacity and uses IOPS.

2. **Read data back from disk later.**

   The data then needs to be touched again by the post-process deduplication engine, consuming yet more IOPS and CPU resources.

3. **Invest CPU to deduplicate or compress.**

   After the data is read, it needs to be processed again using more CPU.

This means that the data will be replicated before deduplication, and then all the dedupe work to save capacity must happen at both the primary and disaster recovery sites. This process consumes additional resources, including CPU, IOPS, capacity, and network bandwidth. Post-process deduplication invests all these resources to get a reduction in storage capacity. The trade-off isn't a positive one.

The results: greater costs and lower efficiency. The best outcome in any environment is to eliminate writes to disk before they even happen. In a hyperconverged environment, because of caching in random access memory (RAM), many operations don't have to touch storage.

In the modern data center, data efficiency is about IOPS and WAN bandwidth, not storage capacity. Capacity has become plentiful as vendors release bigger drives (16TB and more!). Originally, data efficiency technologies were focused on the backup market. The objective was to provide an alternative for tape-based technology. In order to make the economics work, the primary objective was to fit more data on the disk while delivering the throughput needed for backups. In short, pack 20 pounds of data into a 5-lb. bag. This was the right solution at the time.

Although spinning disk capacities have gotten bigger, spinning disk performance still faces challenges, which is a key reason that all-flash systems are starting to dominate. In general, organizations may not have a storage capacity problem, but they could have an IOPS problem, which manifests as poor performance.

Some administrators try to solve the performance problem by adding caching layers or fast (but expensive) NVMe storage. Unfortunately, these patchwork solutions can introduce additional complexity or expose other bottlenecks in the architecture. A patchwork architecture isn't an ideal scenario.

Storage and networking are becoming increasingly comingled. The network has long been the holdout when it comes to hyperconverged infrastructure. Hyperconvergence handily takes care of servers and storage, but the network has been limited to internode communication. Today, the network needs to take center stage for a lot of reasons. First, with the addition of disaster recovery to remote locations or to the cloud now a part of most customer environments, the need for a robust networking solution that is well integrated into a hyperconverged platform is critical. The demand for WAN bandwidth has also increased and we now have a bandwidth challenge. Data reduction technologies, such as deduplication, can help address this situation, but it's only a first step.

To bring networking to the hyperconvergence game as a first-class citizen requires the deployment of an intelligent network fabric to relieve stress on the systems and produce more agile and performant environments. These software-defined networking systems provide programmable and adaptable network services that can handle the on-demand services, dynamic workloads, and diverse traffic flows of the contemporary data center. By treating the network as an integral component of a hyperconverged system — with unified administration and automated provisioning — IT organizations can bring cloud agility, scalability, and simplicity to the data center.

Back on the storage front, in a primary storage environment, infrastructure needs to optimize for performance and latency, not capacity and throughput. This requires new technology and an approach to data efficiency that is systemic — which is one of the hallmarks of hyperconverged infrastructure.

Inline deduplication provides the level of efficiency needed and consists of only two steps: Process data and write data. Inline deduplication invests CPU just one time and gets a reduction in IOPS, WAN bandwidth, and storage capacity. These are critical resources, but these resources are only conserved when the data efficiency is delivered inline.

In the modern data center, data efficiency is also about mobility and data protection, particularly when talking about online backup and restore. Data efficiency saves all the IO traditionally required to do backup and restore operations.

Policy Misalignment

In addition to facing the performance challenges of the post-virtualization world, virtualized organizations face policy challenges in both the physical and virtual worlds.

Physical

Physical servers have a direct mapping from application to server to storage array to LUN to storage policy. This results in an environment where an application policy is directly linked to an internal construct of the storage array. There is no abstraction. This approach is what makes storage upgrades so complex. For example, a replication policy is applied to a LUN in storage array X at IP address Y and tells that LUN to replicate to storage array A at IP address B. Imagine the complexity of an array replacement when there are a couple of arrays in a couple of locations, and the replication policies are all tangled together. No wonder there are so many storage administrators in IT.

Virtual

In the virtualized world, there are many applications on a host and many hosts on a single LUN. It isn't efficient to apply a policy to a single LUN if that LUN represents the data for many applications (and hosts). In a hyperconverged environment, backup and replication policies are applied directly to individual applications (or VMs). There are no LUNs or RAID sets to manage. Replication policies specify a destination — in this case, a data center — that is abstracted away from the infrastructure. This allows an administrator to perform a platform upgrade without any policy reconfiguration or data migration, which increases efficiency and decreases risk.

IN THIS CHAPTER

» Virtualizing everything

» Introducing automation

» Using IT as a Service

Chapter 3 Welcome to Software-Defined Infrastructure

What a concept! Software for the modern data center!

Consider the situation just ten years ago. Legacy data centers were hardware-centric places. Storage companies created their own chips and boxes to ship to customers. Networking vendors took a similar approach, creating individual circuits and arrays for their products. Although this approach wasn't necessarily bad, the resulting hardware products were relatively inflexible, and the flexible software layer played a supporting role.

In this chapter, I discuss what has become the modern data center standard: the software-defined data center (SDDC) and the software-defined infrastructure (SDI) that operates inside it. In these environments, software becomes the focus over hardware. Because SDI has several defining characteristics, including virtualization, automation, and the use of IT as a Service (ITaaS), I take a look at these characteristics in detail.

Virtualization

Every SDI environment employs a high degree of virtualization, which goes beyond virtualizing servers. Everything is sucked up into the virtualization vacuum cleaner: storage, servers, and even supporting services such as load balancers, wide area network (WAN) optimization devices, and deduplication engines. Nothing is spared. This eliminates the islands of central process unit (CPU), memory, storage, and networking resources that are traditionally locked within a single-purpose device, such as a backup-to-disk device, and creates a single shared resource pool for both business and infrastructure applications.

Virtualization abstracts the hardware components of the data center and overlays them with a common software layer: the virtualization layer, which manages the underlying hardware. The hardware can be a mix-and-match mess, but it doesn't matter anymore, thanks to the virtualization layer. All the data center administrator has to worry about is making sure that applications are running as expected. The virtualization layer handles the heavy lifting.

Automation

Many boardrooms today are asking companies to do more with less. One of the fastest ways to improve efficiency (and reduce costs) is to automate routine operations as much as possible.

Until now, many legacy IT architectures have been so varied and complex that automation has only been a pipe dream. SDI brings the dream one step closer to reality.

Software-imposed normalization of the data center architecture permits higher degrees of automation. Plus, the software layer itself is often chock-full of automation helpers, such as application programming interfaces (APIs). With this kind of assistance, automation becomes far easier to achieve.

IT as a Service

When resources are abstracted away from hardware and plenty of automation techniques are in place, companies often discover that they can treat many IT services as exactly that: services.

As they do with all other services, companies that use ITaaS have certain expectations:

- **Predictability:** The service should operate in a predictable way at a predictable cost. SDI can provide this conformity.
- **Scalability:** Business needs today may be very different from tomorrow, and the data center can't be a limiting factor when expansion becomes necessary. In fact, a data center should be an *enabler* of business expansion.
- **Improved utilization:** Companies expect to get maximum benefit from the services they buy. Because hyperconvergence-powered SDI is built on common components that eliminate the islands of resources traditionally trapped within infrastructure appliances, high utilization rates are exceedingly easy to achieve.
- **More efficient personnel productivity:** With SDI, staff can operate a data center far more efficiently. The reason is simple: SDI banishes traditional resource islands in favor of the new software-powered matrix. In fact, according to IDC, more than 60 percent of organizations using an HPE SimpliVity hyperconverged infrastructure solution were looking for ways to improve their staff productivity. This has real potential for outcomes: HPE SimpliVity customers spent just 14.8 percent of their time on innovation and new projects before deploying HPE SimpliVity compared with 28.1 percent of their time on innovation and new projects after deploying HPE SimpliVity, a gain of 91 percent.
- **Reduced provisioning time:** A company that invests in a software-defined infrastructure expects to receive business benefits. SDI offers agility and flexibility, which reduce provisioning times for the new services that business units require. Provisioning time comes *way* down with software-defined networking (see Chapter 7).

HARDWARE IN A SOFTWARE WORLD

When people hear the term *software-defined data center*, their first question usually concerns where the software for SDI is supposed to run. The answer is simple: The software layer runs on hardware.

But if SDI is software-centric, why is hardware still required? Again, the answer is simple: You can't run SDI without hardware.

Most hardware in SDI looks quite different from hardware in traditional environments. Whereas legacy data centers have lots of proprietary hardware to manage myriad devices, SDI uses fewer, more consolidated, more efficient hardware components.

If SDI contains any proprietary hardware, the software leverages it to carry out important functions. In the world of hyperconvergence, this kind of hardware essentially becomes part of the data center's standard operations. Because it's identical hardware (and not unique to each device), it scales well as new appliances are added to the data center. Software is still top dog in such an environment, but without the hardware, nothing would happen. And the better the hardware, the better your software can perform.

IN THIS CHAPTER

» Boosting efficiency

» Reducing risk

» Getting — and staying — agile

Chapter 4
What Businesses Want from IT

Did you know that the IT department doesn't exist just to play with technology? Who knew? Apparently, it's far more important for this increasingly critical department to take its eye off the gear and turn it more toward the business.

This attention shift isn't just a good idea; it's a trend being driven hard by CEOs and business-unit leaders who have important needs to meet. Technology pros who yearn to stay ahead of the curve need to hone their business chops.

Expectations of high returns on large data center investments are climbing ever higher, and companies are much less willing to assume risk. They want a data center that

- Improves IT operational efficiency and performance
- Reduces business risk
- Is flexible and agile enough to support changing business needs

This chapter examines these characteristics.

Increased Efficiency

Does your boss ever walk into your office and say something like this:

> We need to have a talk about your performance. You're just too darn efficient, and we need you to dial that back a few notches. If you could do that on Saturday, that would be *great.*

I didn't think so. If anything, IT departments are under growing pressure to *increase* efficiency. Improving efficiency generally means changing the way that IT operates — changes that involve anything from small course corrections to major initiatives.

TIP

One of the greatest benefits of hyperconverged architecture is that it generates efficiency benefits while also streamlining operations.

Using time better

As poet Delmore Schwartz put it, "Time is the fire in which we burn." For those who slog through repetitive, mundane tasks every day, truer words were never written. When it comes to business, any time wasted on mundane work really is burned time — time that could've been spent advancing business objectives.

Management wants IT to spend its time wisely. Traditional IT operations simply won't cut it anymore. Neither will prolonged product evaluation and integration processes or extended return-on-investment metrics. IT needs to be faster and leaner than ever before.

Matching skills to tasks

Step back for a second to think about what the IT staff really has to deal with on a day-to-day basis: servers, hypervisors, storage devices, network accelerators, backup software, backup appliances, replication technology, and a whole lot more. Forget for a moment about the physical effects of this plethora of equipment on the data center. Instead, consider the human toll. Every one of these devices has a separate administrative console that operators have to learn. Also — let's face reality — not every device plays nicely with every other device.

When each device requires vastly different sets of skills to operate, each skill requires ongoing training. Even when you can get a few people in IT trained on everything in the data center, at some point those people may move on, and you may have trouble finding new employees who have the same set of skills.

In addition, every time you bring a unique resource into the environment, you need staff to manage it. As that resource grows, you may need even more staff to keep up with the workload. In essence, you're creating resource islands as you forge ahead.

REMEMBER

Resource islands are inherently inefficient. The broader you can make the IT environment, the easier it is to achieve operational economies of scale.

The bottom line: IT staffs are being crushed under the weight of legacy infrastructure. Each unique resource requires unique skills, and companies aren't adding IT staff at a pace that keeps up with technical needs.

Managing resources wisely

The laws of physics are immutable forces in the land of IT, and these natural laws become apparent when you walk around a data center:

- You find that two objects can't occupy the same space at the same time, which is why equipment occupies separate racks.
- You discover the laws of thermodynamics as you walk behind a rack and feel the blast of hot air, and then walk under the cooling equipment and feel a chill.
- You encounter electromagnetism as you watch electricity-sucking hardware whir away. (On the plus side, all the blinking LEDs put on a pretty impressive light show.)

All these physical resources — space, power, and cooling — require cash to keep running. Every piece of equipment added to a data center requires the use of all these resources.

WARNING

Continuing to add equipment without considering your use of resources will do nothing to increase your overall efficiency.

Meeting budgets

Think back to your first virtualization project. Why did you undertake it? I'm willing to bet that you wanted to make better use of your resources and improve on what used to be a 15 percent average utilization rate for servers.

If you're not using all your equipment at a reasonable level, you're leaving money on the table — money that could be used to fund new projects and innovation. You may not be enjoying the maximum return on an incredibly expensive investment.

TIP

You can make your IT budget more efficient and serve the business better by rethinking the way you provide data center services. Don't think about each individual resource as its own island. Instead, think at a higher level. Instead of focusing on individual resources, focus on overall scale of all resources as your standard operating procedure.

Reduced Risk

Risk can worm its way into an otherwise flawless IT system in several places:

- **Procurement:** With so much hardware to maintain in the data center, staying on top of all the little details can be difficult. Before you purchase any new equipment, ask yourself questions such as these:
  - Are you sure that the storage area network (SAN) you've selected has enough capacity in terms of terabytes *and* performance in terms of input/output operations per second (IOPS)?
  - If you're expanding an existing system, will the expansion create a risk of downtime or data loss?
  - If you're upgrading an existing storage array, are all the component firmware revisions the same or supported by the new hardware?
- **Operations:** Generally, you need two of everything in a data center to maintain the levels of availability that the business expects. Without that redundancy, you run the risk of prolonged outages, which management tends to dislike.

REMEMBER

Redundancy is the norm in IT, but it's an expensive standard. Plus, it requires personnel who have specialized skills to maintain the different high-availability features that ship with each product.

- **Data protection:** Too many companies don't plan their data protection mechanisms carefully, or they rely on many services provided by many companies. As a result, the blame game typically occurs when something unexpected happens. When recovery is job one, no one wants vendors to fight about who's at fault. Keep in mind: Data protection isn't about backup; it's about *recovery.*

CIOs and IT staff want — and need — to lower risk in their organizations, especially in these times with ransomware and other cyberattacks on the rise. Systems and applications must remain highly available, and data must be safe. Unfortunately, as more diverse hardware is installed, achieving these critical goals becomes more difficult.

TIP

Companies can reduce risk by adopting a hyperconverged infrastructure. Hyperconverged systems include all the components required to make a data center operate without the complexity of legacy solutions.

Improved Agility

Quickly getting a new product or service to market often results in long-term business advantages. Today's commerce markets are cutthroat places, and IT issues can't be allowed to get in the way of critical business operations. It's one of the reasons that the public cloud is such a draw for business users.

Any lag in delivering what the business needs could result in the business taking IT matters into its own hands — the so-called "shadow IT" problem, which happens more often than you think. If a business constituent requests a function or service that IT can't be as responsive to, then the business constituent, with the swipe of a credit card, can procure a cloud-based service without IT's involvement — or IT's knowledge that it exists. Shadow IT practices have the potential to introduce risk.

To meet the demands of business units for more speed and agility, many IT departments simply build larger and larger infrastructures, which are inflexible and difficult to manage. Eventually, such a system turns into a house of cards, and the slightest problem can topple it. This scenario is hardly the hallmark of an agile IT infrastructure.

REMEMBER

Agility and *speed* are two mantras that every IT person should adopt. Ensuring that the infrastructure is agile is critical in order for IT to be able to quickly and easily deploy new applications and services in response to business demands.

IN THIS CHAPTER

» **Employing shared, scalable resources**

» **Emphasizing automation**

» **Understanding cloud economics**

Chapter 5
How the Cloud Has Changed IT

Virtualization transformed the data center. The cloud aims to transform the whole of IT, and that transformation is well on its way. Consider this:

- The hybrid cloud has become the new normal. It's more unusual to find a company that *isn't* using cloud in some way than it is to find ones that have adopted it as a part of their workload support strategy.
- Ease of acquisition and an operationalization of expenses were original key drivers for cloud adoption. Today, it's about reducing time to value for new services, optimizing spend, and enabling a broader set of workload opportunities.
- According to research by ActualTech Media, 54 percent of organizations have adopted a hybrid cloud architecture. Some analyst firms estimate that this figure will jump to 80 percent or even 90 percent over the next few years. Estimates vary depending on how people define *cloud,* but the fact remains that hybrid cloud has cemented itself as the new normal.

- On-premises portions of hybrid cloud environments have been architected to include compelling public cloud characteristics in order to improve operations. This is a significant way in which the public cloud has impacted the modern hybrid cloud world.
- Just about every data center architecture available on the market today has built-in integration with various public cloud providers, enabling a seamless hybrid cloud fabric that spans both on-premises and public cloud environments.

For hyperconverged infrastructure, only the last point is really pertinent, and it's what this chapter is about.

Scaling and Sharing Resources

The hallmarks of Google's and Facebook's environments are, among other things, sheer scalability and reasonable economics. Many of these cloud principles have been adapted for use in smaller environments and packaged in hyperconverged products that any company can buy.

Software-centric design

As I mention in Chapter 3, software overtaking hardware in the data center has the potential to lead to very good things. Companies such as Google discovered this potential years ago and tamed their hardware beast by wrapping it inside software layers. A data file inside Google is managed by the company's massively distributed, software-based global file system. This file system doesn't care about the underlying hardware. It simply abides by the rules built into the software layer that ensures that the file is saved with the right data protection levels. Even with expansion of Google's infrastructure, the administrator isn't concerned about where that file resides.

Economies of scale

In a legacy data center environment, growing the environment can be expensive due to the proprietary nature of each individual piece of hardware. The more diverse the environment, the more difficult it is to maintain.

Commodity hardware

Companies such as Google and Facebook scale their environments without relying on expensive proprietary components. Instead, they leverage commodity hardware. To some people, the word *commodity,* when associated with the data center environment, is a synonym for *cheap* or *unreliable.* You know what? To a point, they're right.

REMEMBER

When you consider the role of commodity hardware in a hyperconverged environment, however, keep in mind that the hardware takes a back seat to the software. In this environment, the software layer is built with the understanding that hardware can — and ultimately will — fail. You still want to deploy the best hardware you can find to optimize performance and reduce the chance of failure, but software-based architecture is designed to anticipate and handle any hardware failure that takes place.

Bite-size scalability

Think about how you procure your data center technology now, particularly when it comes to storage and other nonserver equipment. For the expected life cycle for that equipment, you probably buy as much horsepower and capacity that you need, with a little extra "just in case" capacity.

How long will it take you to use all that prepurchased capacity? You may never use it. What a waste! But on the other hand, you may find it necessary to expand your environment sooner than anticipated. Cloud companies don't create complex infrastructure update plans each time they expand. They simply add more standardized units of infrastructure to the environment. This is their scale model; it's all about being able to step to the next level of infrastructure in small increments, as needed.

Resource flexibility

Hyperconverged infrastructure takes a bite-size approach to data center scalability. Customers no longer need to expand just one component or hardware rack at a time; they simply add another appliance-based node to a homogenous environment. The entire environment is one huge virtualized resource pool. As needs dictate, customers can expand this pool quickly and easily, in a way that makes economic sense.

Enabling Automation

Do you think that the typical cloud provider spends an inordinate amount of time rebuilding individual policies and processes each time it makes changes or adds equipment to its data center? Of course not. In the cloud, change is constant, so ensuring that changes are made without disruption is critical. In the world of enterprise IT, things should work the same way. A change in data center hardware shouldn't necessitate reconfiguration of all your virtual machines (VMs) and policies.

Abstracting policy from infrastructure

Because hardware isn't the focus in the software-defined data center (SDDC; see Chapter 3), why would you write policies that target specific hardware devices? Further, because enterprise workloads leverage VMs as their basic constructs, why should a VM be beholden to policies tied to underlying infrastructure components?

Consider a scenario in which you define policies that move workloads between specific logical unit numbers (LUNs) for replication purposes. Now multiply that policy by 1,000. When it comes time to swap out an affected LUN, you end up with LUN-to-LUN policy spaghetti. You need to find each individual policy and reconfigure it to point to new hardware.

Policies should be far more general in nature, allowing the infrastructure to make the granular decisions. Instead of getting specific in a LUN policy, for example, policies should be as simple as, "Replicate VM to San Francisco."

Why is this good? With such a generalized policy, you can perform a complete technology refresh in San Francisco without migrating any data or reconfiguring any policies. Because policy is defined at the data center level, all inbound and outbound policies stay in place. Beautiful.

Taking a VM-centric approach

The workload takes center stage in the cloud. In the case of enterprise IT, these workloads are individual VMs. When it comes to policies in cloud-based environments, the VM is the center of the world. It's all about applying policies to VMs — not to LUNs, shares, datastores, or any other constructs. Bear in mind the plight of the VM administrator, who is VM-centric. Why wouldn't the administrator assign backup, quality-of-service, and replication policies to a VM?

The need to apply policies across individual resource domains creates fundamental operational issues in IT. In the cloud and in hyperconvergence, policies are simply called *policies.* There are no LUN policies, caching policies, replication policies, and so on — just policies.

Understanding Cloud Economics

Cloud providers and enterprise IT organizations operate their environments using very different economic models. CIOs expect enterprise IT infrastructure to last many years, so they buy enough capacity and performance to last that long. In many cases, however, the full potential of the infrastructure buy is never realized. CIOs often overbuy to ensure that capacity lasts the full life cycle.

Conversely, by design or by mistake, CIOs often underbuy infrastructure. The organization then needs to buy individual resources when they begin to run low. This leads to watching individual resources constantly, reacting when necessary, and hoping that your existing product doesn't have end-of-life status.

Now consider cloud vendors, who don't make one huge buy every five years. Doing so would be insane in a few ways:

- A *lot* of hardware would have to be purchased upfront.
- Accurately planning three to five years' worth of resource needs in these kinds of environments may be impossible.

Instead, cloud organizations pay as they grow. Operational scale and homogeneity are core parts of cloud providers' DNA, so they simply add more standard resources as needed.

The public cloud is highly appealing to the enterprise. The instant-on service is elastic and costs only a few cents an hour. What could be better? It's not for all applications; it presents major challenges for many, particularly when it comes to cost predictability. The true cost of public cloud increases dramatically when compared to the cost of predictable storage performance, high availability, backup, disaster recovery, private networking, and more. IT ends up paying for a server that's running at 15 percent utilization, and the cloud provider is benefiting from packing those VMs onto a single host.

IT applications are designed expecting high-availability infrastructure, disaster recovery, backup and recovery, and other necessary services. (This is why internal IT isn't like Facebook and Google.) IT applications place a different set of demands on the infrastructure. Therefore, any hyperconverged infrastructure *must* deliver on these requirements.

Business units may not understand these nuances and may be compelled to buy cloud services without having a grasp on the full picture. This rise of *shadow IT* — where non-IT units create their own systems (see Chapter 4) — is real, and the cloud is complicit in enabling this trend. Shadow IT exists, however, either because IT isn't able to provide the kinds of services business units demand or because they aren't responsive enough. So, these units turn to the cloud to meet individual needs, giving rise to fragmented services, potential security breaches, and overall data-quality issues.

Hyperconvergence brings cloud-type consumption-based infrastructure economics and flexibility to enterprise IT without compromising on performance, reliability, and availability. Instead of making huge buys every few years, IT simply adds building blocks of infrastructure to the data center as needed. This approach gives the business much faster time to value for the expanded environment.

Figure 5-1 shows integrated systems (which scale via large steps; see Chapter 6) and hyperconvergence (which scale via bite-size steps; see Chapters 6 and 8), with wasted resources above the wavy line.

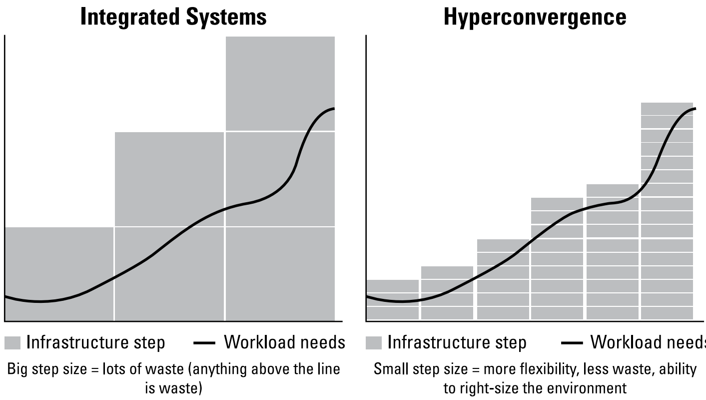


FIGURE 5-1: The unit of scale has a significant impact on economics.

IN THIS CHAPTER

- » **Seeing how convergence evolved**
- » **Introducing hyperconverged infrastructure**
- » **Making the most of differences**

Chapter 6 Understanding Converged Infrastructure

The IT infrastructure market is undergoing unprecedented transformation. The most significant transformation is reflected by two major trends: convergence and software-defined infrastructure (SDI). Both trends are responses to the IT realities of infrastructure clutter, complexity, and high cost; they represent attempts to simplify IT and reduce the overall cost of infrastructure ownership. Data center innovations in recent years have served to consolidate and converge what were once fractured traditional infrastructure environments. Of course, such traditional environments still exist, but they're being increasingly displaced by converged, hyperconverged, and other software-driven infrastructure options.

The benefit of these modern approaches over traditional infrastructure are clear, particularly as you peel back the multitude of layers that traditional environments are composed of to see what makes them tick. Such environments are typically composed of

8 to 12 hardware and software products from nearly as many vendors, with each product offering a different management interface and requiring different training. Chaos ensues.

To make matters worse, each product in the legacy stack is often grossly overprovisioned, using its own resources (central processing unit [CPU], dynamic random access memory [DRAM], storage, and so on) to address the resident applications' intermittent peak workloads. The value of a single shared resource pool, offered by server virtualization, is still limited to the server layer. All other products become islands of overprovisioned resources that aren't shared. Therefore, low utilization of the overall stack results in the ripple effects of high acquisition, space, and power costs. Simply put, too many resources are wasted in legacy environments.

This chapter explores one leading solution: convergence, which ultimately leads to hyperconvergence.

The Evolution of Convergence

The following sections describe the evolution of convergence over the past few years.

Integrated systems

Early iterations of infrastructure convergence solutions had complete network, compute, storage, and virtualization capabilities, but in many instances they were simply conglomerations of existing hardware and software, with little to no actual innovation in product features to be leveraged. Although still in use in some solutions, these types of solutions are quickly falling out of the mainstream as more capable solutions enter the fray.

But integrated systems did offer a few benefits, which carry forward to a number of modern solutions. Most notably, customers gained a single point of contact for their infrastructure, from purchase to end of life. These systems are always tested and almost always arrive at the customer site fully racked and cabled, so they're ready to go.

On the downside, these systems often have a big step size and a price tag to match. When you need more power, you have to buy a big chunk of infrastructure. Also, these products don't always solve the serious challenges that so many organizations face (see Chapter 2).

Converged infrastructure

Converged infrastructure products combine the server and storage components in a single appliance, effectively eliminating the need for dedicated storage area network (SAN)–based storage.

These systems provide a localized single resource pool solution, offering simplified management and faster time to deployment. They have effectively virtualized the storage layer and allowed it to run in the virtualization platform. Overall acquisition cost is lower, and management (at least, for the server and storage resources) is simplified. With these systems, overall resource utilization is higher than with a legacy island-based infrastructure.

Converged infrastructure has some limitations, however:

- The systems often include just the server and storage resource components.
- Many of the fundamental data management challenges haven't been solved. It's the functionality of a traditional storage array migrated into the virtualization platform.
- Resource ratios (such as CPU-to-storage-to-network) are fixed, making them less flexible than some organizations require.
- The products can't always be used by existing infrastructure. In other words, you may not be able to use a converged infrastructure appliance's storage from existing legacy systems. In essence, you're forced to create a separate resource island.

For these reasons, converged infrastructure systems don't sufficiently address performance problems in every legacy infrastructure.

Likewise, on the data front, the systems don't address all data problems, because not all data efficiency appliances are converged. Management may be improved, but it's not unified global management.

Hyperconverged infrastructure

Enter hyperconvergence, also known as *hyperconverged infrastructure,* which represents the logical next step in the evolution of infrastructure convergence. Hyperconvergence delivers simplification and savings by consolidating all required functionality into a single infrastructure stack running on an efficient, elastic pool of x86 resources. The underlying data architecture has been completely reinvented, allowing data management to be radically simplified. As a result, hyperconverged infrastructure delivers on the promise of the software-defined data center (SDDC) at the technological level. It also carries forward the benefits of convergence, including a single shared resource pool.

Hyperconvergence goes far beyond servers and storage, bringing into the convergence fold many features that make standalone appliances and services obsolete in some IT environments, including the following:

- Data protection products (backup, replication)
- Deduplication appliances
- Solid-state drive (SSD) arrays
- SSD cache arrays
- Public cloud gateways
- Replication appliances or software
- Certain networking appliances, replaced by intelligent network optimization services

In Chapter 8, I delve into how hyperconvergence takes convergence to the next level in the data center and provides a plethora of benefits to both IT and to the business.

Convergence Characteristics

The preceding convergence options build on one another, with each having key differences. Figure 6-1 illustrates the high-level characteristics that define each convergence type.

Each characteristic is critical to realizing all the traits that business demands of IT in the modern era: lean, mean, and green.

	Technical Attributes			Organizational Benefits	
	Data Efficiency	Single Shared Resource Pool	Global Management	TCO Improvements	Simplification
Integrated systems	☐	☐ Resource pooling limited to server layer	☐	☐	⊡ Some time to deployment and administrative gains
Convergence	☐	⊡ Limited to primary server and storage resources; other resources not included	☐	⊡ TCO gains primarily due to reduction of legacy gear; does not address backup, replication, and DR	⊡ Fewer products to manage
Hyperconvergence	▣ Data architecture begins with one-time deduplication, compression, and optimization of data	▣ All data center resources are brought into the resource stack	▣ Complete management of all infrastructure resources and virtual machines; single point of administration	▣ Major TCO gains through reduction of hardware resources, streamlined operations, and automation	▣ Reduces hardware littered across data centers, eases management, VM-centricity

☐ ***Not Supported*** ⊡ ***Partially Supported*** ▣ ***Fully Supported***

FIGURE 6-1: Comparing improvements as convergence has evolved.

IN THIS CHAPTER

» Understanding why your network may not support your initiative

» Recognizing the barriers and bottlenecks and how to remove them

» Deploying an intelligent fabric that can scale with your business

» Ensuring long-term success with a few simple steps

Chapter 7
The Importance of the Network for Hyperconverged Infrastructure Agility

Legacy network solutions and disjointed administrative practices can impair large-scale adoption of hyperconverged infrastructure solutions. When it comes to hyperconvergence, networking used to be a bit like the weather. As Mark Twain famously said, everybody talks about it, but nobody does anything about it. That's because there weren't any intelligent networking alternatives that could boost hyperconvergence performance and agility. Thanks to continued innovation in the world of software-defined infrastructure (SDI), that's all changing. Many businesses are now considering data center networking implications and requirements when planning hyperconverged infrastructure initiatives, and for good reason: Intelligent network fabrics can help boost resource utilization, increase agility while ensuring workload support and security, and scale lock-step with the business.

Traditionally, many organizations have operated — and, all too often, continue to operate — the network as a discrete technology silo, managed by a separate team using distinct administrative tools and practices. Networking infrastructure has typically been deployed and provisioned independently of compute and storage infrastructure, consuming time and human resources.

Such legacy data center networking solutions and fractured administrative practices can hinder IT service agility and can actually hold back large-scale adoption of hyperconvergence. Whereas hyperconverged infrastructure provides nimble deployment and scaling capabilities, it can take days or even weeks to stand up associated networking infrastructure to support new applications and projects. Worse still, legacy data center networks designed to support traditional client–server applications can't handle the dynamic workloads and diverse traffic flows that dominate today's data center.

In short, in traditional IT infrastructures the network has become a barrier actively holding back newer data center opportunities. To take full advantage of all the benefits of hyperconvergence, it's time to bring the network to the current era. The modern data center demands a modern data center network.

Understanding Traditional Networking Challenges

Customer demands have raised the bar for increased agility, scalability, cost and performance requirements for hyperconverged infrastructure and the private cloud. Most incumbent data center networks are based on traditional *leaf–spine topologies* composed of distinct top-of-rack (ToR) switches and aggregation switches designed as shown in Figure 7-1.

These two-tier network designs are inherently inefficient and inflexible and feature a number of limitations, covered in the following sections.

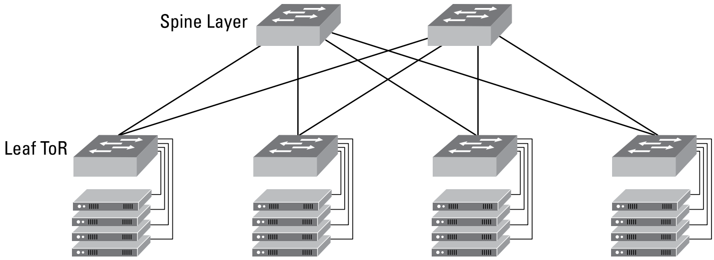


FIGURE 7-1: Traditional leaf-and-spine fabric.

Design constraints

The first limitation revolves around the very design of the network. Traditional data center leaf-and-spine configuration with a set of ToR leaf nodes aggregates rack-based servers and storage. This leaf switch is connected to a cross-rack spine layer to provide broader connectivity to other hosts. This type of design favors workloads that can be contained locally in the rack. Problems arise when heavy rack-to-rack traffic is encountered and one of the nodes has to wait for available bandwidth. Bandwidth crowding can increase transfer times, latency, and potential link timeouts, and applications suffer.

Unfortunately, this design can also introduce load-balancing bottlenecks. Rudimentary load distribution protocols like equal-cost multi-path (ECMP) are not application-aware. ECMP arbitrarily distributes traffic based on network costs, without considering an application's requirements, tendencies, or performance characteristics. As a result, upstream traffic is often distributed unevenly, which leads to hot spots, bottlenecks, and poor service quality.

Protocol boundary constraints

The traditional approach also introduces protocol boundary constraints. Most enterprises use layer 2 protocols in the leaf tier and layer 3 protocols in the spine tier to simplify load balancing. But the approach can lead to IP addressing discrepancies when distributing or moving workloads across the data center, forcing IT teams to build "overlay" networks to make applications or virtual machines (VMs) appear to belong to the same network. These types of overlays create an additional layer of complexity that needs to be managed by specialists. Figure 7-2 provides a look at this situation.

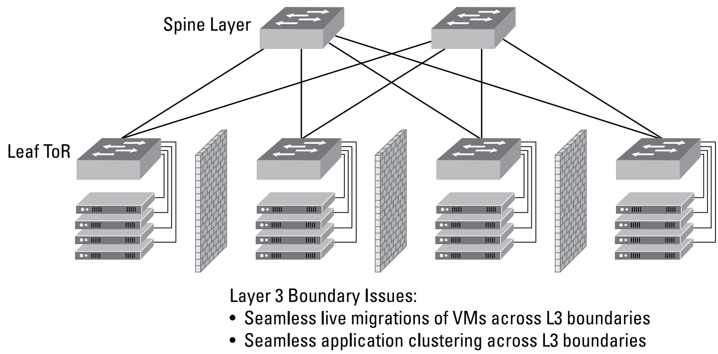


FIGURE 7-2: Layer 2/Layer 3 rack boundaries.

Geographic constraints

Finally, we're at one of the biggest problems: geographic constraints. Thanks to excessive latency, it's exceedingly difficult to address constraints that can restrict scale-out applications and workload mobility in geographically distributed data centers. IT organizations deploying hyperconverged infrastructure for disaster recovery can't leverage multiple data centers for load distribution of active workloads without complex workarounds.

Improving Private Cloud Performance

Private clouds perform better with integrated, intelligent data center networking. Traditional approaches don't cut it. As you may have guessed, there is a solution. There are technologies that break away from the traditional leaf-and-spine model, instead going with a software-defined *composable networking fabric.* This software-centric approach facilitates network management, workflow automation, dynamic and multi-purpose path segmentation, and no-touch scalability with commodity hardware and integrated software.

In Figure 7-3, you can see how data traffic patterns have evolved from the early days of hardware-constrained, multi-tier networking, to leaf-and-spine architecture, and ultimately to extremely agile, composable fabric networking. Software-defined networks eliminate the architectural constraints and price-performance limitations of traditional leaf–spine data center networks.

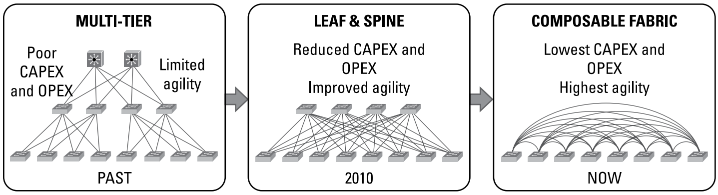


FIGURE 7-3: Traffic pattern evolution from hardware-constrained to software-defined networking.

What makes this method so successful and desirable is the fact that the networking component of the solution is no longer an after-thought, but an integrated, pretested component of the hyper-converged stack that includes compute, storage, *and networking.* That's the key. Even better, it's all managed through a single administration interface. By consolidating the fabric into a single tier, the complexity of the physical network is removed, making it simpler to manage and reducing overall cost. This also enables an agile, integrated network fabric that efficiently supports diverse applications and workloads. The programmable network fabric adapts in real time to ensure high service quality, reliability, and performance for next-generation applications, as well as conventional workloads.

A composable fabric supports an efficient network topology, where ToR switches are directly interconnected using a virtual spine layer. This approach eliminates physical aggregation switches, removes artificial workload restrictions based on rack boundaries, and accelerates application performance and workload mobility. It effectively deconstructs the network into distinct data, control, and integration planes for ultimate scalability, flexibility, and extensibility.

And it results in good news for hyperconvergence. This approach enables better disaster recovery through network awareness and makes remote office/branch office (ROBO) deployments faster with zero-touch configurations and automated workload optimization. In English, it means you can deploy hyperconverged infrastructure solutions faster, cheaper, and with less effort in the data center and at the edge.

Delineating Key Features and Benefits

A number of benefits come from integration of intelligent software-defined networking (SDN) with a hyperconverged infrastructure solution:

- **Rapid time to value:** IT generalists can commission a complete hyperconverged infrastructure platform with SDN quickly and easily, reducing private cloud turn-up times from hours to minutes.
- **On-demand service activation:** Operators, developers, and test engineers can instantly provision IT services (including network connectivity) to streamline application life-cycle management.
- **Unique network design:** SDN topology obviates the need for physical aggregation switches and eliminates artificial workload restrictions based on rack boundaries.
- **Programmable and adaptable network fabric:** SDN can deliver a dynamic network fabric that automatically adjusts bandwidth to satisfy specific application needs and changes in the infrastructure.
- **Elastic resource pools:** Operators can add nodes as needed, to any rack, at any time. SDN provisions network resources automatically to satisfy real-time application requirements and workload demands.
- **Enable true pay-as-you-grow economics:** A scale-out design minimizes upfront investments and tightly aligns ongoing equipment and operations expenses with business demands.
- **Friction-free operations:** Operators can transparently expand and upgrade infrastructure, roll out new projects, and move applications from development to test to production without disrupting services.
- **Radically simplified administration:** Unified administration of the SDN with the rest of the infrastructure eliminates swivel-chair management and manually intensive, error-prone configuration and troubleshooting tasks for ultimate simplicity and convenience.

Ensuring Long-Term Success

Mashing up SDN and hyperconverged infrastructure helps to ensure the long-term success of your hyperconverged deployment. Businesses are turning to hyperconverged infrastructure to jump-start private cloud initiatives and accelerate digital transformation. Unfortunately, legacy data center networking solutions and fractured administrative practices hamper IT service agility and impede large-scale hyperconverged infrastructure rollouts.

Fix it! Enterprise architects deploying hyperconverged infrastructure can take a number of different steps to ensure the long-term success of their hyperconverged infrastructure deployments:

- When designing the hyperconverged infrastructure cluster and implementation, make the network design an integrated element. This will ensure performance and reliability.
- Mandate that all hyperconverged infrastructure vendors integrate with the network operating system to deliver application and workload views of system performance.
- Chart the network design to the traits of your hyperconverged infrastructure deployment and supported applications.

If you take these steps, you'll have an infrastructure deployment that is fast, robust, cost effective, easy to manage, and eminently scalable.

IN THIS CHAPTER

» Focusing on software

» Centralizing management

» Enhancing data protection

Chapter 8 Ten Things a Hyperconverged Infrastructure Can Do for You

How does a hyperconverged infrastructure bring together all the important trends that enterprise IT struggles to handle? Here's a look:

- **Hyperconvergence is the embodiment of software-defined infrastructure (SDI; see Chapter 3).** Based in software, it provides the flexibility and agility that business demands from IT.
- **Cloud operators have their economic model figured out.** Hyperconvergence brings to enterprise IT a cloudlike economic model that provides faster time to value for data center expenditures and lower total cost of ownership (TCO) for the entire solution. A hyperconverged infrastructure offers the economic benefits of the cloud while delivering the performance, availability, and reliability the enterprise demands.

- **Flash solves performance issues but is not the answer for every performance problem.** Hyperconverged solutions reduce storage, bandwidth, and input/output operations per second (IOPS) requirements, eliminating resource islands (see Chapter 2) and boosting overall infrastructure performance and efficiency.
- **The converged infrastructure market provides a single-vendor approach to procurement, implementation, operation, and even networking.** There's no more vendor blame game, and there's just one number to call when a data center problem arises.

In this chapter, I dive a bit deeper into hyperconvergence, showing you ten ways that it solves the challenges inherent in virtualized data centers (see Chapter 2).

Software-Defined Infrastructure

Hyperconvergence is the epitome of SDI (see Chapter 3). The software-based nature of hyperconvergence provides the flexibility required to meet current and future business needs without having to rip and replace infrastructure components. Better yet, as vendors add new features in updated software releases, customers gain the benefits of those features immediately, without having to replace hardware.

Centralized Systems and Management

In a robust hyperconverged infrastructure, all components — compute, storage, intelligent networking, backup to disk, cloud gateway functionality, and so on — are combined in a single shared resource pool with hypervisor technology. This simple, efficient design enables IT to manage aggregated resources across individual nodes as a single federated system.

Mass centralization and integration also happen at the management level. Regardless of how widespread physical resources happen to be, hyperconverged systems handle them as though they were all sitting next to one another. Resources spread across multiple physical data centers are managed from a single, centralized interface. All system and data management functions are handled within this interface, too.

Enhanced Agility

Agility is a big deal in modern IT. Business expects IT to respond quickly as new needs arise, yet legacy environments force IT to employ myriad resources to meet those needs. Hyperconverged infrastructure enables IT to achieve positive outcomes much faster.

Part of being agile is being able to move workloads as necessary. In a hyperconverged world, all resources in all physical data centers reside under a single administrative umbrella (see the preceding section). Supporting workload migration and automated DevOps processes in such environments is a breeze, particularly in a hyperconverged solution that enables consistent deduplication as a core part of its offering. Reduced data is far easier to work with than fully expanded data and helps IT get things done faster. When the hyperconverged infrastructure is deployed on an intelligent composable fabric (see Chapter 7), an IT generalist can administer the whole system from one console.

Scalability and Efficiency

Hyperconvergence is a scalable building-block approach that allows IT to expand by adding units, just like in a LEGO set. Granular scalability is one of the hallmarks of this infrastructure. Unlike integrated systems products, which often require large investments, hyperconverged solutions have a much smaller step size. *Step size* is the amount of infrastructure that a company needs to buy to get to the next level of infrastructure. The bigger the step size, the bigger the upfront cost.

TIP

The bigger the step size, the longer it takes to fully utilize new resources added through the expansion. A smaller step size results in a far more efficient use of resources. As new resources are required, it's easy to add nodes to a hyperconverged infrastructure. The composable network will recognize and configure new devices for you, as you add them.

Very Borg-like, eh (in a good way)?

Cost-Effective Infrastructure

Hyperconverged systems have a low cost of entry and a low TCO compared with their integrated system counterparts and legacy infrastructure. (For more on integrated systems, see Chapter 6.)

Easy Automation

Automation is a key component of SDI (see Chapter 3) and goes hand in hand with hyperconvergence. When all resources are truly combined and when centralized management tools and robust application programming interfaces (APIs) are in place, administrative functionality includes scheduling opportunities, as well as scripting options. Modern infrastructure solutions feature extensible APIs that expose most or all of the solution's capability via programmatic interfaces that make automation a breeze.

Also, IT doesn't need to worry about trying to create automated structures with hardware from different manufacturers or product lines. Everything is encapsulated in one nice, neat environment.

Focus on Virtual Machines

Virtualization is the foundation of SDI (see Chapter 3). Hyperconverged systems use virtual machines (VMs) as the most basic constructs of the environment. All other resources — storage, backup, replication, load balancing, and so on — support individual VMs.

As a result, policy in the hyperconverged environment also revolves around VMs, as do all the management options available in the system, such as data protection policies, which are often defined in third-party tools in legacy environments. With hyperconvergence, integrated data protection policies and control happen right at the VM level. (I discuss data protection later in this chapter.)

VM-centricity is also apparent as workloads need to be moved around to different data centers and between services, such as backup and replication. The administrator always works with the VM as the focus, not the data center and not underlying services, such as storage.

Shared Resources

Hyperconvergence enables organizations to deploy many kinds of applications in a single shared resource pool without worrying about the dreaded IO blender effect (see Chapter 2), which wrecks VM performance.

Of course, with all-flash storage systems, the IO blender is a thing of the past, right? Not really. All-flash storage has eliminated certain *aspects* of the IO blender — the elements that were present due to the mechanical nature of spinning disks do, in fact, disappear, but they're replaced by other bottlenecks in the architecture, such as controllers and the storage fabric itself. In a traditional storage design, absent some technology that provides insight into storage traffic origination, a block-based storage layer doesn't have deep insight into where workloads came from, so they're still blended together and treated in a common way.

With a well-designed hyperconverged infrastructure solution, the tight coupling between the hypervisor and the underlying storage layer effectively eliminates the input/output (IO) blender. In fact, hyperconverged storage handles both random and sequential workloads with ease. The nature of an all-flash hyperconverged infrastructure solution means that there are more than enough input/output operations per second (IOPS) to support even the most intensive workloads — including virtual desktop infrastructure (VDI) boot and login storms (see Chapter 2).

The shared resource pool also enables efficient use of resources for improved performance and capacity, just like those very first server consolidation initiatives that you undertook on your initial journey into virtualization. Along the way, though, you may have created new islands thanks to the post-virtualization challenges discussed earlier. Resource islands carry with them the same utilization challenges that your old physical environments featured. With hyperconvergence, you get to move away from the need to create resource islands just to meet IO needs of particular applications. The environment itself handles all the central processing unit (CPU), random access memory (RAM), capacity, and IOPS assignments and, in some instances, even graphics processing unit (GPU) capabilities, so that administrators can focus on the application and not individual resource needs.

The right mix of on-premises IT based on a software-defined architecture can enable you to deliver IT-as-a-Service (ITaaS) to your business, while avoiding public cloud risks. The business benefits because IT spends less while providing improved overall service. On the performance front, the environment handles far more varied workloads than legacy infrastructure can. IT itself performs better, with more focus on the business and less focus on the technology.

Data Protection

Although it's not always the most enjoyable task in the world, protecting data is critical. The sad fact is that many organizations do only the bare minimum to protect their critical data. That's because comprehensive data protection can be really expensive and really complex.

To provide data protection in a legacy system, you have to make many decisions and purchase a wide selection of products. In a hyperconverged environment, however, backup, recovery, and disaster recovery are built in. They're part of the infrastructure, not third-party afterthoughts to be integrated.

The benefits of hyperconvergence are clear:

- Comprehensive backup and recovery and affordable disaster recovery
- Efficient protection without data rehydration and re-deduplication — and the inefficient use of resources that result

- A single centralized console that enables IT to respond quickly
- Disaster recovery automation and orchestration tools to automate failover and failback of workloads from primary to secondary sites and to reduce errors in the recovery process
- An intelligent networking fabric that automatically redirects traffic in case of a hardware failure

Immediate Gratification

The great thing about hyperconvergence is that it doesn't necessarily require you to replace existing infrastructure in order to be of immediate value. You can gain the benefits of a hyperconverged infrastructure starting now:

- **Consolidating servers and data center:** Are you tackling a new consolidation project or building a new data center? Leading hyperconvergence vendors provide products that integrate seamlessly with your existing environment — and services to help you plan and deploy systems in more complex environments.
- **Modernizing technology smoothly:** The beauty of hyperconvergence is its nondisruptive implementation. The hyperconverged environment is part of your overall environment, so you can phase in new architecture while you phase out the old, implementing and expanding as funds allow. If applications in the legacy environment need the storage performance provided by the hyperconverged environment, they can leverage those resources.
- **Deploying new tier-1 applications:** Is your existing environment suitable for new tier-1 workloads? Instead of simply throwing more resources at an outdated environment, deploy the new workload in a hyperconverged environment to gain the inherent operational benefits.
- **Managing sites remotely:** In a hyperconverged environment, the entire infrastructure is controlled by a single management system. You don't need to have remote personnel perform manual operations such as running backup jobs or creating logical unit numbers (LUNs) or quality-of-service policies.

- **Performing testing and development:** Many organizations operate test and development (test/dev) environments so that bad code isn't released into production. Hyperconvergence supports test/dev and production needs, with management tools that can help you create logical separations between these functions.
- **Modernizing backup and implementing disaster recovery:** If you don't do a good job with either backup or disaster recovery, run — don't walk — toward hyperconvergence as your infrastructure architecture.

Transform your data center with hyperconvergence

You've converged your workloads, so why not your infrastructure, too? Collapsing the entire IT stack into a single hyperconverged system streamlines operations. Hyperconvergence resolves interoperability issues, simplifies management, reduces costs, and brings cloudlike efficiency to your enterprise. This book shows you the ins and out.

Inside...

- Hyperconverge everything — not just servers and storage
- Take a VM-centric approach
- Introduce a new economic model
- Create new levels of efficiency
- Identify the challenges of virtualization on siloed infrastructure
- See how hyperconverged goes above and beyond converged

Hewlett Packard Enterprise | intel

Scott D. Lowe is a former CIO, the founder and managing consultant of The 1610 Group, and cofounder of ActualTech Media. Scott has written thousands of articles for various publications and regularly contributes to VirtualizationSoftware.com and EnterpriseStorageGuide.com.

ISBN: 978-1-119-56249-8
Not For Resale

9 781119 562498