ActualTech Media

SCALE COMPUTING

Going Small to Get Big

Why SMB and the Edge have Emerged as Pivotal Use Cases for Hyperconverged Infrastructure

Scott D. Lowe

CONTENTS

The Solution Landscape....2

Technical Decisions....2

From Continents to Islands....2

Hyperconvergence Builds Huts....3

Grow Small to Get Big....3

Cutting Down the Resource Jungle....3

Scale Computing at the Edge....3

Summary....4

INTRODUCTION

For more than a decade, the term *hyperconverged infrastructure* has graced the digital pipes of the internet as the technology it describes takes the world by storm. Hyperconverged infrastructure is a technology that collapses the previously disparate server, storage, and hypervisor silos that plagued the data center landscape. The introduction of hyperconverged infrastructure forever transformed this landscape and brought to the market an option that drives cost and complexity out of the data center equation.

The Solution Landscape

As happens with revolutionary technology categories, a number of solutions jumped into the market, each with critical nuance that defined their path to the customer data center. The technical decisions made by vendors have become key points in how they've gone to market and how their solutions are consumed by their customers.

At the same time, the enterprise IT landscape is constantly undergoing tectonic activity, shifting as new technologies, new uses cases, and new needs act as geological forces. It is the intersection of these decisions and the trends that have emerged in the past decade that have brought us to where we are today.

TECHNICAL DECISIONS

At its core, hyperconverged infrastructure solves challenging storage issues. The technology is clearly more than just storage, but this was the key driver behind its initial development. As a result, as a part of the architectural decisions being made in product development, how storage management is accomplished is a milestone design decision. The choice impacts the solution's ultimate scale, performance, and cost.

> The IT landscape is constantly undergoing tectonic activity, shifting as new technologies, new uses cases, and new needs act as geological forces.

In general, there are two methods by which most hyperconverged solutions on the market choose to handle storage management. The first is to create a virtual machine dedicated to managing all of the storage resources on a local node. This *controller virtual machine* is in the data path for all I/O operations that take place on the host and between hosts in the cluster.

The second method is to make use of a hypervisor operating system construct – a kernel model, for example – to handle storage operations. Rather than a controller virtual machine, this kernel module is the key player in the data path to handle storage operations.

Each approach has pros and cons. The controller virtual machine approach provides more portability and flexibility when it comes to hypervisor choice. The kernel module approach provides a bit more performance and integration with the rest of the I/O stack. But, vendors on both sides have made architectural decisions that are now impacting their ability to meet emerging demands in the market.

From Continents to Islands

As the transformation process continues its inexorable journey forward, new workload needs have become front-and-center issues for IT in organizations large and small.

At one time, the data center was the Pangea-like supercontinent that had to be governed by IT. Eventually, this data center environment drifted into separate continents as the hybrid cloud was born. Under this paradigm, organizations sought to embrace the public cloud while also retaining on-premises data centers.

In recent years, series of islands have emerged in the enterprise IT geography and have become known as the *edge*. The edge is a roll-up of a number of formerly separate use cases, including remote office and branch office (ROBO) environments. ROBO environments have always been difficult and expensive to support as central IT has had to place a full stack of data center hardware on-premises in these locations or try to work out ways whereby these locations can consume centralized IT resources, a scenario that introduces a precarious dependence on a never-disrupted connection to the internet.

ROBO environments carry with them some critical challenges that aren't always present in more environmentally-controlled centralized data center locations:

- **Lack of support.** Very few ROBO environments have dedicated support staff, so outages may take longer to repair and be more expensive
- **Uncertain environment.** Some locations simply don't have space for a full-height rack of servers, nor do they have the dedicated cooling that would be necessary to support it

- **Expensive deployments.** Traditional multi-tier IT deployments aren't generally inexpensive, making it difficult to scale to hundreds or thousands of individual sites in a cost-effective way
- **A need for speed.** Having small workload requirements needs doesn't equate to a lack of concern for performance

HYPERCONVERGENCE BUILDS HUTS

There are a number of sought-after outcomes that organizations try to achieve when they adopt hyperconverged infrastructure solutions:

- **Ease of scale.** Start small and go bigger as workload demands dictate
- **Ease of operations.** Centralize administrative functions for servers, storage and the hypervisor
- **Ease of budgeting.** Look for ways to reduce capital expenditure costs as well as total cost of ownership
- **Ease of maintaining performance.** Workloads maintain acceptable performance levels

At first glance, these hoped-for outcomes would seem to align perfectly to solve the challenges with edge and ROBO deployments. And this has been true, at least to a point. Part of the problem is the product development decisions and directions chosen by hyperconverged infrastructure vendors.

In fact, there are three key problems:

1. The underlying assumption is that hyperconverged solutions will start small and grow big
2. The storage management construct – a controller virtual machine or kernel module – consumes an inordinate amount of resources, leaving less for workloads
3. There is a hypervisor licensing cost that can become significant at scale

So, what's the answer?

Grow Small to Get Big

Edge and ROBO needs are often (but not always) lesser than those that exist in the primary data center. Workloads may include point of sale systems, security cameras, and timeclocks. In other words, there is a limited universe of applications that operate in such environments, so the infrastructure doesn't need to be as substantial as that of the centralized data center.

With this in mind, it becomes clear that the node size needs to be smaller and the stack still has to be highly available. And, to solve the expense problem, the hardware and software need to be aligned to these miniaturized workloads.

Going small doesn't just mean supporting ROBO environments, either. There are tens of thousands of small and medium sized businesses (SMBs) that require robust data center architecture that is cost-effective and simple to use. Many of the hyperconverged infrastructure solutions on the market today are focused on larger companies, as evidenced by the design decisions that have been made in the product architectures.

CUTTING DOWN THE RESOURCE JUNGLE

In many hyperconverged infrastructure solutions, making nodes smaller – less CPU, less RAM, and less disk – simply isn't feasible, making them difficult to justify in a ROBO or SMB scenario. The storage construct has been architected in a way that makes it a resource sinkhole. In fact, in some architectures, this controller construct can consume a whopping 24 GB of RAM. In a 32 GB node, this would leave a pitiful 8 GB of RAM for workloads. In a 64 GB node, you're left with 40 GB, so overhead is still substantial.

Moreover, these controller virtual machines require CPU cores to operate, making it difficult to scale these systems down to ROBO- and SMB-sized chunks.

Instead, you need an architecture that minimizes processing overhead and brings a cost-centric focus to the ROBO and SMB equation.

Scale Computing at the Edge

This is exactly where Scale Computing brings a welcome sunrise to the ROBO islands and SMB mini-continents that dot the sprawling business landscape. Scale Computing has perfected a storage management model that results in the consumption of a fraction of the processing overhead of competing solutions. In **Figure 1**, you can see that, for a 32

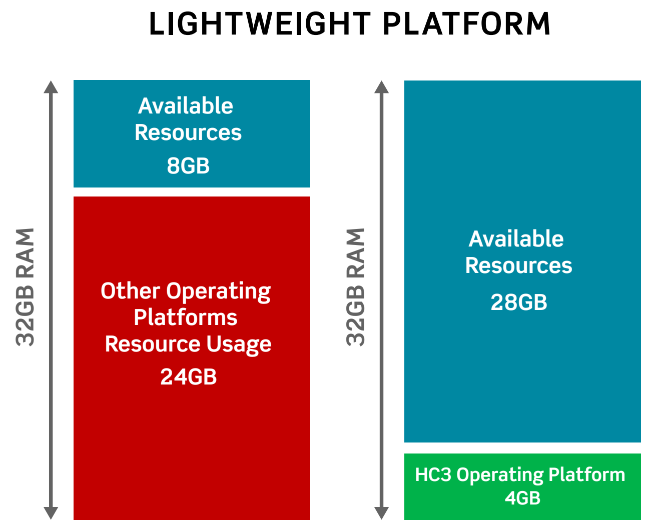

Figure 1: Node storage management overhead comparison

GB node, Scale Computing's approach leaves available 28 GB of RAM for workloads on a node whereas a competing solution's leaves just 8 GB.

If this feels a little like magic, rest assured that it's not. The Scale Computing Reliable Independent Block Engine (SCRIBE) storage engine was purpose-built to provide highly available, scalable compute and storage services at the core and at the edge with as little overhead as possible. If you'd like to learn more about how SCRIBE works and how Scale keeps overhead so low, read the HC3, SCRIBE and HyperCore Theory of Operations paper.

This architecture makes it possible to place more workloads on individual nodes and enables a solution that meets every requirement of sprawling ROBO environments. Scale Computing allows customers to match the right storage configuration against their workloads. Storage solutions start with appliances that include four hard drives anywhere from 1 TB to 8 TB in size each up through an all-flash solution that sports up to four 960 GB SSDs... and everything in between.

> This architecture makes it possible to place more workloads on individual nodes and enables a solution that meets every requirement of sprawling ROBO environments.

For those that wish it, all Scale Computing ROBO-based clusters can be configured to back themselves up to a centralized cluster to enable data protection and disaster recovery for edge sites.

For ROBO and SMB needs, cost is a key consideration. Thanks to Scale's inclusion of their purpose-built hypervisor, you don't have infrastructure-centric software licensing costs to deal with. In fact, affordable three-node clusters make Scale Computing eminently suitable for edge computing, ROBO sites, disaster recovery, and core SMB workload needs.

> While many pundits talked about how big hyperconvergence can scale, Scale Computing worked secretly in the background to drive inefficiency out of their architecture in an effort to scale down while everyone else scaled up.

Further on the cost front, the simple nature of Scale Computing's solution means that a store manager or IT generalist can perform maintenance on the cluster, including replacing failed drives, rebooting virtual machines, and rebooting nodes. There's no more need to roll out an expensive technician for routine maintenance tasks, which, in the long-term, can be a serious cost savings.

Summary

While many pundits talked about how big hyperconvergence can scale, Scale Computing worked secretly in the background to drive inefficiency out of their architecture in an effort to *scale down* while everyone else scaled up. This ability to scale nodes down to mini-size is providing Scale with the tools it needs to grow much larger by helping small businesses and sprawling organizations create highly available branch office and edge computing recipes that are affordable, supportable, and dependable.