DELL EMC FLEX FAMILY AND VXFLEX OS: THE SOFTWARE BEHIND THE SYSTEMS

Architecture Matters

ABSTRACT
This white paper explains the business challenges driving the need for a hyper-converged system that delivers the performance, resilience, and flexibility needed in the datacenter to support high performance databases and applications.

It describes how the Dell EMC Flex family systems and Dell EMC VxFlex OS are designed to meet these challenges in both large-scale environments and heterogeneous environments with multiple hypervisors and operating systems.

June, 2018
The information in this publication is provided “as is.” Dell EMC makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any Dell EMC software described in this publication requires an applicable software license.

Dell EMC, EMC², EMC, the EMC logo, ESRS, VxFlex OS are registered trademarks or trademarks of Dell EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners. © Copyright 2018 EMC Corporation. All rights reserved. Published in the USA. 06/18 white paper H15359

Dell EMC believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

EMC is now part of the Dell group of companies.
TABLE OF CONTENTS

REVISION HISTORY ........................................................................................................... 4
EXECUTIVE SUMMARY ...................................................................................................... 5
THE STATE OF INFRASTRUCTURE ....................................................................................... 5
Storage Area Networks (SAN) ............................................................................................. 5
Converged Infrastructure .................................................................................................... 6
moving to Software-Defined .............................................................................................. 6
Hyper-converged Infrastructure (HCI) ................................................................................ 6
VXFLEX OS: FLEXIBLE, SOFTWARE-DEFINED, AND MASSIVELY PARALLEL .............. 7
Configuration and Operational Flexibility ......................................................................... 7
Elasticity and Scale .............................................................................................................. 7
Resiliency and Availability ................................................................................................. 7
Performance ....................................................................................................................... 8
VXFLEX OS ARCHITECTURAL OVERVIEW ...................................................................... 9
Storage Data Servers (SDS) ................................................................................................. 9
Storage Data Clients (SDC) ................................................................................................ 9
Meta data Managers (MDM) ............................................................................................. 10
VXFLEX OS DEPLOYMENTS ............................................................................................... 10
Hyper-converged Deployments .......................................................................................... 10
Two-layer Deployments .................................................................................................... 10
VXFLEX OS ADVANTAGES OVER OTHER OFFERINGS ................................................. 11
No Performance Hot Spots ............................................................................................... 11
Provides Shareable Resource Pools Across ESX Clusters .................................................. 12
Shareable Resource Pools Across Hypervisor Clusters and Bare Metal .......................... 13
Operational Flexibility ....................................................................................................... 13
Enterprise Readiness and Resiliency ............................................................................... 13
Flexible Resource Pools .................................................................................................... 14
Manageable at Any Scale ................................................................................................. 14
Reduces Expenditures ...................................................................................................... 14
Granularly Scalable Performance .................................................................................... 15
Granularly Scalable Capacity ........................................................................................... 16
On-Demand Operational Agility ....................................................................................... 16
## Revision history

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Author</th>
<th>Change Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 2016</td>
<td>1.0</td>
<td>Jay Goldfinch</td>
<td>Initial Document</td>
</tr>
<tr>
<td>September 2016</td>
<td>1.0.1</td>
<td>Jason Brown</td>
<td>Formatting, trademarks, and Dell template</td>
</tr>
<tr>
<td>April 2018</td>
<td>2.0</td>
<td>Jay Goldfinch</td>
<td>VxFlex OS Update</td>
</tr>
<tr>
<td>June 2018</td>
<td>2.0.1</td>
<td>Brian Dean</td>
<td>Positioning changes</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY
Expectations of IT infrastructure have changed dramatically in recent years. Cloud computing has created an environment where platforms, applications, and infrastructure must provide a robust array of services at any scale.

These expectations dictate that no component can be a bottleneck to performance or manageability. Simplicity and manageability are required for every piece of the underlying infrastructure. Micromanagement of data storage infrastructure is neither acceptable to users, nor feasible to operations.

The Dell EMC Flex family of products, with Dell® EMC VxFlex OS®, are designed to meet these expectations. The Flex family currently includes the VxRack FLEX engineered systems and the VxFlex Ready Nodes. Each of these has management and automation tools to simplify the operation of every aspect, from hardware up through software layers. This paper will focus on the storage layer software inside these systems.

VxFlex OS (previously known as ScaleIO software) is a scale-out block storage service that enables customers to create a scale-out Server SAN or hyper-converged infrastructure on x86 server hardware. It is platform agnostic, resilient, flexible, elastic, universally consumable, and highly scalable. VxFlex OS is built in a way that solves many problems inherent in traditional SAN. VxFlex OS offers full-stack VMware® integration alongside support for other hypervisors and operating systems. It is a software-defined solution that offers a fully functional enterprise grade SAN equivalent using Dell EMC servers and intelligent software.

Using traditional SAN as a basis of comparison, this paper describes how customers that choose the Flex family systems and VxFlex OS meet operational and storage resource challenges. This paper also describes the challenges administrators and business owners face when using other software-defined offerings, and how VxFlex OS excels in contrast.

THE STATE OF INFRASTRUCTURE

STORAGE AREA NETWORKS (SAN)
Storage Area Networks (SAN) arose as an alternative to direct-attached storage (DAS) because of limitations inherent in the tight coupling between compute resources and their storage in DAS environments.

With DAS, the storage on the local host determined the amount of capacity and performance available to the application. That capacity could not be used by another machine. SAN addressed this by aggregating storage in a RAID-protected array and providing administrators the ability to partition out logical pieces of storage (LUNs) from the array to many compute nodes. Each of these compute nodes could be running a different application. SANs allowed shared access to LUNs from compute nodes, providing access to a LUN from a different node if the original node failed. SANs also aggregated performance and capacity, increasing both storage performance and overall storage capacity utilization.

This was an effective approach for many years. It met the requirement of separating physical storage from logical allocations of that storage, and the requirement that storage be accessible from multiple hosts. SAN has other advantages that persist today. Since SAN serves raw LUNs to compute nodes, storage presented by a SAN can be consumed by disparate clusters of hypervisors and bare metal deployments.

However, SAN brings with it operational problems. Overprovisioning is a common practice both to avoid performance problems and to keep up with expected business growth. But overprovisioned resources are wasted until applications need them. Business growth is not always predictable, and sometimes customers find themselves wasting capex never to realize return on their investment.

Even under the best circumstances, the process of migrating a workload from one storage array to another is inevitable as storage arrays reach the end of their usable life. Since storage arrays add capacity and performance in large chunks, this means forklift upgrades and additional rack space during the transition activity.

Operationally, the nature of SAN is that capacity and performance is unbalanced with demand. These problems exist even in circumstances of purely linear, predictable growth. When individual workloads are unpredictable, fluctuate, or grow massively, the siloed approach becomes even more challenging. Large and overprovisioned SAN arrays are not agile enough to cater to unpredictable and fluctuating business needs.
CONVERGED INFRASTRUCTURE

One of SAN’s biggest challenges is interoperability. Configuring a SAN, compute nodes, and the associated networks for a specific use case requires detailed product knowledge and conformity to complicated support matrices. As a result, vendors provide converged infrastructure offerings that are pre-tested with compute, networking and storage components as building blocks.

Converged infrastructure simplifies the process of acquiring IT infrastructure, and some vendors provide a single pane of management. However, many converged infrastructure offerings inherit the operational problems of their components. Converged infrastructure built using a SAN, for example, inherits the operational problems of SAN.

With converged infrastructure, the customer’s choice of vendors may be limited to those in the reference configuration. Converged infrastructure also requires accurate planning and sizing. Converged infrastructure is typically deployed in coarse grained “pods” tailored to a specific workload requirement.

Converged infrastructure is a huge advancement for overworked IT departments. However, many implementations sacrifice flexibility for manageability by glossing over fundamental problems with their components.

MOVING TO SOFTWARE-DEFINED

Business requirements, storage media, compute resources, and networking infrastructure are very different today than when SAN was conceived. Large capacity hard disk drives have resulted in lower cost for storage capacity but not necessarily lower cost per I/O operation. Large capacity drives have also resulted in very long RAID reconstruct times. Flash has introduced scaling considerations for CPU and memory. Advances in CPU and memory have enabled storage logic to run on compute nodes. Advances in Ethernet have widened the I/O channel to 40 gigabits or more, and enabled network latency in the tens of microseconds.

Because of these advancements in hardware, businesses have come to expect flexibility, scalability, and policy-based management of their data centers. This is delivered through software-based service-level abstractions for compute, storage and networking.

Ideally for many customers, the storage side of the puzzle is achieved without any of the underlying expensive storage or fabric infrastructure which is typical of SAN. These customers have begun adopting Server SAN which is built on standard x86 servers with directly attached storage as an alternative to traditional SAN. They expect the same level of full-stack interoperability and storage resource flexibility provided to them with SAN but with much lower cost and lower operational overhead.

HYPER-CONVERGED INFRASTRUCTURE (HCI)

As mentioned above, HCI combines compute, network, storage and hypervisor for ease of deployment and management. HCI eliminates impediments and simplifies workflows by delegating authority previously held by storage administrators to virtualization administrators. They eliminate the need for dedicated storage arrays and switches, using standard hardware instead. They also provide granular scalability, and reduce deployment time.

On the other hand, HCI offerings may trespass on organizational boundaries by ceding control of storage to virtualization administrators. Many companies have separate teams dedicated to managing compute and storage resources. These teams of professionals have distinct budgets and responsibilities, and the demands of existing infrastructure may require these teams to remain intact.

In addition to organizational considerations, the fact that many HCI offerings do not allow compute and storage to be decoupled limits options for independently scaling compute and storage resources. Since many hyper-converged implementations require that compute and storage be on the same set of nodes, scaling compute or storage alone is not possible.

The majority of HCI offerings eliminate the possibility of running applications on bare metal. Although there is no engineering limitation that ties software-defined storage to virtualized environments, many HCI offerings cannot run outside of a hypervisor. Support for bare metal deployments is increasingly important as developers move toward abstraction at the operating system level using containers.

Simply put, many HCI offerings do not carry with them the full value proposition provided by traditional SAN. SAN offers a pool of universally accessible storage. In contrast, these offerings tightly couple hypervisor implementations and clusters of hypervisors with storage capacity and performance. This creates islands of stranded and unutilized performance and capacity, re-introducing the same set of problems SAN was created to solve.

Software-defined storage is clearly a step in the right direction. It is a revolution in the storage industry, and is changing how storage will be deployed and managed in the future.
VxFLEX OS: FLEXIBLE, SOFTWARE-DEFINED, AND MASSIVELY PARALLEL

Consolidated and agile datacenters have requirements that neither traditional SAN nor many HCI offerings were designed for. As discussed in the previous section, offerings must scale in both capacity and performance, be flexible, elastic, and provide ease of management and integration. VxFlex OS was designed to meet these requirements through superior configurability and operational flexibility; elasticity and scale; resiliency and availability; manageability; and performance.

CONFIGURATION AND OPERATIONAL FLEXIBILITY

The common magic behind the systems in the Dell EMC Flex family is VxFlex OS, which is designed to solve the problems inherent to siloed approaches. As with SAN, VxFlex OS provides a universally consumable pool of storage capacity and performance that is independent of the hypervisors, hypervisor clusters, or operating systems using the storage.

Unlike SAN, however, VxFlex OS gives customers consistent, balanced load and utilization at scale. VxFlex OS provides maximum flexibility with minimal management overhead. VxFlex OS is designed to provide bursting capability to seasonal workloads and ephemeral workloads, and allows for enormous and unpredictable growth of any and all applications that use block storage.

VxFlex OS can run in a hyper-converged mode, where storage and compute is hosted on the same set of nodes, in a more traditional “two layer” mode that separates storage and compute infrastructure, or a combination of both.

VxFlex OS supports bare metal OS deployments, native hypervisor deployments, and deployments that combine the two. VxFlex OS natively allows multiple ESX clusters to share the same pool of shared storage, just as customers do today with SAN.

VxFlex OS is natively multi-tenant. And its distributed nature and Ethernet-based fabric make it ideal for containerized applications requiring persistent storage.

VxFlex OS supports any media combination, including spinning hard disk-only, SSD-only, PCIe flash, NVMe flash, hard disks boosted by flash media, and mixed clusters containing different media types for varying service levels. In some systems, asymmetric nodes are supported, where sets of nodes contain different numbers and types of disks, varying CPU types, different amounts of RAM, and different levels of network connectivity.

Partitioning storage resources is flexible, and can be tailored – even on the fly – to performance, capacity, and availability requirements.

ELASTICITY AND SCALE

A Flex family deployment can grow seamlessly from systems as small as three nodes to systems with 1,024 nodes. Expansion of capacity or performance is achieved by adding additional nodes. Likewise, nodes can be removed when capacity or performance requirements change.

VxFlex OS automatically rebalances the data layout as nodes and disks are added or removed. Scale-out and scale-in are fast, automatic operations that result in a perfectly balanced pool of distributed storage.

Rebalancing operations are fast because data layout is designed to keep data transfer to a minimum, and is parallelized among all members of the storage pool. Disks and nodes can be rapidly moved between pools of distributed storage.

It is important to note that the elasticity with VxFlex OS can be as fine grained as a single disk or node, or as coarse grained as dozens of nodes at a time. This allows customers to manage pools of storage infrastructure, rather than individual workloads or infrastructure components. It also eliminates operational impediments to workloads that grow or rapidly fluctuate.

RESILIENCY AND AVAILABILITY

Flex family systems with VxFlex OS can be configured to provide 6x9s or more of availability.

When a node or drive fails, VxFlex OS will begin to auto-heal within seconds. Rebuild operations, like scale-in and scale-out, are massively parallel, evenly balanced, and fast. When a node or drive goes offline then comes back online, VxFlex OS can rebuild in a backwards manner. A backwards rebuild leaves good, unchanged data in place on the restored drive or node, further reducing the rebuild time. Rebuild operations after planned hardware maintenance operations are optimized and faster still.

Redundancy in VxFlex OS can be location-aware. This prevents redundant data copies from residing in the same location inside a datacenter. VxFlex OS can continue to serve data after the loss of an entire rack or the loss of multiple nodes in the same enclosure.
VxFlex OS supports rolling, automated non-disruptive upgrades, simplifying the upgrade process, and reducing the likelihood of human error. It also provides advanced snapshot functionality and groups of snapshots synchronized inside the storage cluster.

**PERFORMANCE**

Flex family systems offer supreme I/O performance with a very small footprint. VxFlex OS is able to deliver tens of millions of IOPS, high bandwidth, and predictable low latency (sub millisecond latencies with flash configurations).

With VxFlex OS, data is striped across large pools of disks. Pools of disks may be grown or shrunk, without disruption, depending on the performance and capacity requirements of the applications. Because of this architectural principle, all data volumes created with VxFlex OS enjoy the benefit of all the devices in the storage pool, and there is no unpredictable performance variance among the volumes. Wide-striping of data sets between server nodes also helps provide consistent utilization of storage media capacity and performance by eliminating bottlenecks and hot spots.

Performance scalability is near linear. Performance increases predictably and consistently when nodes or disks are added. Individual nodes and disks can be added granularly. Unlike most modern storage systems, VxFlex OS can maintain peak performance even when capacity utilizations of available storage exceed 90%.

VxFlex OS is extremely lightweight in terms of CPU and memory consumption, freeing memory and compute resources for user workloads when deployed as an HCI solution. This increases the efficiency of the overall infrastructure stack.

VxFlex OS performance capability allows customers to consolidate workloads, use standardized hardware resources, and prevent situations where storage capacity is not fully utilized because of performance constraints. Most importantly, applications owners are free from performance concerns and infrastructure owners are free from constant juggling of workloads in an effort to optimize performance and meet demands.

![Figure 1. A VxFlex OS traffic pattern. I/O, rebuild, and rebalance operations are massively distributed and parallel. This illustrates a two-layer VxFlex OS deployment. Hyper-converged deployments, rebuild, and rebalance operations would be represented as a complete graph, where all nodes are logically connected to all other nodes (not shown).](image)

**NETWORK PLANNING AND INTEGRATION**

A properly planned and implemented network is essential for VxFlex OS to optimally achieve the capabilities outlined above. In the case of VxRack FLEX systems, this is engineered into the system itself. Compare the integrated networking of VxRack FLEX to an appliance or build-it-yourself scenario. Many things need to be considered for this solution to work over time. Onsite resources; a network expert to design and plan for future growth; multiple IT teams for everything to work; complexity that often gets overlooked like bandwidth, ports, oversubscription rates, and maintaining the same performance over time. Those worries are removed with an engineered system with integrated switches – superior performance at scale without the worry.
VxRack FLEX uses a set of top of rack switches connecting the nodes to each other at 10GbE or 25GbE. This ensures plenty of bandwidth as VxFlex OS pools capacity and storage. As the VxRack FLEX is filled with enclosures and the need to move to a second rack is necessary, another set of switches is used. This delivers 40GbE for east-west connections across racks, ensuring plenty of bandwidth as system scales. The engineered system removes the need for complicated calculations that might have to be done with stand-alone software defined solutions because the integrated network has already been done.

In the case of VxFlex Ready Nodes, please refer to our Dell EMC VxFlex OS Networking Best Practices for guidance on building the network needed for optimal performance, scalability, and manageability.

VxFLEX OS ARCHITECTURAL OVERVIEW

VxFlex OS fundamentally consists of three types of software components: the Storage Data Server (SDS), the Storage Data Client (SDC), and the Meta Data manager (MDM).

STORAGE DATA SERVERS (SDS)

The Storage Data Server (SDS) consumes raw local storage in a node and serves it up as part of a VxFlex OS cluster. The SDS is the server-side software component – a server that takes part in serving data to other nodes has an SDS installed on it. A collection of SDSs form the VxFlex OS persistence layer.

SDSs maintain redundant copies of the user data, protect each other from hardware loss, and reconstruct data protection when hardware components fail. SDSs may leverage SSDs, PCIe based flash, spinning media, RAID controllers, available RAM, or a combination of the above.

SDSs may run natively on Windows or Linux, or as a virtual appliance on ESXi. A VxFlex OS cluster may have up to 1024 nodes, each running an SDS. In a minimal configuration, Each SDS requires only 500 megabytes of RAM.

SDS components can communicate directly with each other. They are fully meshed and optimized for rebuild, rebalance, and I/O parallelism. Data layout between SDS components is managed through storage pools, protection domains, and fault sets.

Client volumes used by the SDCs are placed inside a storage pool. Storage pools are used to logically aggregate types of storage media at drive-level granularity. Storage pools provide varying levels of storage service priced by capacity and performance.

Protection from node, device, and network connectivity failure is managed at node-level granularity through protection domains. Protection domains are groups of SDSs where replicas are maintained.

Fault sets allow large systems to tolerate multiple simultaneous failures by preventing redundant copies from residing in a single rack or chassis – whichever way the fault set is defined for a given system.

STORAGE DATA CLIENTS (SDC)

The Storage Data Client (SDC) allows an operating system or hypervisor to access data served by Vx Flex OS clusters. The SDC is a client-side software component that can run natively on Windows, Linux, or ESX. It is analogous to a software initiator, but it is optimized to use multiple networks and endpoints in parallel.

The SDC provides the operating system or hypervisor running it access to logical block devices called “volumes”. A volume is analogous to LUN in traditional SAN. Each logical block device provides raw storage for a database or a file system.

The SDC knows which Storage Data Server (SDS) endpoint to contact based on block locations in a volume. The SDC consumes distributed storage resources directly from other systems running Vx Flex OS. SDCs do not share a single protocol target or network end point with other SDCs. SDCs distribute load evenly and autonomously.

The SDC is extremely lightweight. SDC to SDS communication is inherently multi-pathed across SDS storage servers, in contrast to approaches like iSCSI, where multiple clients target a single protocol endpoint.

The SDC allows for shared volume access for uses such as clustering. The SDC does not require an iSCSI initiator, a fibre channel initiator, or an FCoE initiator. Each SDC requires only 50 megabytes of RAM. The SDC is optimized for simplicity, speed, and efficiency.
META DATA MANAGERS (MDM)

MDMs control the behavior of the VxFlex OS system. They determine and provide the mapping between clients and their data, keep track of the state of the system, and issue reconstruct directives to SDS components.

MDMs establish the notion of quorum in VxFlex OS. They are the only tightly-clustered component of VxFlex OS. They are authoritative, redundant, and highly available. They are not consulted during I/O operations or during SDS to SDS operations like rebuild and rebalance. When a hardware component fails, the MDM cluster will begin an auto-healing operation within seconds.

MDMs can be run alongside the SDS and/or the SDC components or on separate hardware, for better protection and network isolation. In a VxRack FLEX system, the MDMs live on multiple nodes (known as controllers) to ensure their uninterrupted and optimal operation.

VXFLEX OS DEPLOYMENTS

HYPER-CONVERGED DEPLOYMENTS

In hyper-converged deployments the Storage Data Clients (SDCs) and Storage Data Servers (SDSs) run on the same set of nodes. This maximizes hardware utilization and reduces infrastructure requirements.

In hyper-converged environments where the storage servers and storage clients run on the same physical nodes, the resource utilization efficiency of the overall stack (storage, CPU and memory) is extremely high. This approach works very well in environments where a single group manages both storage and compute.

Figure 2. A hyper-converged VxFlex OS deployment. All the nodes are both providers and consumers of storage.

TWO-LAYER DEPLOYMENTS

In many environments, traditional SAN is supported by storage team, and the compute resources are managed by another team. Flex family systems work well in these environments because of the unique way VxFlex OS can be deployed on the nodes. In a two-layer (also known as storage-only) VxFlex OS deployment, compute resources exist on one set of nodes and storage resources exist on another set of nodes.

A two-layer deployment provides operational autonomy between existing teams of application owners and storage professionals. Two-layer deployments also allow compute and storage resources to grow independently. In two-layer deployments, existing compute and storage teams can work together to build a large-scale, software defined infrastructure.
The RAM cache on each individual storage node may be leveraged concurrently. This distributed cache architecture provides extremely high performance in two-layer deployments.

Finally, deployments can be mixed between two-layer and hyper-converged. This allows storage and compute to scale independently, while retaining the full-stack efficiency benefits of a hyper-converged deployment. In contrast, some HCI offerings require that additional storage resources be accompanied by additional compute resources, whether additional compute is needed or not.

Because VxFlex OS is software-based, it is possible to convert a two-layer deployment to a hyper-converged deployment over time. Converting a two-layer deployment to a hyper-converged deployment is simply a matter of adding SDS components to existing compute nodes, and adding SDC components to existing nodes.

![Diagram](image)

Figure 3. A two-layer VxFlex OS deployment. The SDS software components that make up the persistence layer reside on a different set of nodes than the applications. The applications use SDC components to consume the storage.

**VxFLEX OS ADVANTAGES OVER OTHER OFFERINGS**

**NO PERFORMANCE HOT SPOTS**

The innovative wide-striping design leverages the performance of hundreds of devices resulting in maximal and predictable performance without performance variability or hot spots. This allows VxFlex OS to achieve maximum performance even under near-full capacity utilization.

All storage resources, including HDDs, SSDs, and read caches are leveraged concurrently by every volume in the storage pool. Storage resources and I/O requests are perfectly balanced to achieve the best utilization of all the hardware components in the storage pool. This eliminates unusual performance variance between individual volumes, simplifying operations and satisfying users.

VxFlex OS will have superior performance even in case of lower skew values (high randomness of workload). This is because I/Os can be serviced by hundreds of devices. This design characteristic results in very small variability in performance and no hot spots even in case of highly random workloads and large working set sizes.

VxFlex OS uses hardware evenly and efficiently. VxFlex OS eliminates concerns of erratic individual volume performance and minimizes the impact of workload skew. With VxFlex OS, hardware resource sharing is a design principle. VxFlex OS is well suited to performance-oriented applications and unpredictable I/O patterns.
As with SAN, VxFlex OS allows native sharing of storage resources between and across ESX clusters. With VxFlex OS this workload distribution is done automatically and seamlessly. There is no need to vMotion, Storage vMotion, or rebalance the nodes in any way, nor is additional management tooling required to perform these functions.

VxFlex OS eliminates silos, reduces wasted storage capacity, provides a single pool of free space, increases operational efficiency, and delivers increased ease of use. This can result in higher performance, because I/O is distributed among a larger shared pool of storage; lower TCO, because hardware is more effectively utilized; and lower opex, because management overhead is reduced.

With VxFlex OS, administrators can host multiple ESX clusters in the same pool of shared storage, while sharing and balancing across all of them. VxFlex OS enables customers to share storage to non-ESX platforms and applications (such as databases, containers, and bare metal operating systems) even if the storage infrastructure is built on top of ESX.

Figure 4. Multiple ESX clusters, each with its own siloed storage.
SHAREABLE RESOURCE POOLS ACROSS HYPERVERSOR CLUSTERS AND BARE METAL

VxFlex OS can provide shareable resource pools between different types of hypervisors and bare-metal compute resources. Proprietary and open source clusters of hypervisors can coexist on a VxFlex OS cluster. Clusters of hypervisors and groups of machines running on bare metal can access and contribute to the same underlying pool of storage resources.

VxFlex OS allows customers to maintain a single pool of storage, while leveraging all of its available free space and performance potential, without having to manage additional storage infrastructure that comes with hypervisor or platform specific storage.

OPERATIONAL FLEXIBILITY

With VxFlex OS, media types, two-layer configurations, and hyper-converged configurations can be mixed to optimize pools of resources. This allows infrastructure to align with application performance requirements.

Unlike SAN, VxFlex OS does not require infrastructure such as fibre channel switches, cables, or HBAs. VxFlex OS can use existing, commodity Ethernet infrastructure. Old and new hardware can be mixed over time. No flash cache is required, and media types can be mixed in a cluster. No special HBAs are needed.

VxFlex OS uses large wide stripes across storage servers and is very light-weight. This gives VxFlex OS the ability to leverage all available hardware resources.

ENTERPRISE READINESS AND RESILIENCY

In traditional SAN and many SDS offerings, reconstruction of data protection can be a lengthy process because of the amount of data to be rebuilt and the number of devices participating in the rebuild process. Because of the lengthy duration of the rebuild, there is concern of additional failures occurring while data protection is in a degraded state.

VxFlex OS avoids this problem by distributing data across a larger pool of storage media. When a storage device inside a VxFlex OS storage pool fails, VxFlex OS will swiftly re-protect any affected data. VxFlex OS’s natively balanced layout means that every storage device in a storage pool will have roughly equivalent levels of primary data, mirrored data, and free space. The wide-striping, mirrored mesh layout of VxFlex OS volumes means that the data will be striped across every storage device in the storage pool.

Since the loss of a device means that the data layout in a storage pool will change, every storage device in the storage pool will participate in the rebuild concurrently, directly contacting its peer SDS and transferring only what’s necessary to make sure that there is a mirrored copy of every block.

This results in a huge reduction in rebuild time, reduces exposure to additional failures during rebuild, and reduces the severity and duration of impact to client I/O.
FLEXIBLE RESOURCE POOLS
Some hyperconverged implementations prefer a copy of data to be local. This ties CPU to storage. This means that data resides in the same physical node as the compute resource that accesses it. This also means that performance is not predictable when a workload moves to another physical node due to a failover or planned migration. This is also a problem when a workload spans multiple nodes because an application does not fit in a single node due to a compute, memory or storage limitation.

In addition, localized architectures limit storage resources to those available on the local node. Adding more nodes in these environments does not help storage performance.

Dell EMC Flex family systems with VxFlex OS have a highly distributed data layout that maintains high performance regardless of the location of a compute resource. VM guest performance remains high throughput and low latency, even as operations such as vMotion occur. This is in contrast to approaches that prefer I/O to be local. These approaches lose performance when a compute resource is moved.

With VxFlex OS, all the user volumes that use the underlying pool of distributed storage enjoy the benefits rapidly as physical resources are added to the pool.

MANAGEABLE AT ANY SCALE
VxFlex OS scales to 1024 nodes per cluster. This provides an order of magnitude more scalability than the biggest external SANs, along every dimension. VxFlex OS can provide far more raw capacity, raw throughput, and IOPS than traditional SAN.

In addition to rich management interoperability discussed above, VxFlex OS has numerous manageability advantages.

- Unlike SAN, VxFlex OS self-heals and balances data layout by itself. In contrast, modern RAID implementations result in stranded storage assets and increasingly slow reconstruct times.
- Unlike SAN, VxFlex OS enables server administrators to manage the storage as “just another application”. This provides more time for existing staff to manage other infrastructure components.
- Unlike hyper-converged infrastructure offerings, VxFlex OS can reduce team silos that can result in risk caused by poor communication of needs between teams.
- Unlike SAN, VxFlex OS requires minimal operational requirements. Familiar compute and networking infrastructure is used, and complicated storage protocol configuration is not required.

Additionally, VxFlex OS:

- Provides roles that determine which parts of the system an administrator can configure. Roles include front end configurator, back end configurator, administrator and monitor.
- Supports snapshots that can be mounted and used as clones.
- Supports QoS limits by IOPS or raw throughput. QoS limits give service providers an opportunity for revenue, and businesses the opportunity to control bully workloads and noisy neighbors.
- Provides hardware awareness when the underlying disks and RAID-controller are known by the system.

REDUCES EXPENDITURES
CPU and memory utilization is high with many other hyperconverged products leaving fewer resources for applications.

VxFlex OS increases the efficiency of the overall infrastructure stack in part because it can be scaled quickly and evenly, but also because the VxFlex OS software components are extremely lightweight. A VxFlex OS SDC requires only a few megabytes of RAM, and very little CPU in order to operate.

VxFlex OS can reduce per-CPU licensing expenditures. There is no proprietary hardware infrastructure required. Client side hardware requirements are very small. This ensures optimal use of CPUs in systems where software is licensed by CPU.
Additionally, many consumers of storage are doubling up on compute resources like RAM and CPU, increasing capex and opex. Hardware is not fully utilized, resulting in additional power, cooling, and rack space requirements. Traditional monolithic SAN arrays often require large amounts of power and cooling regardless of their utilization.

In hyper-converged configurations Flex family systems with VxFlex OS are extremely efficient, reducing rack space, power, and cooling requirements. Performance capability and capacity over provisioning is not required.

**GRANULARLY SCALABLE PERFORMANCE**

Performance scalability is near linear. Because performance grows automatically as hardware is added to a storage pool, VxFlex OS allows administrators to manage classes of service levels, rather than infrastructure components.

The sheer amount of I/O VxFlex OS can handle effectively eliminates concern of increased latency at scale. Some business requirements grow slowly to this kind of demand, others grow rapidly or fluctuate. VxFlex OS will maintain high performance throughout the lifecycle.

When new nodes are added to an existing VxFlex OS system, the performance and capacity available to existing compute nodes is increased. VxFlex OS therefore provides an enormous operational runway long after the scalability of any storage array would have been reached.

In contrast, the performance of traditional storage arrays is often described using a “latency curve”. A latency curve is a graph that illustrates the relationship between the load on the array and the speed at which the array responds to read and write requests. The more load placed on the array, the slower the array gets, until the array becomes saturated and responds very slowly.

![Figure 6. A typical latency curve. The X axis represents the amount of I/O operations per second processed by the array. The Y axis represents the amount of time it takes for the array to respond. The array slows down as more work is demanded from it. The box represents the “knee of the curve” where price and performance are optimal.](image)

The point where the storage array is nearing its maximum performance is called the “knee” of the latency curve. This is the point where a storage array is serving as much data as possible, while still performing quickly. This is where storage hardware is most efficient, but it is also where workloads and devices need special attention to prevent performance problems.

VxFlex OS manages performance and capacity in a different way. Performance growth is managed in small chunks of infrastructure. There are no monolithic, underutilized pieces of infrastructure at any point on a VxFlex OS deployment timeline.
Figure 7. A set of latency curves representing the growth of a VxFlex OS storage pool. Automatic rebalancing of data across new nodes enables optimal performance and hardware utilization throughout the deployment timeline.

VxFlex OS allows operators to maintain service levels on any scale, without micromanaging the hardware components that provide the services. As administrators add hardware, data is redistributed to take advantage of the new performance and capacity. Growth from a mid-size deployment to a datacenter scale deployment is granular, and allows operators to maintain an exceptional price/performance ratio throughout the deployment lifecycle.

GRANULARLY SCALABLE CAPACITY
SAN requires that capacity grow in course-grained chunks. VxFlex OS allows you to buy only what capacity you need, when you need it. Fast, native rebalancing allows capacity growth to closely match capacity requirements.

Figure 8. A timeline showing an increase in capacity requirements (green). To meet these requirements using SAN (red), capacity is overprovisioned as each new array or storage shelf is added. To meet these requirements using VxFlex OS, capacity overprovisioning is nearly eliminated by managing growth at a fine granularity.

ON-DEMAND OPERATIONAL AGILITY
Unlike SAN, Flex family systems with VxFlex OS can scale down. Hardware components can be removed from a protection domain or storage pool just as easily as they can be added. Just as when components are added, removal of components triggers a rebalance operation inside the storage pool, preserving data protection and a balanced layout.

This allows VxFlex OS to adapt to seasonal and short-lived storage requirements, allows administrators to adapt to changing business requirements, and provides the freedom to use hardware resources where they are needed.
Replacement of large storage arrays is never an issue with VxFlex OS. Components can be added, removed, and replaced in a controlled, granular, and nondisruptive manner. Because of near linear and enormous scalability of performance and capacity, concerns in those areas can be eliminated quickly.

**CONCLUSION**

Dell EMC Flex family systems with VxFlex OS are the foundations of a scale-out block storage service that enable customers to create a scale-out Server SAN or hyper-converged infrastructure using x86 server hardware.

VxFlex OS is flexible, high performance, and interoperable. It exhibits balanced and predictable behavior, and allows for varying performance and capacity ratios, decouples the scalability of compute and storage resources, and can scale enormously and nondisruptively. Additions to an existing VxFlex OS cluster achieve a rapid return on investment without operationally burdensome procedures like cut overs.

VxFlex OS provides a completely distributed pool of storage capacity and performance. It can provide consistent IOPS and latency, eliminating hotspots, no matter the load.

Flex family systems with VxFlex OS are capable of empowering environments where compute and storage teams are separate as well as environments where a single team manages both compute and storage. This allows organizations to manage their transition to a fully hyper-converged environment on their own terms and time scale.