SCALE COMPUTING

# GROWING MANAGED SERVICES PROFITABLY BY DOING MORE WITH LESS

## EXECUTIVE SUMMARY

The market for managed services has grown fiercely competitive, forcing Managed Service Providers (MSPs) to seek ways to remain competitive while becoming more profitable. The hype surrounding the simplicity hyperconvergence, the integration of the compute, storage and virtualization layers of infrastructure into a single solution architecture, makes the architecture seem promising, but what more do MSPs need to know to guarantee success?

Scale Computing has a decade-long track record of innovation in hyperconvergence, storage architectures, cloud-based disaster recovery, ease-of-use, and most recently, product packaging and pricing that match MSP business models. Together these innovations have given hyperconverged systems the high-availability, scalability, versatility, and affordability MSPs need to become and remain competitive while increasing profits.

This white paper, intended for CEOs and VPs of Sales, Marketing and Business Development, describes how MSPs can support more applications for more customers with less investment than is possible with traditional datacenter and other hyperconverged architectures. The content is organized into three sections followed by a brief conclusion. The first section provides context by highlighting some of the challenges confronting MSPs. The second section describes how the design of Scale Computing's Hyperconverged Compute Cluster (HC3®) solution addresses these challenges to give MSPs a compelling and enduring competitive advantage. The third section identifies the most profitable opportunities for selling managed services and systems based on the HC3 solution.

## CHALLENGES CONFRONTING MSPs

MSPs are challenged to provide reliable IT services for a reasonable price. Using many traditional technologies, MSPs face complexity, high costs, and support challenges in combining those technologies into solutions their customers need. One of the most pervasive yet challenging technologies MSPs have deployed is the traditional 3-2-1 infrastructure.

### *Complexity*

The traditional 3-2-1 datacenter architecture consists of 3 or more servers connected by 2 network switches to 1 or more storage area network (SAN) or network attached storage (NAS) appliance(s). This layered architecture presents some serious challenges, the most obvious being considerable complexity in bringing together the multiple layers of individual components. Using best-in-class systems is normally a best practice, but having multiple hardware and software vendors makes it necessary for staff to be trained on and use a separate management system for each layer of the stack.

Virtualization creates what should be an additional and versatile layer of abstraction, but because the stack is built with layers of components from different vendors, the complexity of the integration is not optimized for performance. These solutions usually try to overcome the inefficiencies with extra RAM and SSD caching for storage performance improvements but this consumes extra system resources that should instead be available for running more VMs.

### *Single Point(s) of Failure*

The use of a SAN or NAS (the "1" in "3-2-1") creates another challenge: single point of failure. Eliminating all single points of failure in a SAN/NAS solution further increases complexity – and cost. And because SAN/NAS typically has a monolithic architecture, it can be difficult and expensive to scale for capacity and/or performance. Providing disaster recovery (DR) protection can also be expensive because it often requires "doubling down" on the infrastructure footprint by deploying a fully redundant configuration at a separate site.

Some converged infrastructure solutions attempt to mimic this 3-2-1 architecture. One such approach for converging compute and storage resources relies on virtual storage appliances (VSAs) that run as VMs. But because the approach is similar to the way SAN and NAS controllers function (a potentially worthy objective), VSAs usually suffer from the same resource utilization and performance problems that plague the 3-2-1 architecture (the reason why is covered in the next section).

### *Going Small*

Beyond building out their own internal infrastructure or building 3-2-1 infrastructure for customers, a further challenge faced by MSPs is delivering right-sized infrastructure for customer sites where a multi-node cluster may be too large for their needs and budget. These may be primary sites for smaller customers, ROBO sites, or edge sites. Being able to deliver turn-key infrastructure onsite that can be managed remotely with inherent replication back to the MSP datacenter for DR is more challenging with traditional infrastructure and hypervisors. These solutions often require additional levels of licensing and management tools for the MSP.

MSPs relying on traditional hypervisors and virtualization infrastructures confront these additional challenges:

- Differentiating MSP business services can be difficult when every MSP is using the same commoditized hypervisor solutions
- The enormous economies of scale in the cloud apply downward pressure on the pricing of managed services hosting and traditional on-prem infrastructure
- The complexity of traditional 3-2-1 virtualization solutions requires more highly-trained and certified professionals to maintain, increasing costs.
- Protecting data and applications from disaster and attacks like ransomware can be more challenging with traditional virtualization infrastructures that are more vulnerable and harder to recover due to complexity
- Creating scalable infrastructure that can provide only the resources (and cost) desired initially but that can grow rapidly with ease as it is needed

These challenges erode the profit potential of managed services, giving MSPs a potentially existential reason to rethink the traditional 3-2-1 architecture, or poorly converged solutions based on that architecture, for something more suitable to managed services offerings.

## CONFRONTING THE CHALLENGES WITH HYPERCONVERGENCE

With no industry standards for hyperconverged systems, all are designed differently, even using different definitions of what constitutes “hyperconvergence”. When originally coined, the word meant the convergence of compute, storage and/or networking in a single solution that also included the hypervisor. So the “hyper” had real meaning, and not just the “hype” it is today with some solutions.

Scale Computing defines “hyperconvergence” as “the integration of the compute, storage and virtualization layers of infrastructure into a single solution architecture.” Scale Computing Hyperconverged Compute Cluster architecture also adds backup and disaster recovery capabilities to the compute, storage and virtualization layers, resulting in a fully converged, highly reliable “datacenter-in-a-box” solution. Such genuine hyperconvergence, implemented in a family of standalone and clustered appliances, dramatically simplifies IT infrastructure, making it easier to deploy and manage, and substantially reduces the total cost of ownership.

Scale Computing also made the design suitable for use in private, public and hybrid clouds. The latter two are particularly important to MSPs because they enable offering managed services that reside exclusively in the public cloud, as well as those in synergistic hybrid configurations with systems deployed both on customer premises and in MSP datacenters.

The remainder of this section describes some of the ways Scale Computing has advanced the state-of-the-art in hyperconvergence, with particular attention paid to those capabilities that address the challenges confronting MSPs.

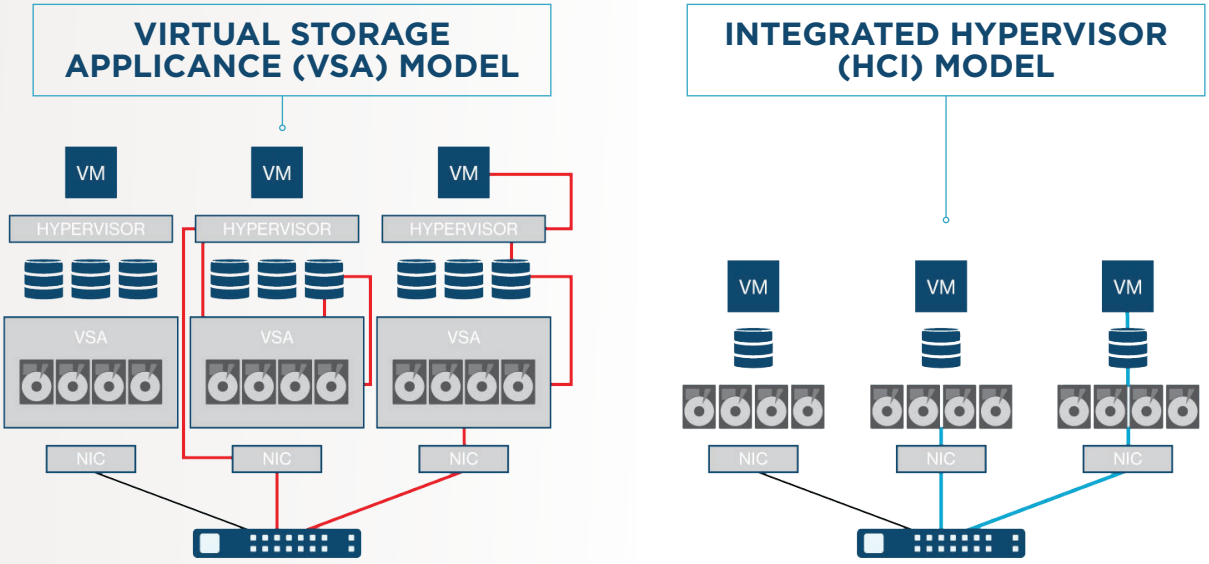
***Architectural efficiency*** forms the foundation of Scale Computing's HC3 solution. The HyperCore™ hypervisor is lightweight with minimal overhead, resulting in more memory being available for system and application software. HyperCore is a Type 1 bare metal hypervisor based on components of the open source KVM hypervisor. Kernel-based Virtual Machine has long been part of the Linux kernel and, as a result, KVM is field-proven in both small- and large-scale deployments.

While other operating platforms typically reserve and consume 24-32GB of memory, with some consuming up to 100GB, HyperCore only reserves 4GB. In a 64GB system, this results in 60GB or nearly 94% of memory being available for revenue-generating applications (vs. only 40GB or 62% for other platforms). In a 32GB system, the results are even more impressive with three times as much memory (28GB or 88%) being available for applications (vs. only 8GB or 25%).

***Streamlined, high-performance storage*** is provided by Scale Computing's innovative and patented SCRIBE (Scale Computing Reliable Independent Block Engine) storage layer. SCRIBE is a carrier-class, clustered, block-level storage layer that is purpose-built to be consumed directly by the HyperCore hypervisor. SCRIBE utilizes a wide-striped storage architecture that automatically discovers all storage resources, including both solid-state drives (SSDs) and spinning hard disk drives (HDDs). It then aggregates the total capacity available across the cluster and presents it to HyperCore as a managed pool of shared storage. With this design, all data written to the pool is immediately available for read and/or write access by every node in the HC3 cluster.

The benefits of SCRIBE derive from the intelligent pooling of storage blocks distributed redundantly across the entire cluster, which maximizes uptime and optimizes utilization. To assure high-availability by protecting against individual drive and node failures, the blocks are striped and replicated across all nodes in the cluster. And to efficiently utilize the storage pool's total capacity, the data is automatically deduplicated.

Performance is enhanced in a variety of ways, including SSD/HDD tiering, intelligent load balancing, and elimination of the inefficiencies inherent in SAN- and NAS-based storage solutions, as well as in Virtual Storage Appliances. HEAT (HyperCore Enhanced Automated Tiering) is a particularly powerful aspect of the SCRIBE storage layer. HEAT intelligently distributes blocks between the fast flash SSD tier and the slower, but less expensive HDD tier based on a heat map that tracks I/O operations for each virtual disk in the pool. The result is a cluster that delivers a competitive price/performance for the full spectrum of different customer needs. For applications where premium performance is warranted, the cluster's appliances can be configured with only SSDs. Conversely, for applications that do not require the high performance of SSD, the cluster's appliances can be configured with only HDDs.




***THE EFFICIENT DESIGN OF A FULLY INTEGRATED HYPERCONVERGED SYSTEM, LIKE SCALE COMPUTING HC3 ON THE RIGHT, ELIMINATES THE COMPLEXITY AND POOR PERFORMANCE FOUND IN DESIGNS THAT DO NOT INTEGRATE THE HYPERVISOR, LIKE THE ONE ON THE LEFT.***

***Built-in high-availability and self-healing,*** complete with disaster recovery (DR) protection, and file-level data backup and recovery, combine to make any managed service carrier-class. An HC3 cluster consisting of three or more nodes provides the native redundancy and resiliency needed to ensure high-availability. VMs can be replicated both locally and remotely, with the latter making it possible to provide DR protection for both the MSP cloud and systems deployed at customer premises.

The redundancy is layered and integral to the HC3 architecture, and includes dual network ports, redundant power supplies, and redundant striping across the cluster's storage pool. Built-in intelligence quickly and automatically recovers from node and drive failures by redistributing VMs and data across other nodes and drives, respectively. The system also automatically absorbs replacement nodes and drives.

Appliances can be deployed on customer premises in clusters with three or more nodes to provide local high-availability, or as "single-node clusters" with drive level redundancy to meet the needs of even the smallest sites and customers. In both configurations, VMs can be replicated to an HC3 cluster in the MSP's cloud as part of a DR-as-a-Service offering or to Scale Computing's Cloud Unity℠ service (covered in the next section). This enables MSPs to quickly and easily failover any VMs affected for protection against both site-level failures and more widespread disasters. Once service is restored, MSPs can just as easily failback the VMs to the customer premises.

Recoveries in all cases are fast and easy. Failures in any drives or nodes are handled automatically by the system. Recovering from site-level failures requires simply cloning VMs to a secondary site. The ability to revert to a prior snapshot or mount a prior virtual disk for file-level recovery is all included, as well. Recovery procedures require just a few clicks on the intuitive management interface, whether the problem is local to where the administrator is, or at a remote customer site or a distant MSP facility.

***Versatile scalability*** affords the ability to add capacity and/or increase performance incrementally and cost-effectively as needed. Scale Computing's efficient architectural design allows HC3 to run on small devices suitable to edge computing, ROBO, or small customer sites with as little as a single appliance or a cluster of small appliances for high availability and then scale out as needed. Other solutions that emulate 3-2-1 storage architectures have to start on biggers systems to support their resource overhead, making them less cost-effective.

Unlike most other clustered solutions, Scale Computing permits dissimilar appliances to coexist in a single cluster, and those appliances can be configured with the CPU, memory and storage resources needed. The more nodes that are added, the greater the aggregate performance and the larger the storage pool. Read/write I/O performance is further enhanced by wide striping across drives in the storage pool. RAID, by contrast, requires allocating drives to a specific RAID set, which limits the ability to scale both capacity and I/O throughput.

New appliances can be added in minutes without any downtime in any existing node in the cluster. The efficient design, combined with the compact appliance form factors, facilitate scaling HC3 clusters to meet the growing and changing needs within existing space, power and cooling constraints.

***Multi-cluster management*** of an MSP's entire distributed, hybrid infrastructure is available (by authorized personnel, of course) via an ordinary browser. The typical hybrid cloud consists of systems deployed both on customer premises and in the MSP's datacenter(s). The intuitive design of Scale Computing's Web-based management system enables IT generalists to configure, monitor, update, failover and failback, and otherwise operate the entire hybrid cloud infrastructure without any special training and with only a shallow learning curve. Multi-user administration supports an unlimited number of administrators, whether by shift or domain of responsibility, each of whom is assigned his/her unique account and login credentials.

An example of the simplicity of Scale Computing's management system is provided by the non-disruptive rolling system updates. A single click is all it takes to update an entire cluster's hypervisor, storage system, firmware and other system software. Once initiated, the automated update process systematically live migrates VMs among all nodes to facilitate updating every node in rolling fashion with no cluster-level downtime. Updating the entire cluster in a single, automated process has another advantage: It avoids problems that can be caused by potential incompatibilities among different versions of the software.


Creating and configuring VMs is just as easy. From a single screen, administrators can spin up a VM, allocate the desired resources, and upload the operating system and other software needed to run the application. Because all HC3 nodes have access to the cluster's entire storage pool, VM placement is determined automatically by the availability of compute (CPU and memory) resources, which also makes it easier to "right-size" existing VMs when needed.

It is important to note that making powerful management capabilities like these so easy to use is possible only with a hyperconverged architecture that integrates the full virtualization-server-storage stack.

## *A Better Bottom Line*

Scale Computing's innovation goes beyond the HC3 architecture to include how appliances are packaged and priced, which has an additional beneficial impact on MSP profitability. For MSPs, margins matter more than prices, but lower prices still matter because they lead to higher margins when using systems optimized for managed services. With an HC3 solution, MSPs can "do more with less" to maximize profits through a combination of increased revenues and reduced costs.

Here is a summary of the ways Scale Computing helps MSPs grow managed service offerings more profitably:

- No licensing fee or separate service fee for the HyperCore hypervisor
- No separate charges for the multi-cluster, multi-user management system
- Pay-as-you-grow pricing that becomes a "pass-through" to the customers
- Per-core pricing models that align better with MSP business models
- A choice of purchasing or leasing systems
- The ability to scale clusters and clustered resources incrementally, including by specifying the amount of CPU, memory and storage configured in new appliances
- The speed and simplicity of scaling clusters that eliminates the need for over-provisioning
- The ease of implementation and operation that eliminates the need for training, making it possible to staff 24x7 with IT generalists
- Built-in high-availability that minimizes downtime making it possible to offer competitive money-back service level agreements (SLAs) with confidence
- The ability to monetize high-availability and disaster recovery, creating even more opportunities for growing revenues
- Simple "out-of-the-box" deployment, remote manageability and dependable operation that combine to minimize the need for truck rolls to customer sites

## PROFITABLE OPPORTUNITIES FOR MSPs

Managed services based on Scale Computing's Hyperconverged Compute Cluster, complemented by standalone HC3 appliances deployed as "single-node clusters" on customer premises, can be employed in virtually all application use cases across all industry sectors. Among the more popular use cases in a hybrid cloud are:

- VMware Replacement – The high cost of VMware gives the license-free HyperCore hypervisor a compelling competitive advantage
- Cloud Migrations – Tools provided by both Scale Computing and its third-party partners make it easy to migrate customer workloads to public or hybrid clouds
- Virtual Desktop Infrastructure – Its lightweight architecture and efficient resource utilization make HC3 systems ideal for VDI, including at smaller offices
- Edge Computing – The compact form factors, self-healing operation and remote manageability make HC3 systems competitive in any edge computing application
- Hardware Refresh – This routine practice affords an ongoing opportunity to sell both systems and services
- Platform Consolidation – HyperCore's versatility enables streamlining workloads from different hypervisors or platforms onto a single, hyperconverged solution.
- Disaster Recovery – Leverage the flexibility of HC3 to protect workloads on multiple hypervisors and platforms.

**These three managed services, described in greater detail below, afford the greatest potential for peak profitability:**

- **Remote Monitoring & Management**
- **Disaster Recovery-as-a-Service**
- **Infrastructure-as-a-Service**

***Remote Monitoring & Management (RMM)*** appeals to customers that want to own or lease their own systems, but lack the personnel and/or expertise needed to manage them. RMM is normally a value-added service for MSPs that resell or (sub)lease the customer's HC3 appliances. Each appliance is normally pre-configured and shipped to the customer site for installation without a truck roll, and is then monitored and managed remotely from the MSP's facilities. Additional value-added services include migrating the customer workloads and Disaster Recovery-as-a-Service.

As mentioned previously, Scale Computing's innovative design enables HC3 appliances to be deployed in "single-node clusters." The smaller form factors are ideal for smaller sites, including remote office/branch office (ROBO) facilities of larger organizations. The smaller, as well as some of the larger appliances, are also suitable for many edge computing applications.
These capabilities make Scale Computing's HC3 solution especially competitive when selling and then remotely monitoring and managing customer premises equipment:

- Right-sized systems based on a choice of appliances and configurations make it possible to match the price to the performance needed at each site
- Quick and easy set-up makes it feasible to guide an untrained user by phone through the physical installation process
- Full remote monitoring and management takes over once the appliance is installed, eliminating the need to have any IT personnel on-site
- Self-healing storage keeps applications running should any drive fail, and automatically adds the replacement drive to the storage pool
- Optional Disaster Recovery-as-a-Service support in a hybrid cloud assures business continuity should the appliance itself experience a failure

***Disaster Recovery-as-a-Service (DRaaS)*** can be particularly profitable because customers are willing to pay a premium for the peace of mind it affords, and providing it in the cloud is relatively inexpensive. Whether the customer owns the HC3 appliance or leases it as part of a managed service, Scale Computing's built-in replication and data-protection features make DRaaS as dependable as it is profitable.

MSPs can also take advantage of Scale Computing's HC3 Cloud Unity service. HC3 Cloud Unity is implemented as a virtual HC3 appliance on the Google Cloud Platform, giving MSPs an alternative to deploying physical HC3 appliances for DR purposes, or hosting a dedicated cluster or single node. The service can be used to provide a full DRaaS offering for any HC3 system, whether deployed on customer premises or in MSP datacenters. HC3 Cloud Unity provides geographic diversity in offering DRaaS offerings, being able to protect workloads to a variety of data centers. And because Cloud Unity service runs the same HC3 software used on the appliances, it is managed exactly as the appliances are.

The HC3 Cloud Unity DRaaS is a full-service offering, complete with planning, initial and ongoing testing, data replication, continual monitoring, failover and failback, and other recovery assistance as needed during and after a disaster.

***Infrastructure-as-a-Service (IaaS)*** is the purely cloud-based model for offering managed services from HC3 clusters deployed in MSP datacenters. In the IaaS model, the HC3 cluster is owned or leased by the MSP, with its versatile and scalable resources shared virtually and securely among multiple customers.

With the HyperCore hypervisor and other aspects of Scale Computing's HC3 architecture making efficient use of all compute and storage resources, IaaS often serves as a core offering that can be value-added with other services, including consulting, migration assistance and DRaaS. The combination of IaaS and DRaaS gives the customer a complete and turnkey cloud experience, while giving the MSP a predictable and potentially lucrative revenue stream. HC3's built-in high availability and self-healing features ensure that there is continuity and availability in applications, workloads, or any other IaaS offerings that customers depend on.

## GETTING STARTED

The efficient architecture, the highly reliable high-performance storage pooling, the self-healing high-availability, the seamless and incremental scale-out, the non-disruptive updates, the ease of implementation and operation all combine to give Scale Computing's Hyperconverged Compute Cluster its industry-leading profit potential for providing managed services.

HC3 appliances have been used in both the smallest and the largest deployments of hyperconverged configurations in the world. The smallest are ideal for "single-node clusters" at customer sites, while the largest give MSPs the confidence needed to grow services without limits. The extensible HC3 architecture, which facilitates adding new and enhanced capabilities, combined with Scale Computing's proven track record of continuous innovation, provides MSPs the additional assurance of having a "future-proof" solution. And when needed, Scale Computing backs the entire HC3 product line with responsive, world-class support.

To help you get started and grow profitably, Scale Computing offers a variety of enablement and engagement resources. Enablement resources include sales and technical training materials, and regional technical, sales and channel personnel. Engagement resources include a Market Development Fund and other marketing programs, customizable marketing and sales materials, sales leads, sizing and configuration tools, and hundreds of customer success stories, at least one is certain to resonate with any prospect you encounter.

**To learn more about how your business can grow profitably and otherwise benefit from the HC3 solution, please visit Scale Computing on the Web at www.scalecomputing.com/partners/managed-service-provider-program, send an email to channel@scalecomputing.com or call 877-SCALE-59 (877-722-5359).**


**CORPORATE HEADQUARTERS**
525 S. Meridian, Suite 3E | Indianapolis, IN 46225
877-722-5359 | **www.scalecomputing.com**