SNIA NSF | NETWORKING STORAGE

Networking Requirements for Hyperconverged Infrastructure (HCI)

Christine McMonigal, Intel
J Metz, Cisco Systems
Alex McDonald, NetApp

February 5, 2019

Today's Presenters

Christine McMonigal
Intel

J Metz
Cisco

Alex McDonald
NetApp

SNIA-At-A-Glance

SNIA NSF | NETWORKING STORAGE

185
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Some Ground Rules!

- This is a ***vendor-neutral*** *presentation*
  - Different vendors have different solutions, there are no recommendations expressed or implied
- This is a ***technology-neutral*** *presentation*
  - No value-comparison with other technologies is expressed or implied
- ***Nothing*** in this presentation supersedes your vendor best practices and recommendations!

Why This Presentation

Source: https://www.networkcomputing.com/data-centers/why-hyperconvergence-needs-networking

The weak point in this new hyperconverged world comes at the interconnect level. Hyperconvergence vendors assume that storage and compute are their playground and more nodes will be sold to satisfy requirements in the future. However, to interconnect these nodes they must rely on existing network infrastructure.

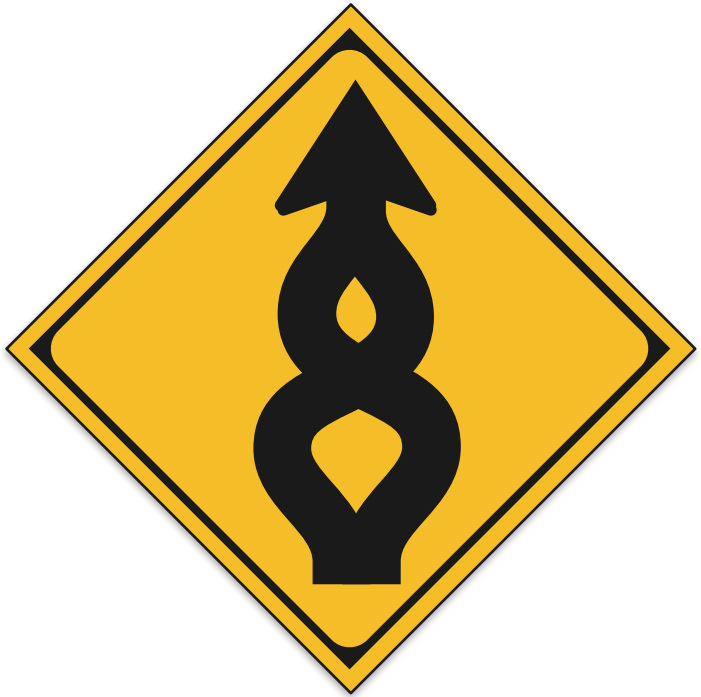
- Tom Hollingsworth, @networkingnerd

- Hyperconverged solutions can be software only, or hardware-and-software
- Focus is on compute and storage
- Networking is often overlooked, as it is assumed to always "be there"
- Things that storage and compute people get wrong about the network:*
  - The network is reliable
  - Latency is zero
  - Bandwidth is infinite
  - The network is secure
  - Topology doesn't change
  - There is one administrator
  - Transport cost is zero
  - The network is homogeneous

* Source: Gosling, James. 1997. The 8 fallacies of distributed computing."

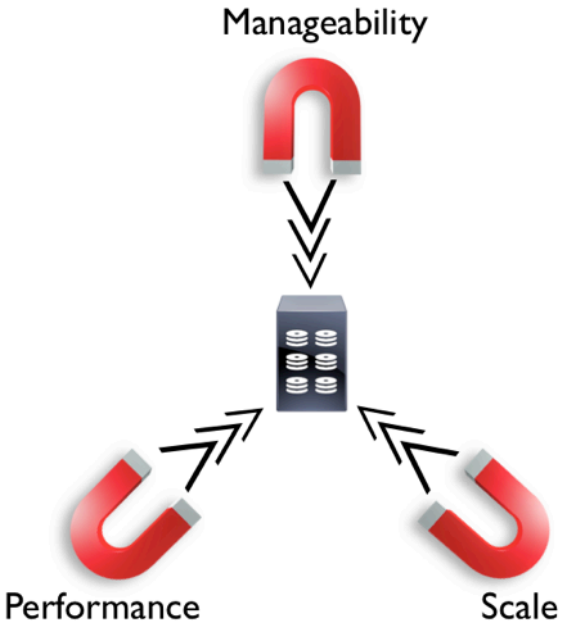
Agenda

- What is Hyperconvergence (HCI)?
  - What is it? SDS and HCI
  - Workloads, size of deployments
- HCI Storage Characteristics
  - Understanding HCI Reads and Writes
  - Other key considerations
- HCI Networking Considerations
  - Network topology options, speeds, classifying traffic flows
  - Sizing
  - Network settings and configurations
  - “Stretching” your cluster



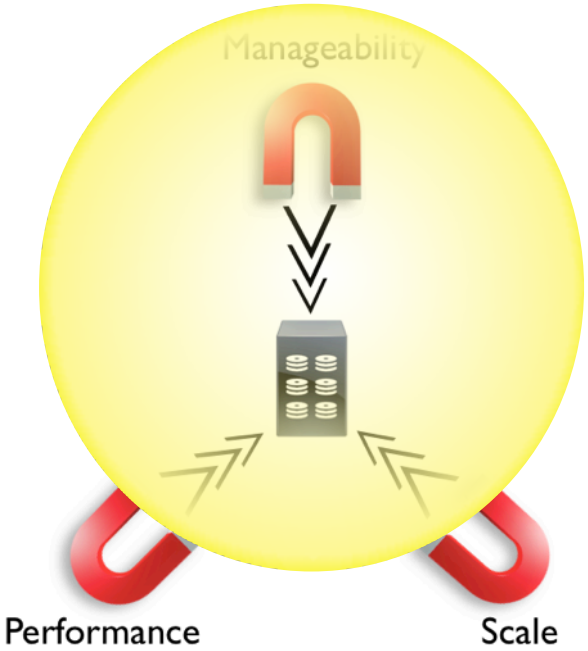
What Is Hyperconvergence?

Storage Trade-Offs



- There is a “sweet spot” for storage
  - Depends on the workload and application type
  - No “one-size fits all”
- Understanding “where” the solution fits is critical to understanding “how” to put it together

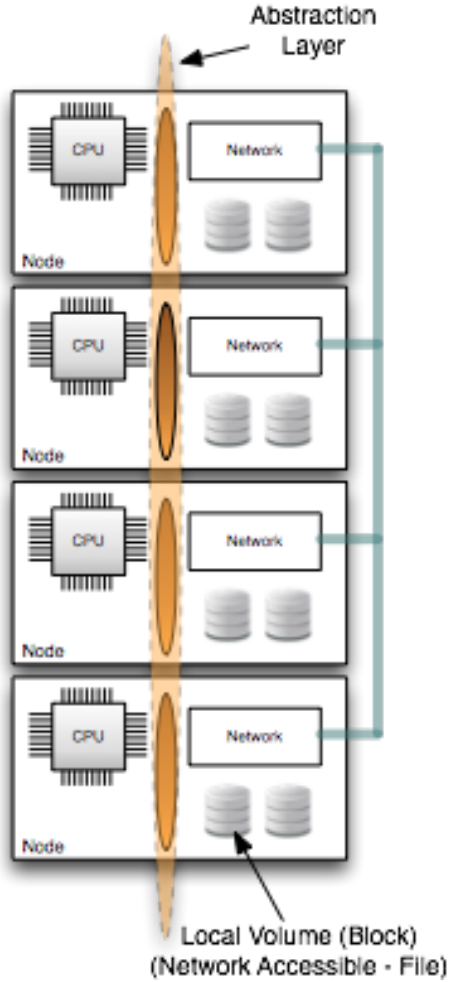
Storage Trade-Offs



- Hyperconvergence fits here
- Not just manageability - it's *uniform* manageability
  - All looks the same

Hyperconvergence - A Specialized Software-Defined Storage

- New infrastructure offering that is a type of Software Defined Storage (SDS)
- Tight Integration of standard servers for compute and storage, networking and virtualization, in an all-in-one appliance
- Integration of hypervisors and physical infrastructure
- Is it Block? File? Can vary depending upon vendor
  - Storage is presented via a distributed filesystem or object store
  - Block: Almost always iSCSI; File: Almost always NFS
- Has an abstraction layer for control plane management
- Magic Sauce: Each node can talk to each other node, centralized management, intuitive UI



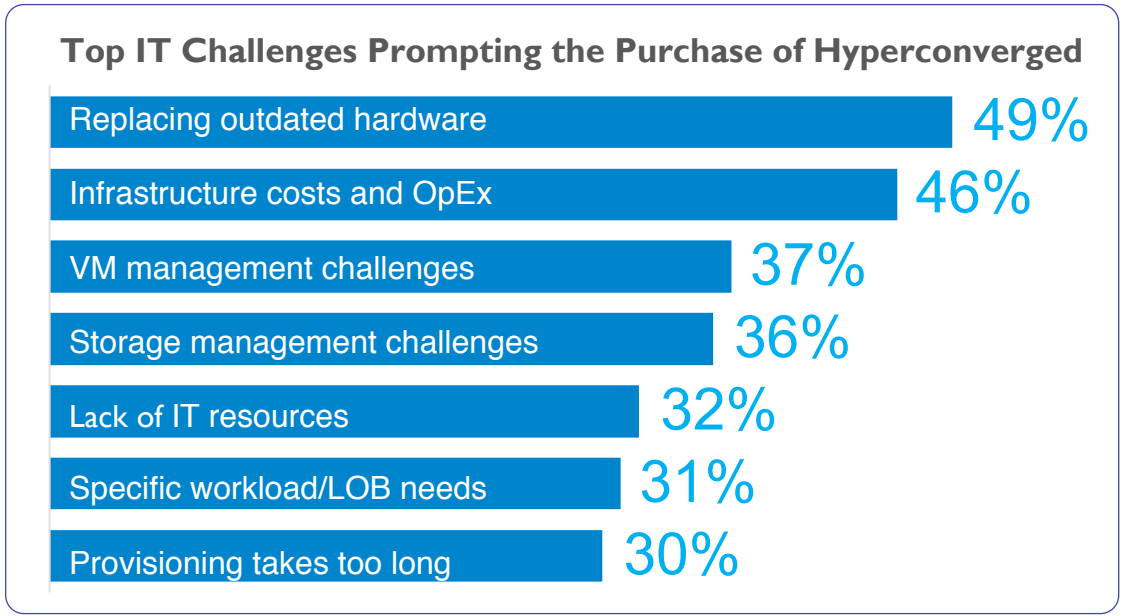
Hyperconvergence and IT Challenges

SNIA NSF | NETWORKING STORAGE

Hyperconverged Solution Adoption Rate among Surveyed Enterprises

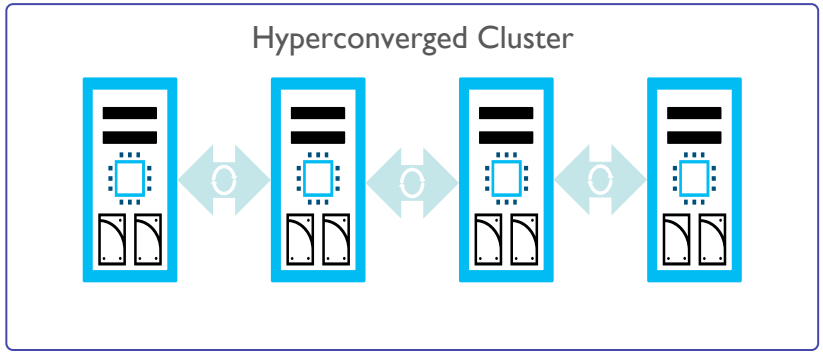
27% already adopted

21% plan to adopt in next year



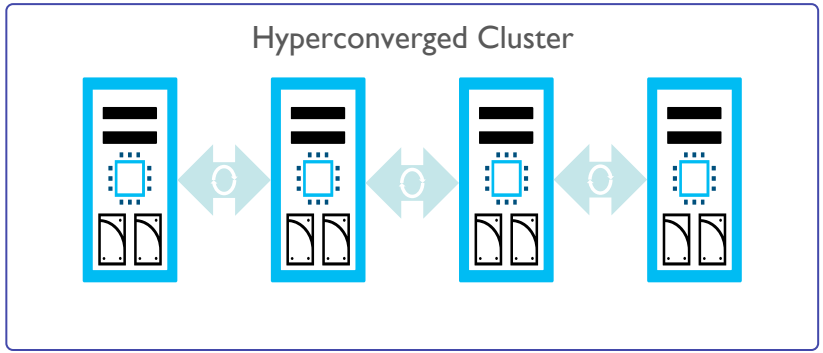
- 2H'18: Hyperconverged Platforms Customer Research, Technology Business Research, Inc.

Trade-Offs - Strengths



- Virtualizes compute and storage – pools the physical storage devices and compute capabilities
- Data distributed across cluster for durability
- Scales performance with capacity
- Supports multiple workloads
- Manage using familiar virtualization tools and resources; reduces “management sprawl”
- Rapid (-ish) deployment times
- Eases refresh budgeting

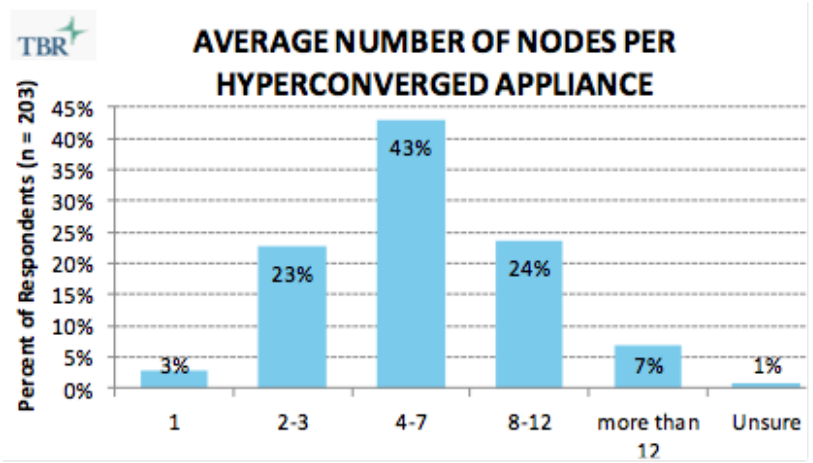
Trade-Offs - Challenges



- One vendor only - cannot mix-and-match software solutions from different vendors; no interoperability
- Hardware flexibility varies
- Balance between compute and storage can be difficult to achieve, especially in changing application environments
- Workloads can compete for resources and cause performance problems
- Some use cases may not be supported (e.g. Big Data, Hadoop, Spark, etc.)
- Growth of clusters places added burden on network

Cluster Size Impacts Network

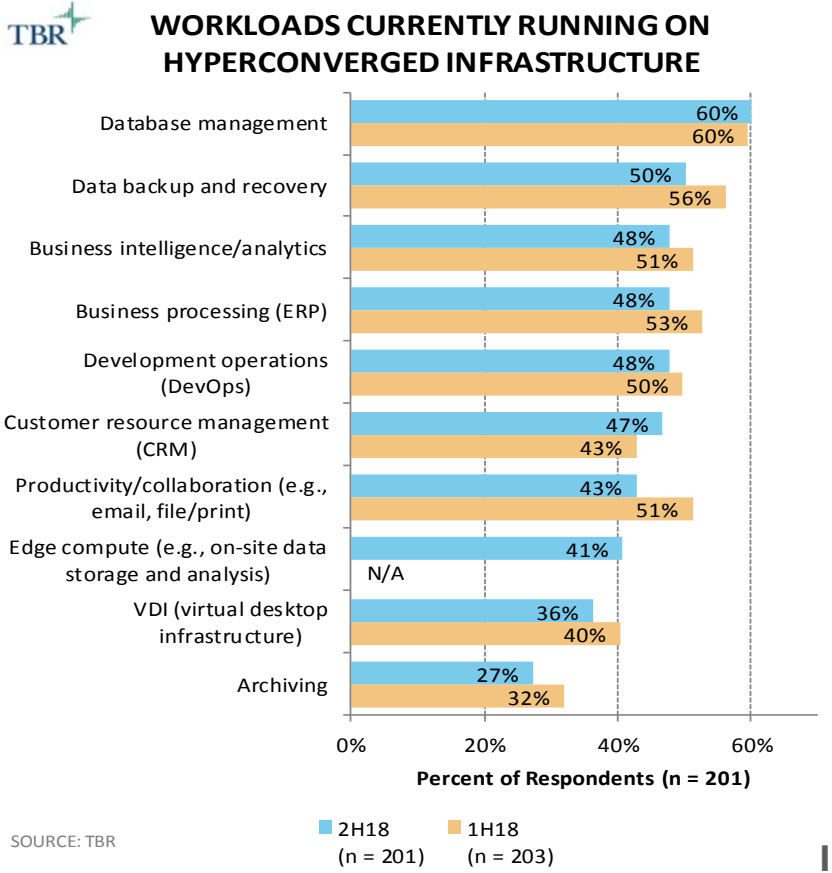
- Clusters are 3 or more server nodes
  - Can be as high as 128 nodes; maximum varies
  - Most deployments typically run between 4-12 nodes, but this varies widely
- Degree of scalability varies between HCI solutions
  - In general, the more nodes = more network demands
- Traffic between nodes places pressure on the network, from a control plane perspective



SOURCE: TBR HYPERCONVERGED PLATFORMS CUSTOMER RESEARCH, 1H'18

Which workloads are run on HCI?

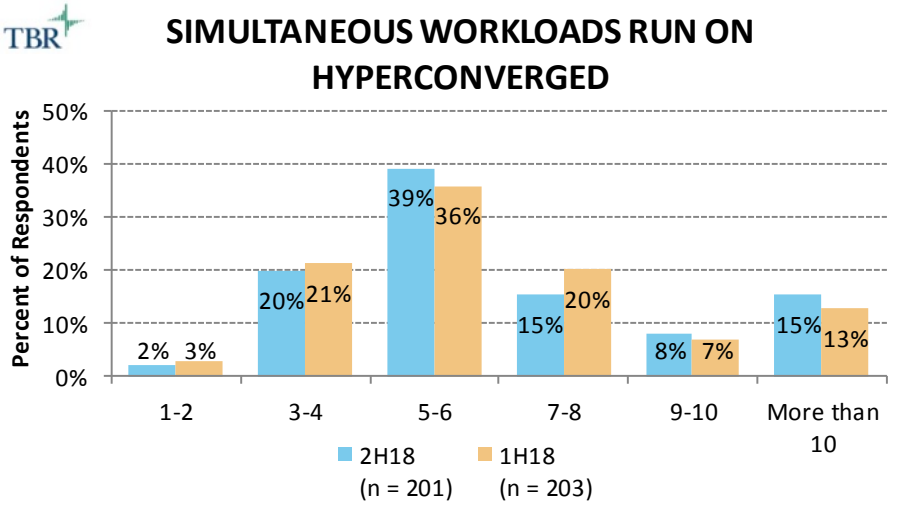
- HCI initially used for secondary workloads
- Now, primary and business critical workloads are most popular
- Networking bandwidth and latency expectations differ for these workloads



SOURCE: TBR

Mixed Workloads Prevalent

- Users run multiple workloads simultaneously on HCI clusters
- Data locality preferences vary between HCI solutions
  - Some HCI solutions try to co-locate data stored for specific workloads on the same servers where the applications are running
  - The greater the number of workloads running, the less likely data locality is achievable
- Most HCI solutions allow prioritization of workloads/apps

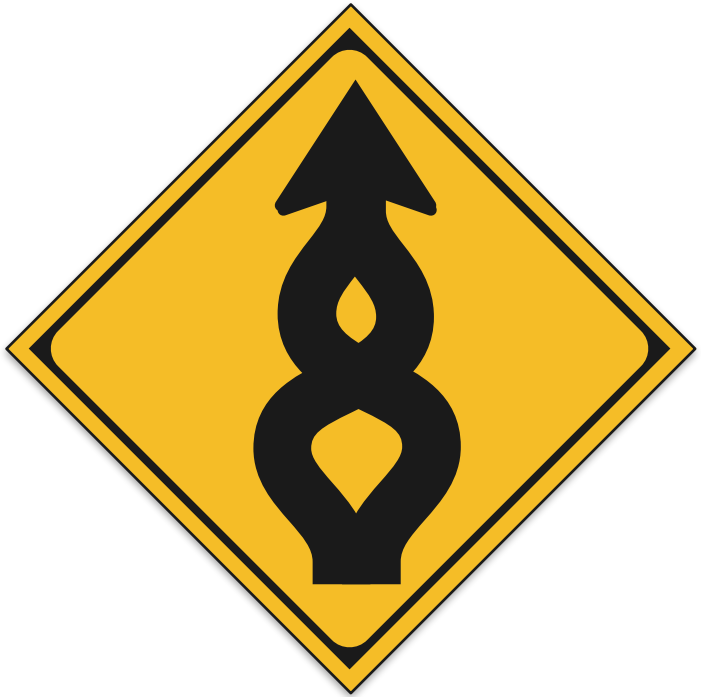


SOURCE: TBR

SOURCE: TBR HYPERCONVERGED PLATFORMS CUSTOMER RESEARCH, 2H'18

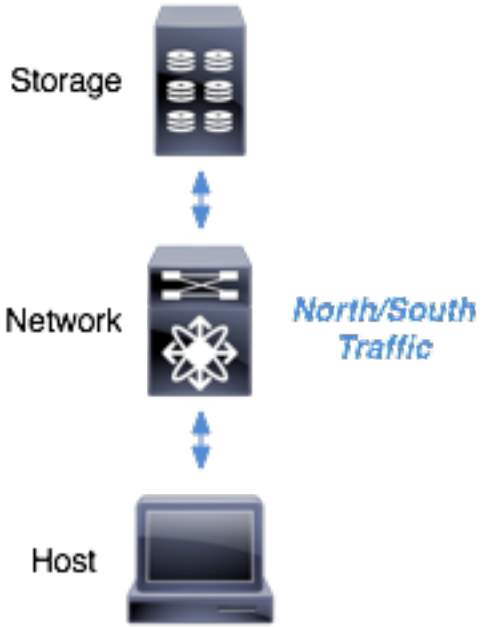
Agenda

- What is Hyperconvergence (HCI)?
  - What is it? SDS and HCI
  - Workloads, size of deployments
- HCI Storage Characteristics
  - Understanding HCI Reads and Writes
  - Other key considerations
- HCI Networking Considerations
  - Network topology options, speeds, classifying traffic flows
  - Sizing
  - Network settings and configurations
  - “Stretching” your cluster



HCI Storage Characteristics

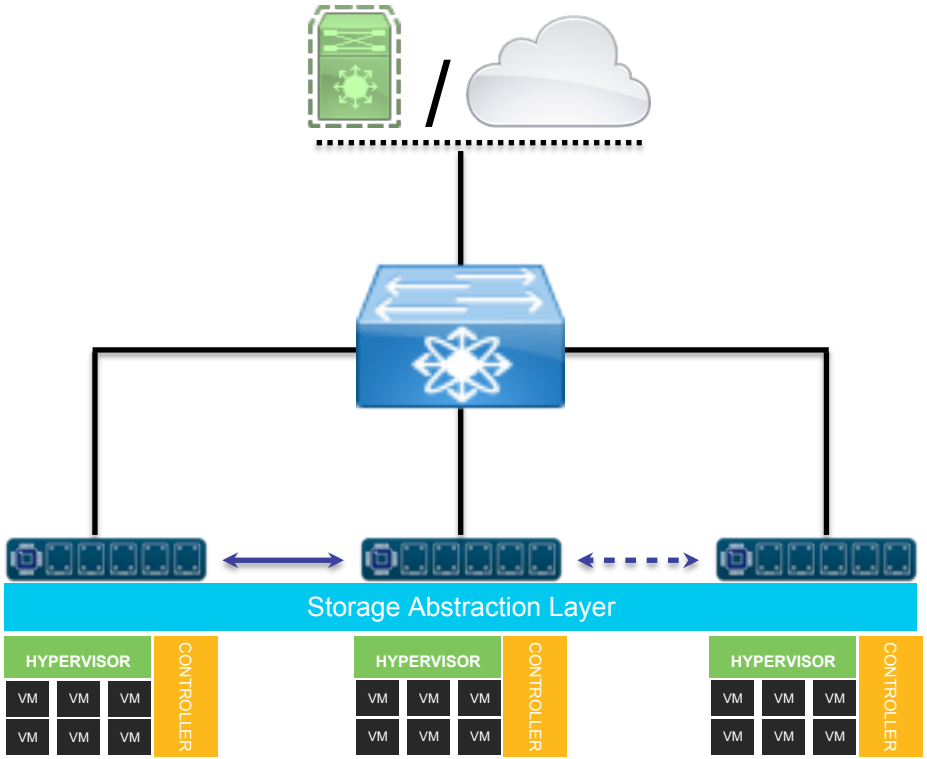
Understanding Networking Differences



Traditional Storage Environments

- "North-South" traffic from host to storage
- Storage is centralized
  - Storage services are maintained inside of storage array
  - Storage Network is tuned for storage traffic
- Management confined to each "layer" (compute, network, storage)
- High Availability managed at each level
  - Active/Active, Active/Passive multipathing
  - NSPOF component architectures for devices
  - Storage resilience maintained in Arrays/Filers
  - Network resilience maintained at L2-L4 or Fibre Channel
  - VM resilience maintained by hypervisor in hosts
  - etc.

Hyperconverged Storage Traffic



- Heavily increased “East-West Traffic”
- Demand for higher bandwidth and lower over-subscription ratios are common, especially for server-to-server communication
- Data is distributed to multiple nodes and load-balanced, creating additional traffic on the network
- Each additional node increases overall bandwidth requirements, especially for load-balancing algorithms

HCI Reads and Writes

- Goal is to make I/O transparent to the application
  - Application just fires I/O at virtualized disks
- Reads and writes require co-ordination between nodes
- Management software (sometimes an accelerator) on each node collects the I/O and decides what to do with it
  - Reads: looks up which node(s) holds and asks for it
    - Delay here isn't helpful
  - Writes: can be "coalesced", cached and subsequently passed to node(s) for writing
    - Optimizations can be performed (compression, deduplication)
    - Delay here can (within limits) be helpful
- Network traffic includes east-west node management "chatter"
- Not all I/O need end up on the network
  - Can require RAID on the node
  - Erasure Coding sharding requires I/O on the network

All Flash & NVMe Solutions

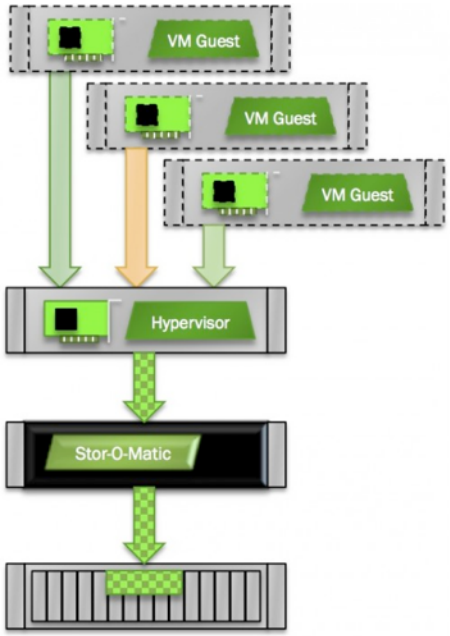
Beware the shifting bottleneck!

- HCI abstracts away the details of storage management
  - …but, you can only virtualize the actual HW underneath
  - Ensure HW selection will meet workload performance requirements
- All-flash and NVMe solutions deliver higher bandwidth, lower latency*
  - …but, the network will become a bottleneck unless users plan for it
  - Newer technologies such as RDMA, NVMe-over-Fabrics (NVMe-oF) and/or higher network speeds can help avoid bottlenecks
    - Watch for stressing the network with these new components

*See SNIA NSF "Under the Hood with NVMe-oF" https://www.brighttalk.com/webcast/663/175515

What is a “Noisy Neighbor?”

- Resource contention between virtual machines
- Different I/O requirements negatively impacting resources (compute/network/storage)
- Trade-offs: Larger Clusters versus Smaller, focused Clusters
- Mitigation:
  - Things like RDMA help
    - RDMA gets rid of the host buffering
    - Makes the network more responsible
    - Relieves the pressure on caching on the host memory and compute cycles, but pushes the burden into the network
  - However…
    - You’re throwing everything onto the network, so you have to make sure the network can cope with it



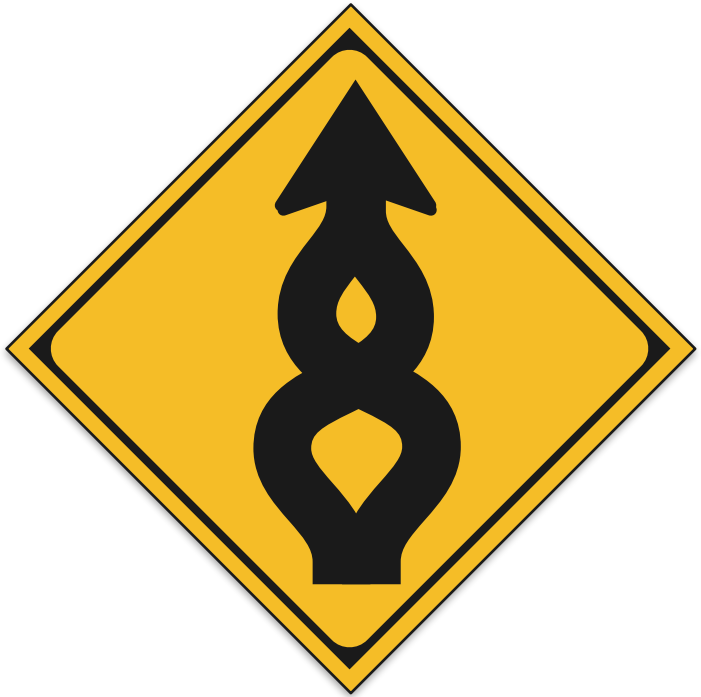
Source: Stephen Foskett. ”The I/O Blender”
https://blog.fosketts.net/2012/05/24/io-blender-part-2-virtualization/

Block or File?

- At this point, with modern kit, block and file don't matter at this scale
- Recall: these workloads are well-understood
- For vendors that offer either option, it becomes dealer's choice
- Performance is roughly the same, and generally not an issue
  - This is not where the performance issues reside

Agenda

- What is Hyperconvergence (HCI)?
  - What is it? SDS and HCI
  - Workloads, size of deployments
- HCI Storage Characteristics
  - Understanding HCI Reads and Writes
  - Other key considerations
- HCI Networking Considerations
  - Network topology options, speeds, classifying traffic flows
  - Sizing
  - Network settings and configurations
  - “Stretching” your cluster



HCI Networking Considerations

Network Quality

Worth Repeating

Worth Repeating

Worth Repeating

Worth Repeating

Worth Repeating

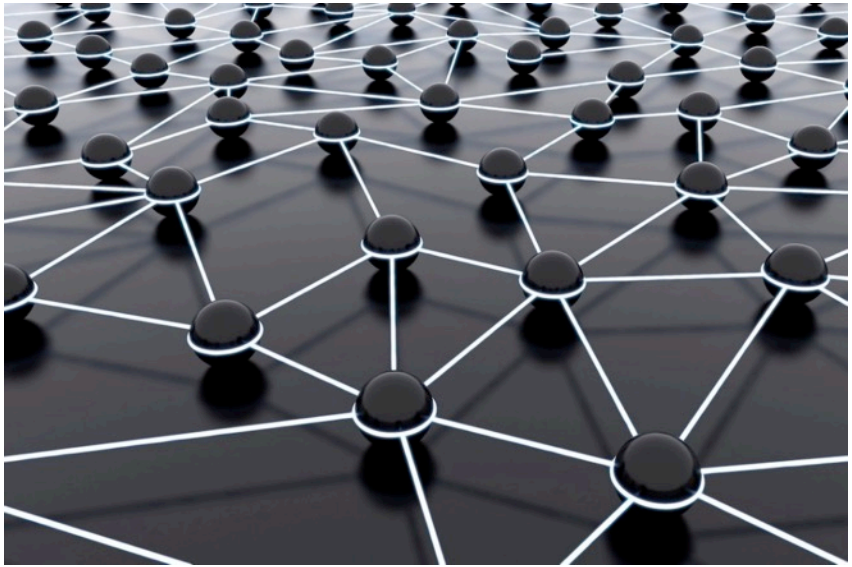
Worth Repeating

Worth Repeating

- The quality of service, reliability, availability, scalability, and overall performance of Ethernet fabric ultimately defines the capabilities of the HCI system
- Things that storage and compute people get wrong about the network:*
  - The network is reliable
  - Latency is zero
  - Bandwidth is infinite
  - The network is secure
  - Topology doesn't change
  - There is one administrator
  - Transport cost is zero
  - The network is homogeneous
  - Source: *Gosling, James. 1997. The 8 fallacies of distributed computing."

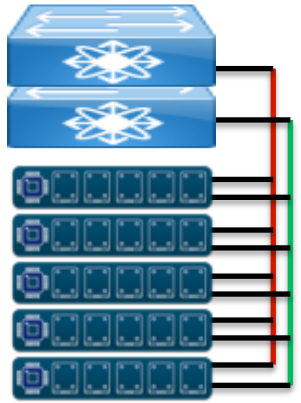
How Many Networks?

- Can be multiple:
  - Management Network
  - VM VLAN Network
  - Storage Network
  - Metadata Network
  - “Motioning” Network
- Different networks have different bandwidth and latency requirements
  - Remember: network bandwidth is measured in *bits*, not *bytes*
    - Easy math: Divide by 10 to get the bytes; 10Gb = 1GB throughput
  - Minimum: 10GbE (some have maximum of 25GbE) - read vendor documentation
- Storage interfaces are generally much faster than network interfaces
  - Consider NVMe impact on network

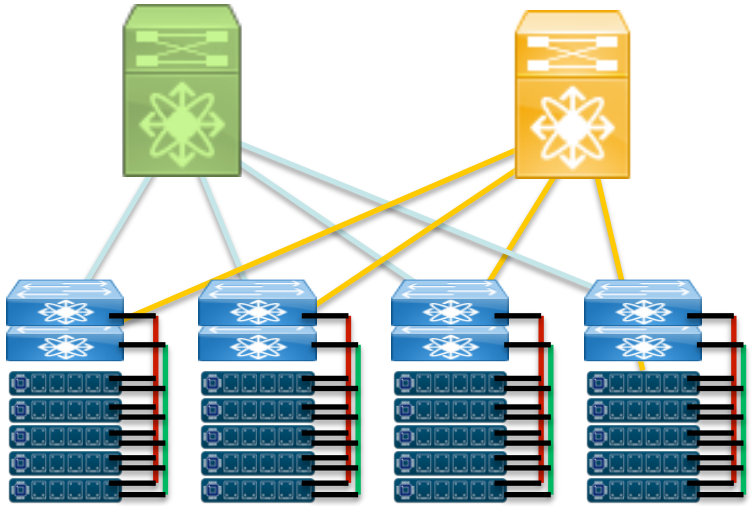


Small, Purpose-Built Cluster

- Reduce Noisy Neighbor Scenarios
- Fine-tune specific services for limited selection of applications
- Smaller fault domains
- Management interface is identical
- Not many networking implications (comparatively speaking) for small, purpose-built clusters



Large Mix-Use Cluster



- More protection and space efficiency
- More options for data protection (e.g., Erasure Coding or RAID levels)
- Better sharing efficiency
  - Reduction of impact on individual resources
- Increase in networking knowledge, management and QoS requirements

Vendors Want You To Consider

- Network Settings
  - Spanning Tree
  - Load Balancing
  - Jumbo Frames
- Link Aggregation/Bonding
  - These things you will want to take your vendor's advice
  - You want the simplicity - you should follow vendors' guidelines to the letter

“Stretching” A Cluster

- Some clusters may need multicast support
- In most cases, latency requirements are strict
  - Often 2-5ms maximum Round Trip Time (RTT)
  - Writes are replicated to the other site, which means that *each* write could take 5ms if that’s your RTT
    - Most solutions have SSD/NVMe caching, so your network will be the biggest problem
    - For every ms of latency, that’s a ms worth of cache
      - Expensive
    - Every ms is a *huge* amount of I/O… that you can’t do
- Because… Physics!

Summary

- HCI is a simple solution with complexity; complexity is under the covers
- Listen to the vendor
  - They've done the blood, sweat and tears
  - Understands what is required from the solution
- Understanding the principles, and what is going on, makes you a better end user

Very Special Thanks:
Duncan Epping
Cormac Hogan
Ivan Pepelnjak
Chris Twigg
Chris Dunk
Tom Hollingsworth
Stephen Foskett
Jase McCarty
Phil White

Our Next NSF Webcast

SNIA NSF | NETWORKING STORAGE

The Scale-Out File System Architecture Overview

February 28, 2019

https://www.brighttalk.com/webcast/663/346111

After This Webcast

SNIA NSF | NETWORKING STORAGE

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Networking Storage Forum (NSF) website and available on-demand at www.snia.org/library
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-NSF blog: sniansfblog.org
- Follow us on Twitter @SNIANSF

SNIA NSF | NETWORKING STORAGE

Thank You