Fibre Channel SAN Workloads

Live Webcast
February 12, 2020
10:00 AM PT

Today's Presenters

Mark Jones
Broadcom
Moderator

Nishant Lodha
Marvell
Presenter

Barry Maskas
HPE
Presenter

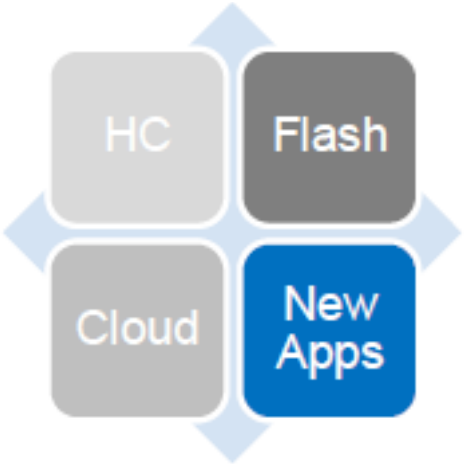
About the FCIA

- The Fibre Channel Industry Association (FCIA) is a mutual benefit, non-profit, international organization of manufacturers, system integrators, developers, vendors, and industry professionals, and end users
  - Promotes the advancement of Fibre Channel technologies and products that conform to the existing and emerging T11 standards
  - Maintains resources and supports activities to ensure multi-vendor interoperability for hardware, interconnection, and protocol solutions
  - Provides promotion and marketing of FC solutions, educational awareness campaigns, hosting public interoperability demonstrations, and fosters technology and standards conformance

https://fibrechannel.org/

Agenda

- Fibre Channel and Business Critical Applications
- Understanding FC SAN Application Workloads
- SAN Application Workloads I/O fingerprints
- How FC Delivers on Application Workloads



Key Tenants of Fibre Channel

- Purpose-built as network fabric for storage and standardized in 1994, Fibre Channel (FC) is a complete networking solution, defining both the physical network infrastructure and the data transport protocols. Features include:
  - **Lossless, congestion free systems**—A credit-based flow control system ensures delivery of data as fast as the destination buffer can receive, without dropping frames or losing data.
  - **Multiple upper-layer protocols**—Fibre Channel is transparent and autonomous to the protocol mapped over it, including SCSI, TCP/IP, ESCON, and NVMe.
  - **Multiple topologies**—Fibre Channel supports point-to-point (2 ports) and switched fabric (224 ports) topologies.
  - **Multiple speeds**—Products are available supporting 8GFC, 16GFC, and 32GFC today.
  - **Security**—Communication can be protected with access controls (port binding, zoning, and LUN masking), authentication, and encryption.
  - **Resiliency**—Fibre Channel supports end-to-end and device-to-device flow control, multi-pathing, routing, and other features that provide load balancing, the ability to scale, self-healing, and rolling upgrades.

How Fibre Channel Compares?

Storage Infrastructure Solutions: Their capabilities compared

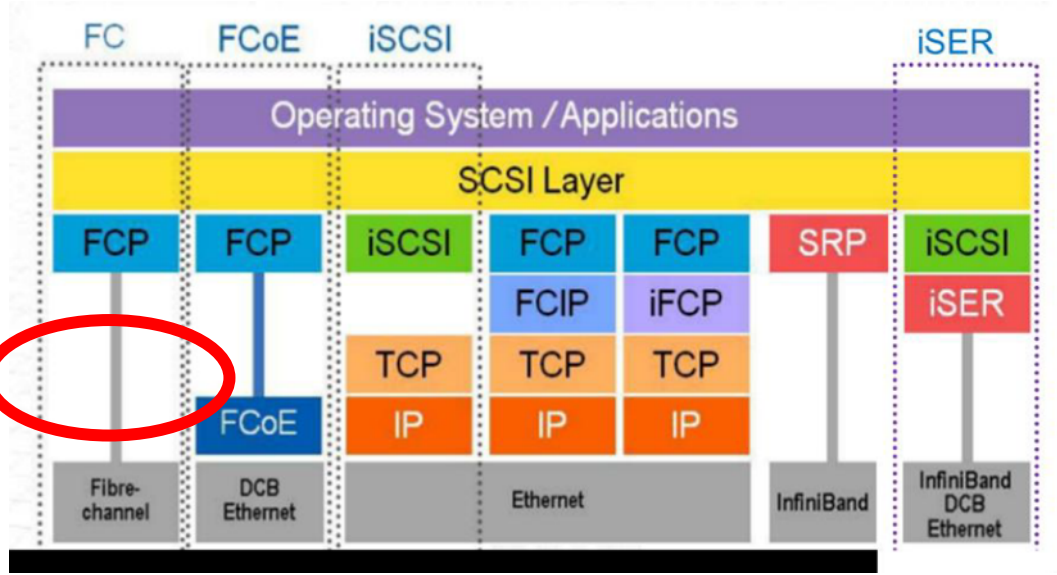
Critical Capability	FC Storage	IP Storage iSCSI/NAS	Converged Infrastructure	Hyper Converged Infrastructure	Software Defined Storage
Availability	●●●●	○○	○○○	○○	○○
Scalability	●●●●	○○○	○○○	○○	○○
Performance	●●●●	○○○	○○○	○○	○○
Agility	○○	○○	○○○	●●●●	○○
Extensibility	●●●●	○○○	○○	○○	○
Manageability	○○○	○○	○○○	●●●●	○
Security	●●●●	○○	○○○	○○	○○
Acquisition and ownership costs	○	○○○	○○	○○	●●●●

Source: Brocade.

FC: Low Overhead

- FC has low overhead in terms of protocol stack
- Enables FC to deliver low latency and low CPU Utilization per I/O

No Overhead!



Fibre Channel Workloads

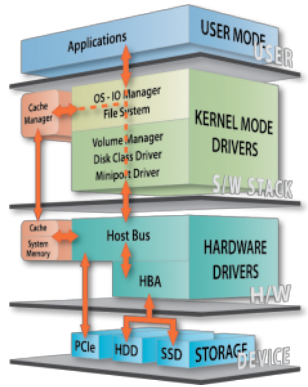
Market Drivers	Applications	Benefits
• Server virtualization • Increasing server workloads • Applications growth • Multi Core processors • NVMe • PCIe 4.0 • Security	• High-end backup • Disaster recovery • Enterprise Databases • Dense Virtualization • Big Data • Remote Replication	• Higher performance • Predictable performance • Reliability • Low Latency • Virtualization Aware • High Availability

Storage is a critical component of Enterprise Applications

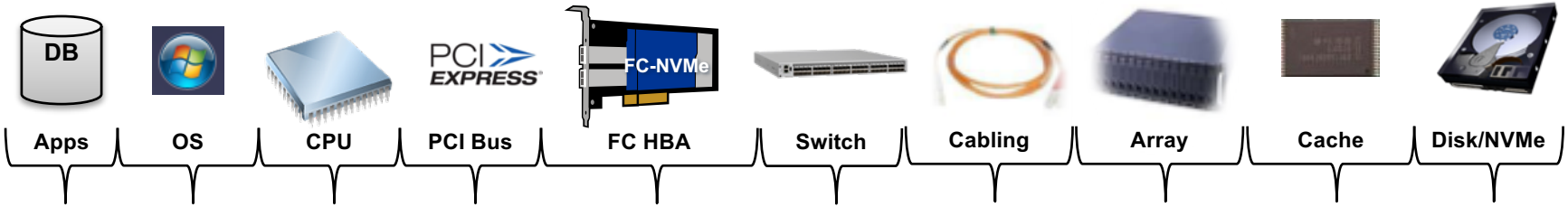
Why is Storage and I/O Important?

➔Business Critical Application expectations from Storage:

- **Performance**
  Uniform application response time under varying workloads
- **Reliability**
  Protection of your data from data loss
- **Availability**
  Data is available to the users



➔Imperative to understand that the full stack for Fibre Channel I/O:



Key Metrics for Measuring I/O?

Throughput

Bandwidth, Data Transfer Rate
Measured in MB/s
Sequential Large Block Workloads

IOPS

Transactional Performance
Measured in Kilo or Million
Random Small Block Workloads

Latency

Response Time
Measured in micro-seconds
Round Trip I/O Completion for sensitive workloads

Maximum Throughput or IOPS = latency continues to with constant throughput

Key Metrics for Specifying I/O?

- **Pattern**:
  - Sequential:
    - Data is read/written from the IO subsystem in the same order as it is stored on the IO subsystem.
  - Random:
    - Data is read/written from the IO subsystem in a different order as it is stored on the IO subsystem.
- **Size**:
  - Specifies size of I/O operations
  - 512Bytes to 1MBytes range
- **Access**:
  - Specifies Read, Write, or mix of both operations
- **Queue Depth:**
  - The number of outstanding I/O operations in flight

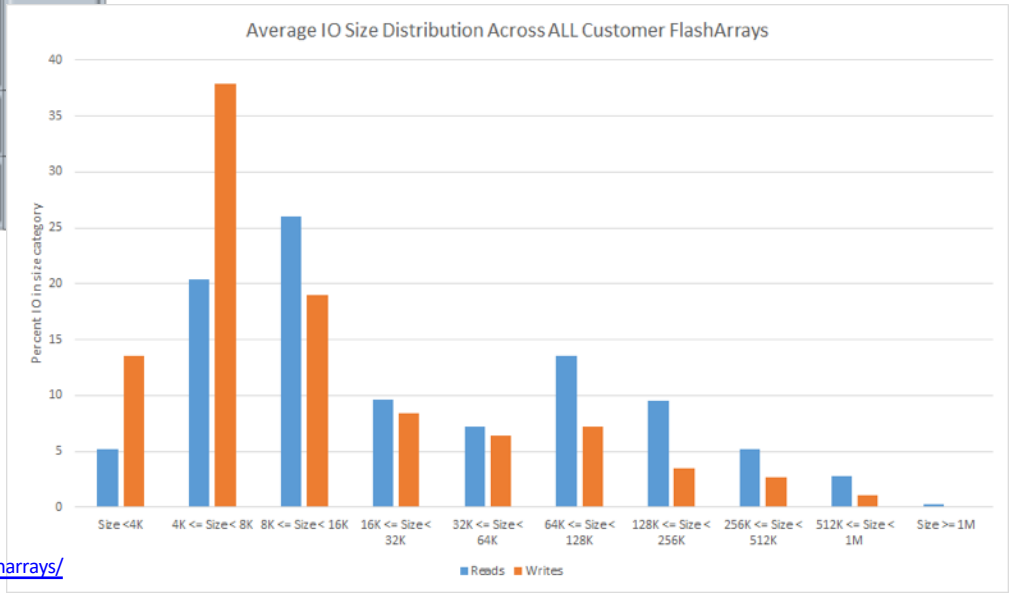
FC Workloads – Data Block Size Survey

Applications	4K	8K	16K	32K	64K	512K	1024K
Oracle Database	✓	✓	✓	✓	✓		
Microsoft SQL	✓	✓		✓	✓	✓	✓
MongoDB database		✓					
HPC for media, genomics, and life sciences	✓	✓			✓		
Microsoft Exchange				✓			
Data reduction		✓					

Source: Marvell survey

- FC applications utilize 4KB or larger block
  - 95% reads; 85%+ writes
  - Almost half (45%+) are 4KB & 8KB block sizes

- FC Workloads typically utilize 4KB or larger block size
  - 5 of 6 applications use 8KB block size
  - 512B micro benchmarks don't represent reality



Source: https://blog.purestorage.com/an-analysis-of-io-size-modalities-on-pure-storage-flasharrays/

I/O Fingerprints

Analytics, Business Intelligence, Data Warehouse, OLAP etc.

- Read-intensive, large block sizes
- Typical 64-256KB sequential reads (table and range scan)
- 128-256KB sequential writes (bulk load)

Transactional or OLTP Processing

- Read (70%) –Write (30%) -intensive, small block sizes
- Typically, heavy on 8KB random read / writes

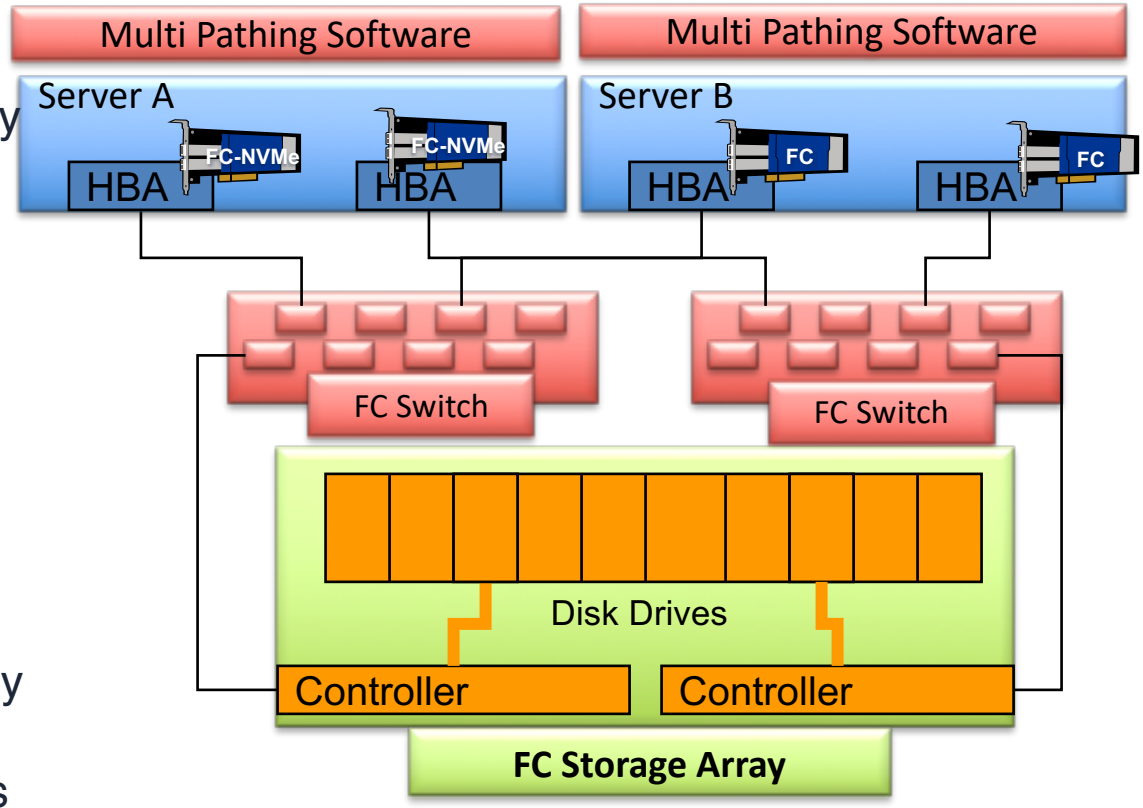
Virtualization and the I/O Blender Effect

- At the hypervisor and storage level
  - The I/O from multiple VMs gets mixed up – as if it were run through a blender
  - The storage system gets random I/O, even though it started out as sequential I/O per VM
  - Virtualization Services e.g. vMotion; HA / Fault Tolerant operation

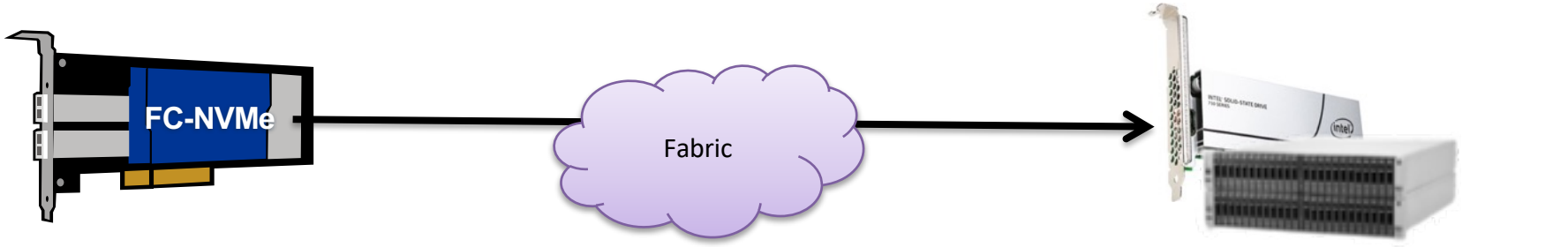
Bursty: Monday morning login problem with virtual desktops, background tasks etc.

Business Critical Apps - High Availability

- Business Critical applications need reliable storage
- In Fibre Channel this is typically achieved by:
- Servers
  - Multiple Host Bus Adapters
- Fibre Channel Switches
  - Two switches for redundancy
- Fibre Channel Storage Array
  - Two Controllers for redundancy
  - Multiple disk drives per array
  - Remote Replication between arrays across sites



New! FC-NVMe!



"NVMe" Over Fibre Channel

Non-Volatile Memory "Express"

Transport NVMe Natively over Fibre Channel

Low Latency

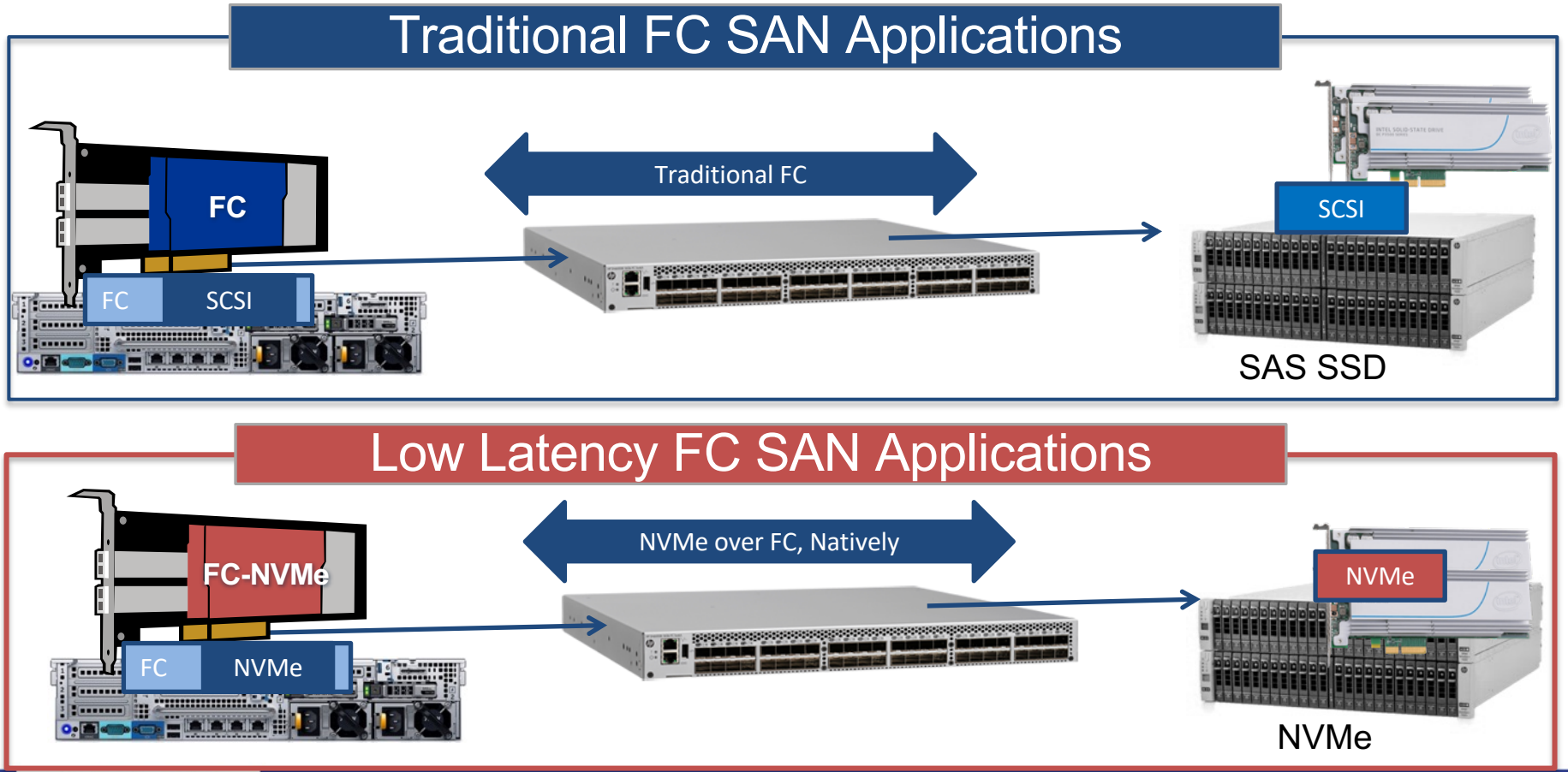
Reliable, Secure, Available

Leverage Existing Investments in Fibre Channel

FC-NVMe v2 near standardization

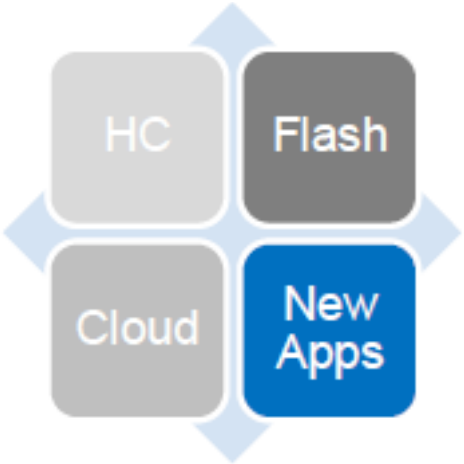
Ecosystem Ready

FC-NVMe – Delivers NVMe Natively



Agenda

- FC Delivers on Application Workloads
- The Stack and Protocol that enables these workloads
- Block Level storage virtualization
  - Multi Pathing, NPIV, Virtual SANs, VMIDs, zoning, B2B Credits, FCIP
- Upcoming standards that further enhance SANs



What is a Workload?

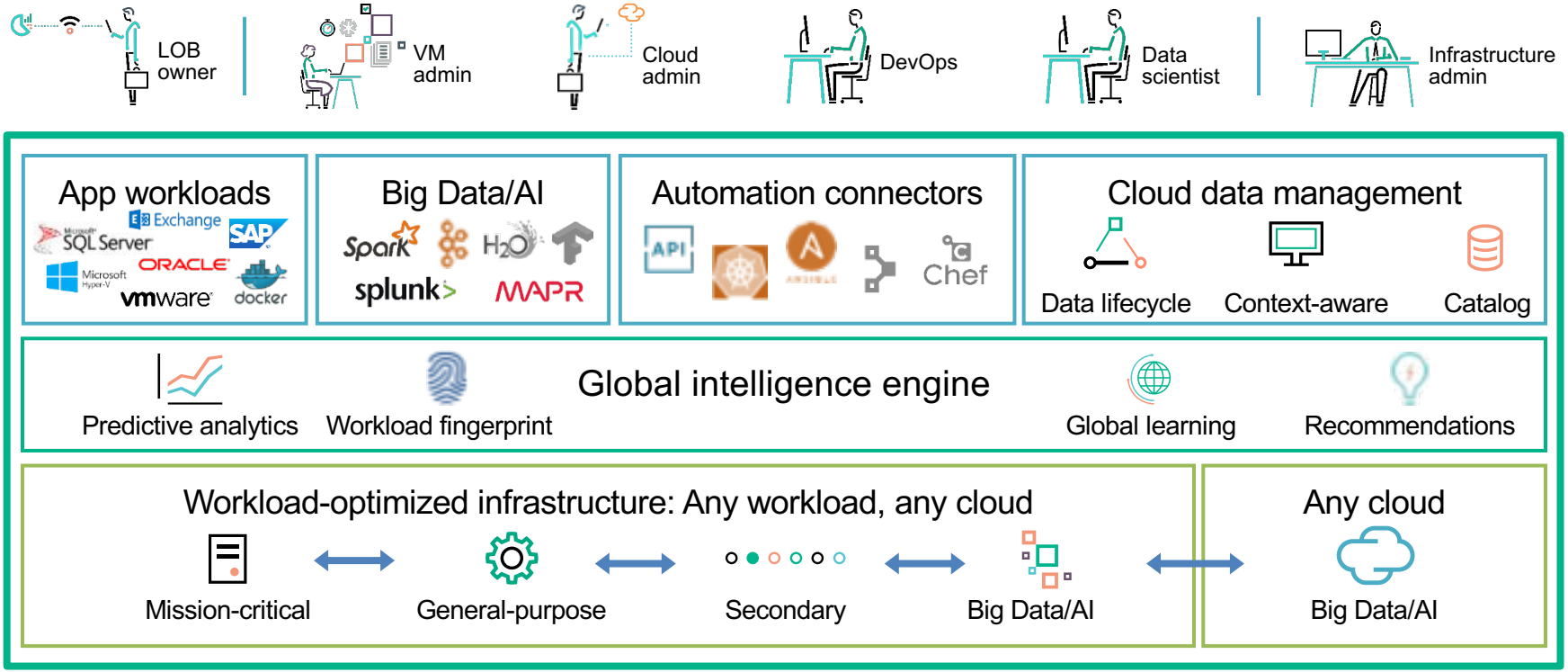
A workload is the combined I/O of the interfaces with network and storage infrastructures of a distributed application, often serviced by multiple servers

- For example, an application workload interacts with a web-server, one or several database servers as well as other application servers.
- The combination of all of these servers and the associated storage and network I/O makes up that application's workload.

It's all about the workloads and enabling a business as a whole.

- OLTP, DB requires random read/write performance & consistently low latency
- Healthcare, finance require resilient connections
- Virtualization requires both performance & resiliency

Workloads Platform



Many Different Software Stacks Running

Big Data - Open Source and/or Community Supported Workload:

IoT:

- Car (simulated): MiNiFi, CARLA, ROS
- Edge (simulated): Edgent, MiNiFi, Zenko, TensorRT+TensorFlow (GPU)
- Core: NiFi, Kafka

Real-Time:

- Streams: Spark Streams, Flink
- Persist: Druid (timeseries), Aerospike, Redis, MemcacheD
- Inference: Spark/Beam, Flink/Beam

Big Data:

- Storage: HDFS, HDFS-EC, Ceph S3, Scality S3
- Batch Processing: Spark, HBase

Deep Learning: PyTorch, Chainer, Caffe2, MXnet, H2O.ai SparklingWater and H2O.ai DeepWater

Storage Fabrics: Alluxio, Apache Ignite, WekaIO

GPU Data Frame: Apache Arrow

Monitoring & Alerting: Elasticsearch Stack, Prometheus, Graphite

Graphing: Grafana, Kibana

GPU Analytics: Knime, GPU/MapD

HPC: OpenMPI

Visualization: Superset, Presto, Hive

Proxy/Load Balancer: Nginx

CI/CD: Jenkins, Git, Garret

Orchestration & Automation: Kubernetes

Packaging: Ansible, Helm

Deployment: Kubespray, HashicorpTerraform

Virtualization: Hashicorp Vagrant and VirtualBox

Containerization: Docker

Installation: PXE, Kickstart, Hashicorp Cobbler

Scheduling: Airflow

HPC Cluster Management: Insight CMU

Security: Kerberos Support, SSL Support, RBAC Support

Key Management: HashiCorp Vault

Governance, Risk & Compliance (GRC): Eramba

The V's of BIG Data

Any one, or all, of the following "V's", and driven by a 5th "**V**" = ***V****alue*.

Businesses must be able to derive *value* from any investment they make in their data

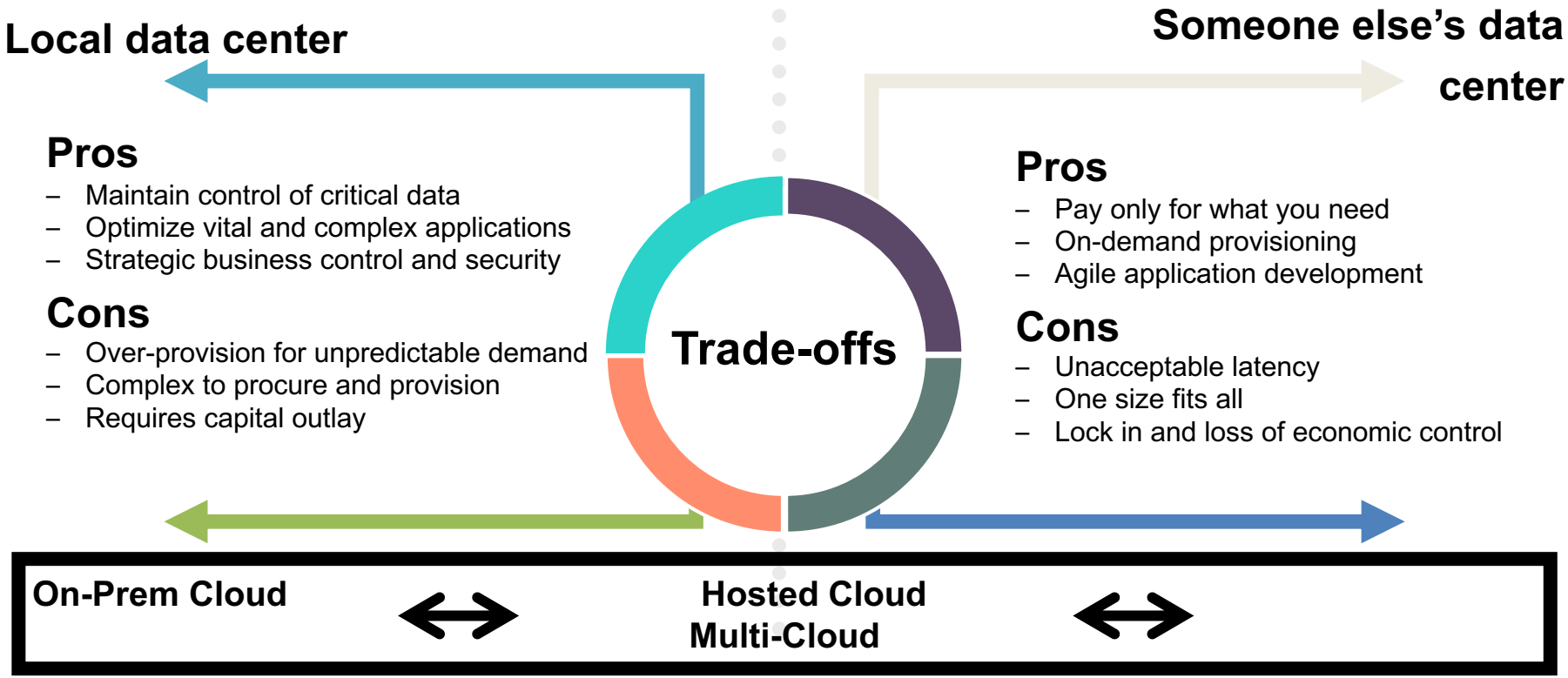
- ✓ this is done through Machine Learning

1. **Volume = Too much data to move around for processing (tens-of-petabytes, at a minimum)**
2. **Velocity = Processing Data in Motion**
3. **Variety = Lots of different data sources providing different kinds of data in different formats, not easily processed due to no universal format**
4. **Veracity = Difficulty deriving any intelligence or truthfulness of the data**

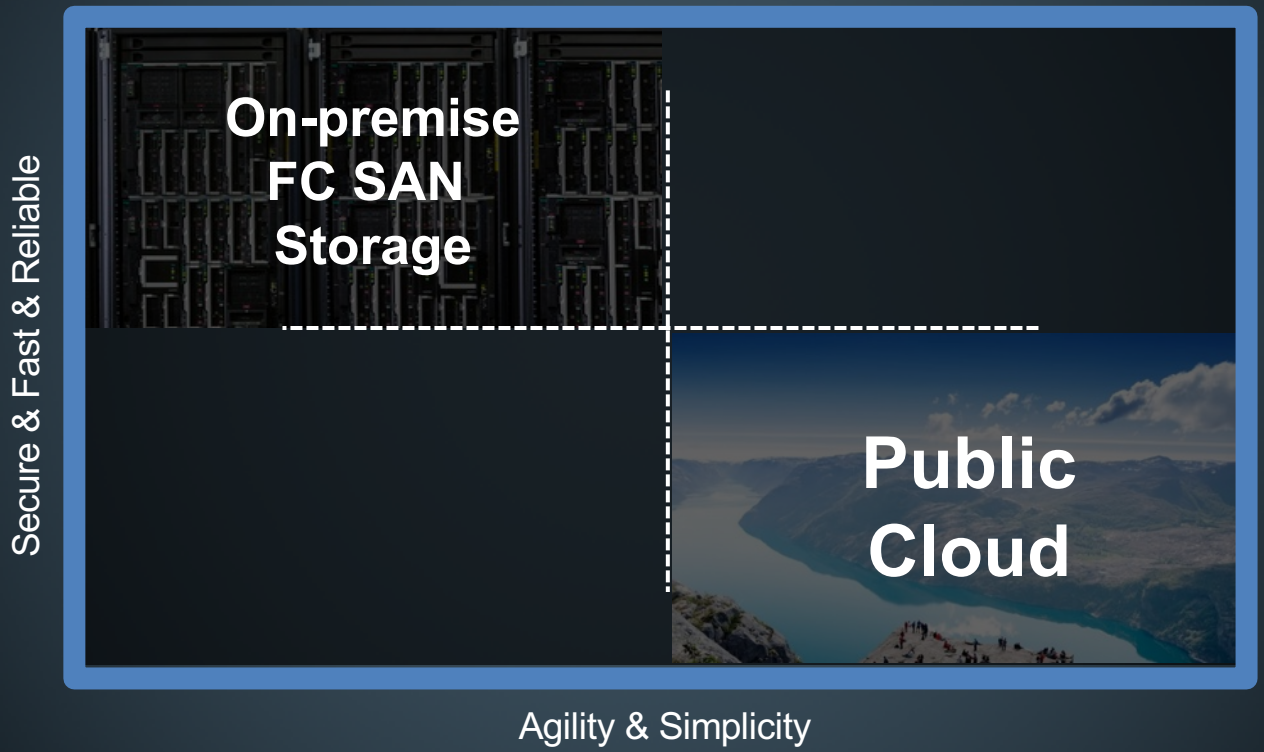
IoT - A subset of Big Data that focuses on all the above-V's

- ✓ specifically the processing of very high velocity – very small data – at extreme scale.

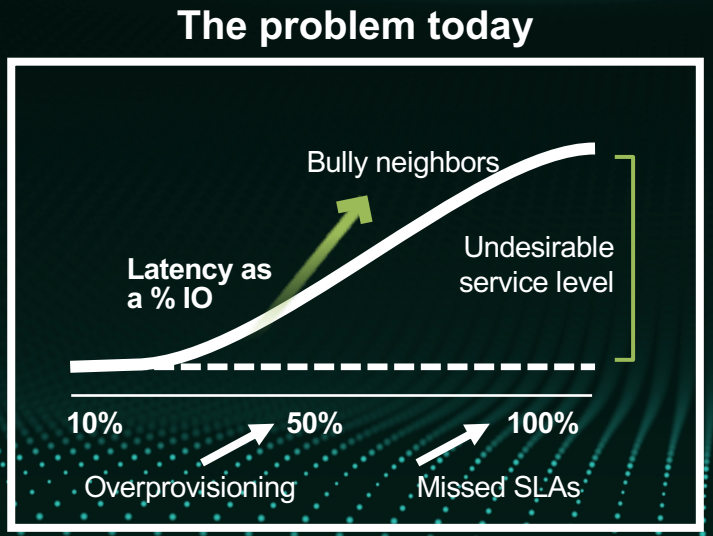
Defining the Right Mix of Hybrid IT is Challenging…

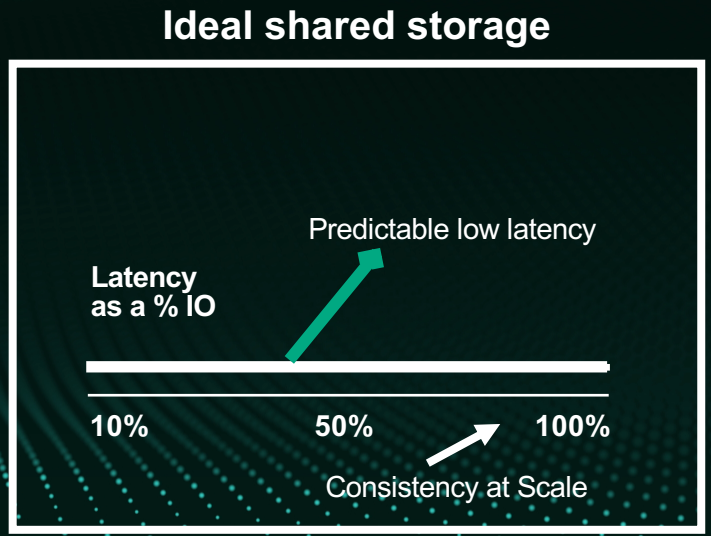


Forced Tradeoff between Agility and Resiliency



Organizations Fly Blind with Inconsistent Latency at Scale





The problem with performance today isn't IOPS and throughput – it's inconsistent latency due to multi-tenant workloads driving up overall response times!

One answer is to isolate mission critical apps.

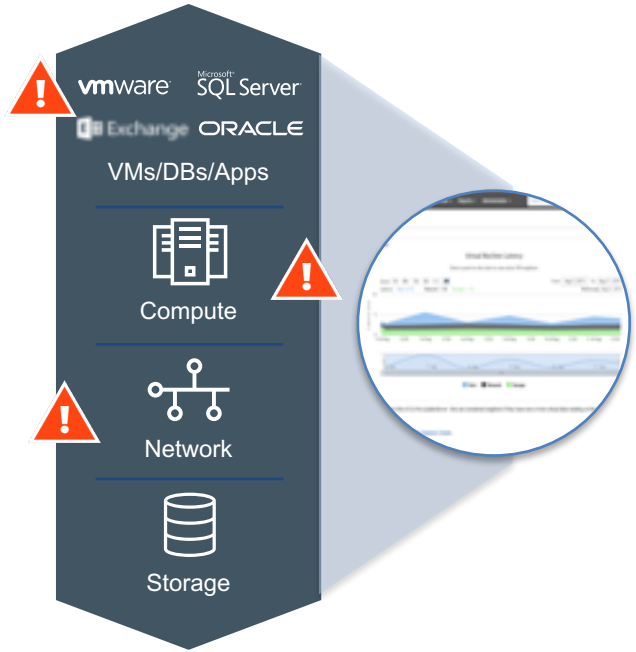
Understanding Workload Characteristics

Without the proper tools, understanding the I/O requirements of multi-tier and multi-tenant workloads is difficult.

  – Comparing the impact of multiple, frequently changing workloads is almost impossible and multi-tenancy adds to too varied latency.

- The same workload in a different company runs differently, each characteristic demands something different from a storage system.
- The ability to capture workload I/O characteristics, analyze that data and regenerate it is a critical capability for data centers to master.
- Workload profiling enables organizations to troubleshoot and optimize their current environment as well as plan for the future.

Cross-Stack Analytics for VMware Environments



Noisy Neighbor

Determine if VMs are hogging resources from another VM

Host & Memory Analytics

Visibility into host CPU and memory metrics

Latency Attribution

Identify root cause across host, storage, or SAN

Inactive VMs

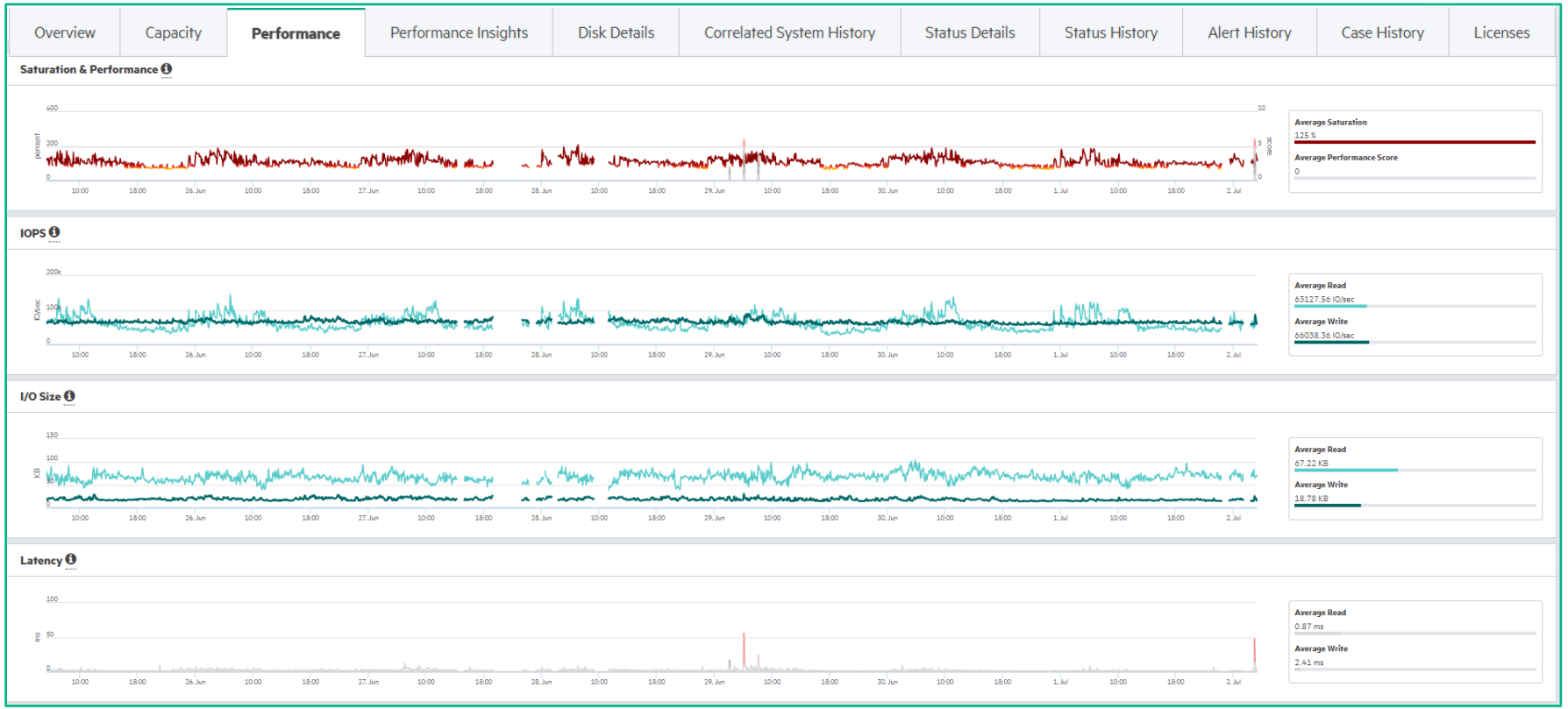
Visibility into inactive VMs to repurpose/reclaim resources

Top Performing VMs

Visibility into Top 10 VMs by IOPs and Latency

Storage System Analytics

Detailed performance overview of storage array



Flash Does Not Change Storage Requirements!

Reducing risk with a comprehensive approach to data integrity

High performance

Flash-optimized architecture

Scalability

Scale out architecture with multiple active/active nodes

Reliability

Proven architecture with guaranteed high availability

Disaster recovery

Data protection with sync and async and multiple sites

Application integration

VMware, Oracle, Microsoft Hyper-V, SQL Server, Exchange integrations

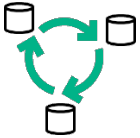
Drive efficiency

Extend life and utilization of flash;

110+% shared storage utilization vs.. >20% per host, each with NVMe drive(s).

Ease of use

Self-configuring, optimizing, and tuning



Data mobility

Federate across systems and sites

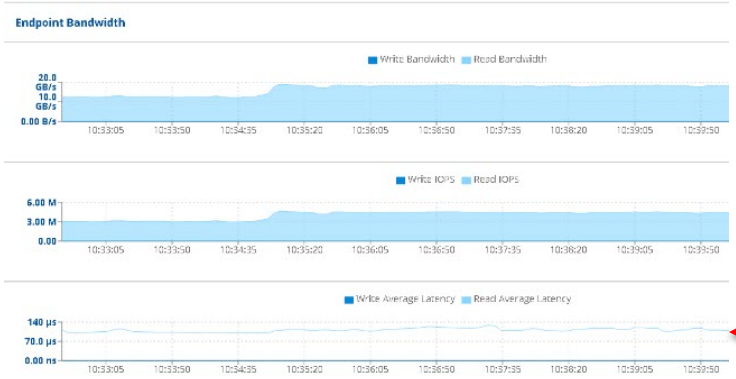
Flash Latency

Start Fast

Up to 50% Lower Latency

Stay Fast

Average 200us Latency or Below



Always Fast

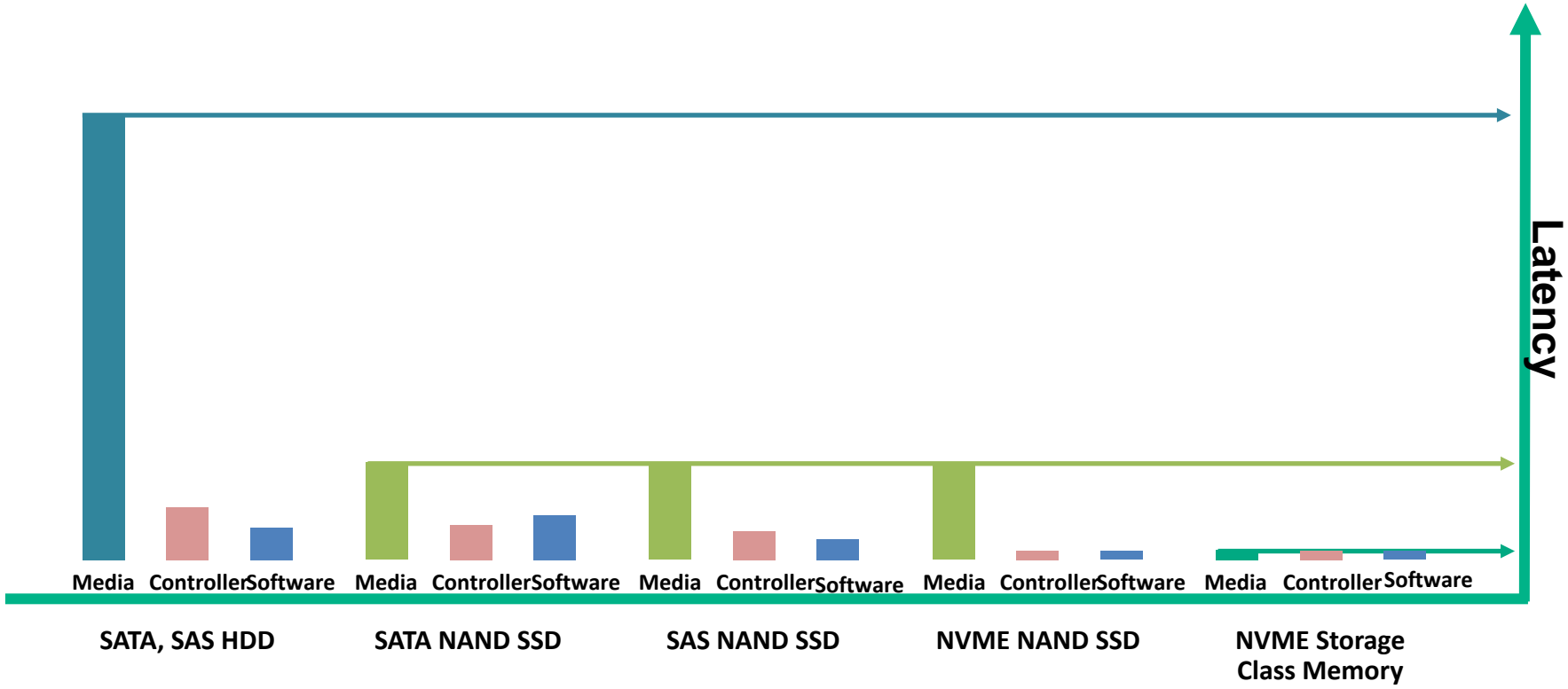
Near 100% within 300us Latency

125us avg. 4M+ IOPS

Measured by the host application using a storage class memory and flash storage system running a 8KB Random read workload and all flash storage system running a 4KB Random read workload.

Truth About an NVMe Storage Back-end

the source of storage latency



Storage Growth

Storage growth is being driven by new and evolving workloads

- mobile computing
- Big Data and analytics
- business-oriented social media
- custom applications - virtualized
- cloud computing
- migration of legacy workloads to virtual infrastructure
- **Migration of workloads back from the cloud to on-premises**
  - ✓ **41%[1] of businesses surveyed brought at least one workload back on-premises in 2018**

Key requirements for storage infrastructure for which virtualization is a major driver of new requirements

- the need to scale out easily and quickly
- Flexable storage media and interconnect support – SSDs, HDDs, mixes of SSD and HDD, NVMe SSD, storage class memory, etc..
- leverage disparate information sources (and pull data in and out of those sources)
- support applications that are geographically distributed
- DevOps oriented and built to support advanced analytics
- storage infrastructure should provide highly available, secure data access, line-rate, non-blocking, high-speed throughput, multi-tenancy, and consistently low latency service times.
- Millions of IOPS and response times in microseconds instead of milliseconds
- Availability is not just about hardware; it's also about a holistic approach with hardware, software, management, and the right architecture.
  - ✓ Fibre Channel SAN is the bedrock of a holistic approach

[1]ESG Market Research

Understanding Workload Characteristics

By identifying workload requirements and their I/O patterns, FC workloads can be mapped to storage and without ever needing to do a comparison to other protocols.

- ✓ However - it is important to gather actual performance metrics for best sizing results.

Each workload has unique characteristics, and each of these characteristics impacts latency, IOPS and throughput. These characteristics include:

1. I/O Mix - is the workload read heavy, write heavy, balanced, or bursty?
2. I/O type - does the workload write or read data sequentially or randomly?
3. Data/metadata mix - does the workload read or manipulate metadata more so than actual data?
4. Block or file size distribution - does the workload write in small or large blocks?
5. Data efficiency appropriateness - does the workload have highly redundant or compressible data so that functions like deduplication and compression work effectively?
6. Is the workload prone to specific hot spots?

How do all of the above characteristics change over a relevant time period?

FC SAN Analytics

Latency is a fundamental reason driving the choice for on-premises over cloud workloads.

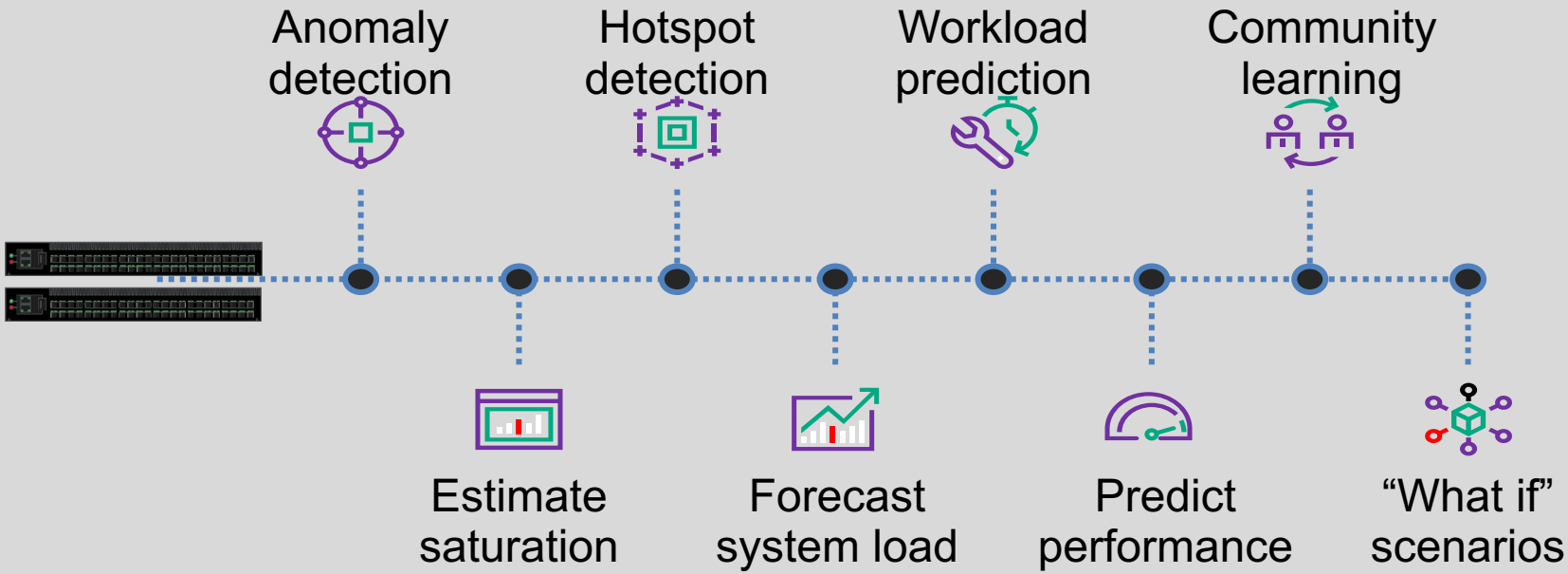
- There are serious operational, geopolitical, performance/latency, and regulatory details to consider before finalizing locality decisions.
- Applications that house very sensitive data, may want to reside on-premises and not beyond the confines of a data center and the FC SAN.

FC SAN Analytics can provide real time workload insight into the causes of performance degradation:

Fibre Channel equipment suppliers have added in-line support to FC for FC-NVMe SAN Analytics that helps with understanding and troubleshooting workloads in real time.

- ✓ FC SAN Analytics programs offer visibility into I/O traffic between compute and storage infrastructures including visibility into individual ports, switches, servers, virtual machines and storage arrays.
- ✓ The information generated by FC SAN Analytics can be used to maintain a performance baseline.
- ✓ A deviation from the historic trend can be used to generate alarms, resulting into proactive troubleshooting.
- ✓ Workload monitoring provides insight into the causes of performance related problems.
- ✓ It is important to gather actual performance metrics for best growth and maintenance plans.

FC SAN Analytics Engine



FC SAN-based Benchmarks Can Help

Benchmarks can provide insights into and set moderate workload expectations. $$

Benchmark = set of programs taken from real workloads ;

Examples:

TPC-C simulates a complete computing environment and involves a mix of five concurrent transactions of different types and complexity.

TPC-DS is the industry standard benchmark for measuring the performance of decision support solutions and is characterized by high CPU and I/O load as volumes of data are examined.

Data Warehouse Workload can be represented by TPC-E and TPC-H:

TPC-E is a "scalable" On-Line Transaction Processing (OLTP) workload using a real (not synthetic) Microsoft SQL Server database to model a brokerage firm and provides a transactional throughput numerical score.

TPC-H is a decision support benchmark which consists of a suite of business oriented ad-hoc queries and concurrent data modifications. Large volumes of data are examined and queried with a high degree of complexity.

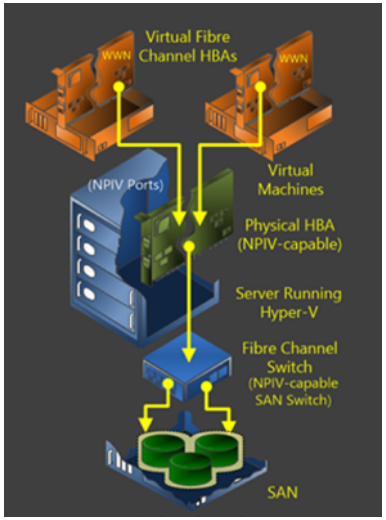
VMmark 3 allows accurate and reliable benchmarking of virtual data center performance and power consumption.

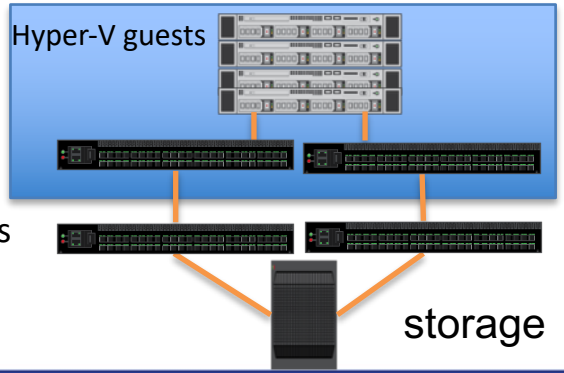
- ✓ High random mix of I/O transfer sizes

SPEC SFS 2014 designed to evaluate performance using file server end-to-end throughput and response time.

FC has Many Virtual Technologies

- VSAN partitions of FC switches provide virtual isolation of a group of ports and their associated traffic, ports within each VSAN can be zoned to provide refined logical connectivity.
- VSANs can support Virtual Fabric Tagging Header which allows FC frames to be tagged with a VF Identifier (VF_ID) of the VF to which they belong – used to share an ISL with other VSANs.
- FC HBA ports (target or server) utilize NPIV - a FC feature whereby multiple FC node ports (N_Port) IDs can share a single physical N_Port and each can be zoned separately.
- Virtual Fibre Channel for Hyper-V guests uses FC HBA NPIV to map multiple N_Port IDs to a single physical Fibre Channel N_port. A new NPIV port is created with each virtual HBA.
- FC switch VSAN partitions and NPIV provisioned HBAs combine to enable a virtual FC system.
- Hyper-V guests directly access FC LUNs as if operating on a physical server.
- Virtual FC in Hyper-V guests includes support for related features, such as vSAN, live and quick migration, MPIO, Import and Export, Save and Restore, Pause and Resume, and guest initiated backups.





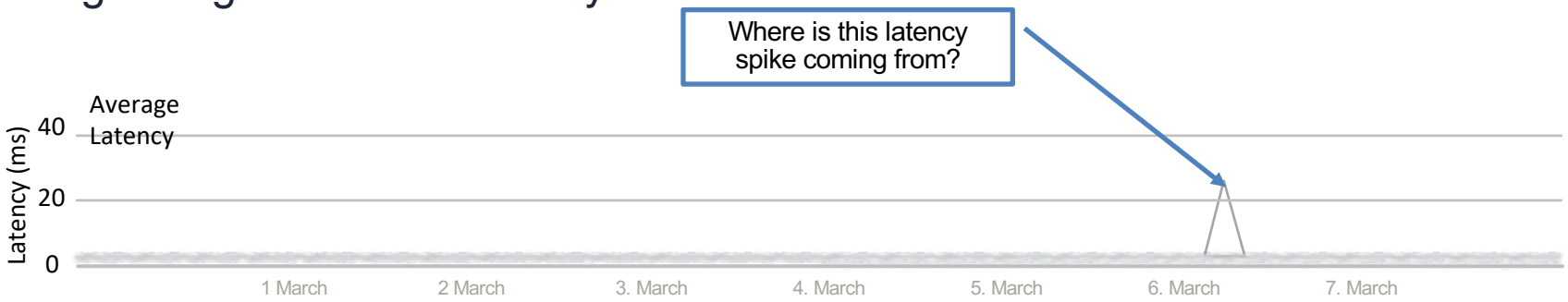
FCF mode switch provides fabric login services

NPV mode switch proxies the FC login to the FCF switch on behalf of each of the attached servers which then utilize NPIV WWNs in zones to connect with storage volumes

NPIV Use Cases

VMID for VMware FC SAN Environments

Diagnosing the traditional way



Traditional monitoring tool graphic

Steps

1. Check application logs and stats
2. Check VMware vSphere logs and stats
3. Check array logs and stats
4. Check network logs and stats
5. Try correlating all events, logs, and performance stats

VMID Based FC SAN Telemetry

VM Datastore volumes are often shared between many VMs:

- FC switch fabrics are equipped to provide monitoring functions
  - making use of VMID technologies requires complementary HBA based telemetry components
- VMID frame tagging provides increased visibility of VM traffic which is associated to different VMs and used by the Hypervisor, FC switches, and storage to understand the data flowing across the SAN(s).
  - Enables End-to-End Quality of Service (QoS): the ability to apply specific levels of QoS on a per workload basis to direct FC traffic from a specific VM through the fabric and onto the end storage device.
- An application Services monitor works by gathering the globally unique ID from the hypervisor, for example from VMware ESX, and interprets the different IDs from every VM to perform intelligent monitoring which enables the application of a QoS policy to each VMs traffic.
- Having the ability to view how different flows from even a single VM can be sent to the associated tiered storage for proper handling allows for improved troubleshooting.

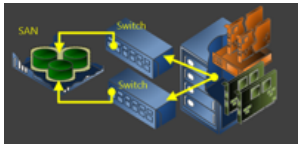
Analytics for VMware Environments

detailed breakdown of a VM

Three Basic FC SAN Topologies

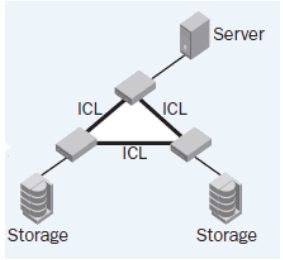
SAN architecture represents a combination of computer system and network characteristics that must work together to sustain the workloads.

1. excluding direct attach and a single switched environment
   - 3 basic fabric categories: cascading, meshed, and core /edge configurations.
2. These categories provide starting points for data center specific customization and workload specialization.
3. SAN expansion and extension and derivations come from one of these three configurations.



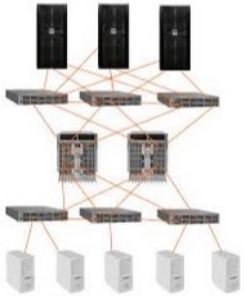
Cascading SAN Configuration

✓ This configuration provides a switch-to-switch connection that allows the number of server and server devices to scale quickly. Also known as dual-fabric configuration.



Meshed SAN Configuration

✓ This configuration provides a performance-oriented system that allows for the quickest path from server to storage, based on FC Fabric Shortest Path First (SPF).



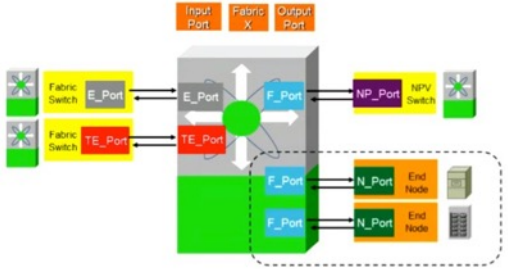
Core/Edge SAN Configuration

✓ Performance and scale oriented configuration takes into account I/O optimization, redundancy, and recovery and is complex in implementation and configuration management.

Fibre Channel Foundational Fabric Services

FC-CT protocol provided fabric services - Technical name is FC-CT protocol (Common Transport for Generic Services)

- ✓ FC Generic services protocol
  - fabric services centrally located in each switch (embedded)
  - Provides the control plane for information transfer between devices
  - Distributed services
- ✓ Principal switch and Domain ID assignment
- ✓ FSPF routing protocol Routing and ISL initialization (synchronization)



Name Server – Fabric, Distributed

Zoning – Security and Access controls, Distributed zoning; Seamless scalability vs. layering over VLAN

High Performance – Flow Control with Buffer Credits – auto recovered, fixed 2K packets, In order Delivery

- ✓ Congestion Control Algorithms that work

Low Latency – Highly optimized (ASIC does I/O transfer, very little software overhead)

Multi generational speed and feature/function and Interoperable auto-negotiated link speeds and signaling

64GFC (PAM4 + FEC), 32GFC (NRZ + FEC), 16GFC (NRZ); 128GFC multi-link speed

High Availability –NSPF with redundant fabrics with MPIO; Extensibility over long distances

Scalability – Easy scalability from single point-to-point links to integrated enterprises with thousands of physical and virtual connections (WWNs)

Virtual Channels for VMIDs and SAN telemetry for differentiating and prioritizing traffic

Centralized FC Fabric Services

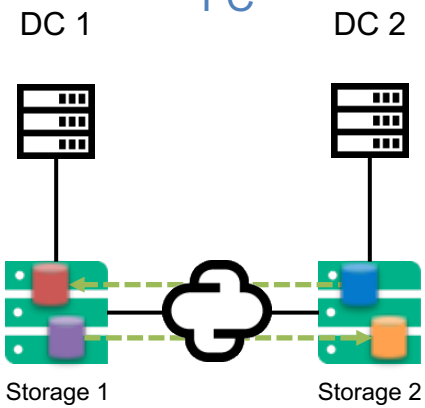
Fabric Login Server	Name Server	Fabric Controller	Management Server
Used during a node's fabric login process. FC address – domain ID, area ID and port ID assigned during login. Located at a well known address FFFFFE Scale: 239 domain ID addresses 239 x 256 areas x 256 ports = 15663104 possible node ports	Responsible for name registration and management of node ports. Located at well known address FFFFFC	Responsible for managing and distributing registered state change notifications (RSCNs) to attached node ports. Responsible for distributing SW-RSCNs to every other switch. SW-RSCNs keep the name server up-to-date on all switches. Located at well known address FFFFFD	Enables FC SAN management using fabric management software Located at well known address FFFFFA

Not available for Ethernet based NVMe-oF and iSNS seldom used with iSCSI

FC Enables Data Availability and Protection

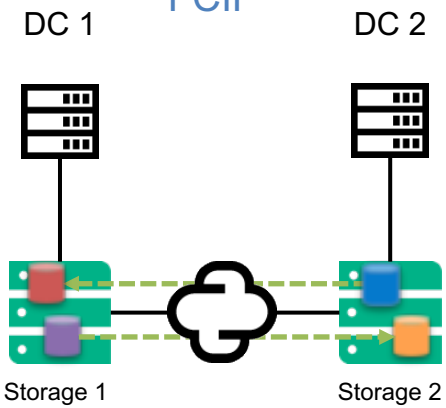
FC distance and FCIP simplify disaster recovery and high availability

Sync Remote Copy FC



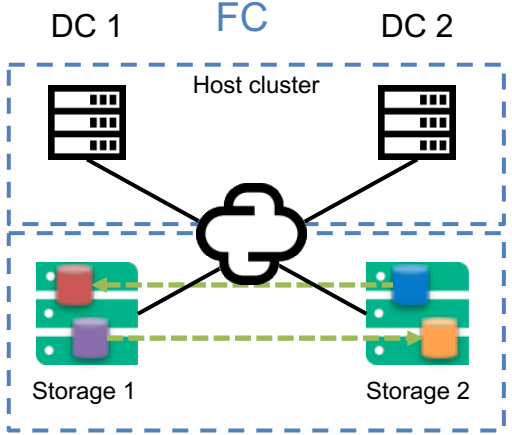
RPO = 0
High availability across metro distances with up to 10 ms RTT (~1000 km)

Async Remote Copy FCIP



RPO = seconds to minutes
Disaster recovery across continental distances with up to 150 ms RTT (~15,000 km)

Metro Cluster FC

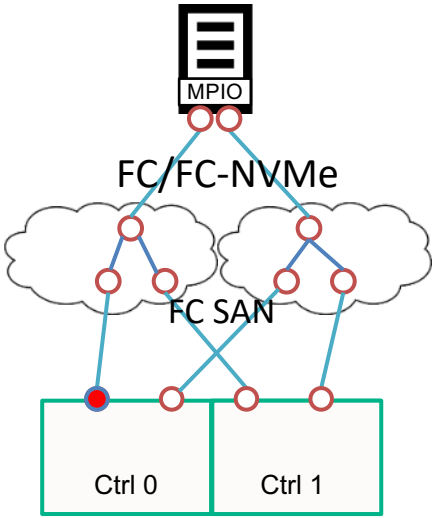


RPO + RTO = 0
High availability across metro distances with up to 10 ms RTT (~1000 km)

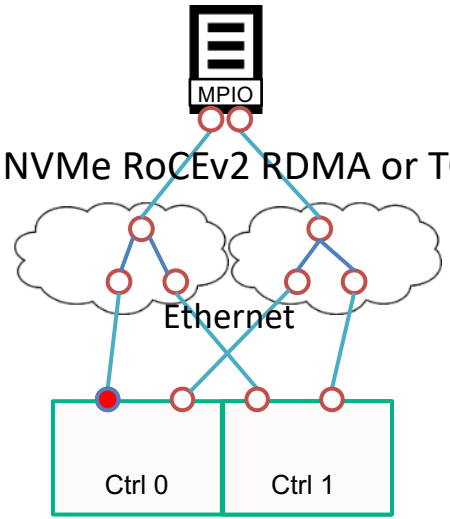
RTT = round-trip-time

Which Solution for Your Workload?

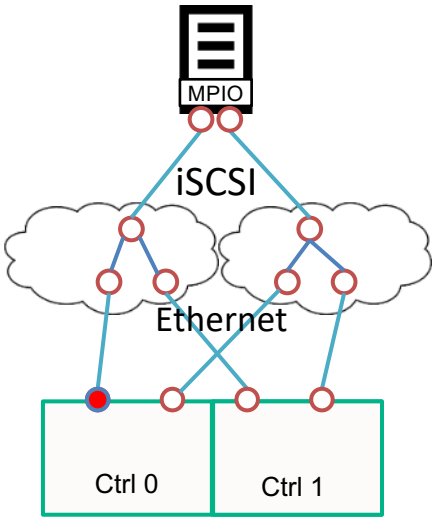
- FC – predictable no risk option; fabric controller facilitates discovery and automated zoning and event notification; scales easily, MPIO, standards based
- RoCEv2 and TCP – evolving; constant change; no fabric services (some proposals); manual IP address assignment; no event notifications; difficult to scale; HA?, new approach to MPIO, open source and proprietary implementations
- iSCSI – no fabric services; manual IP management; difficult to scale



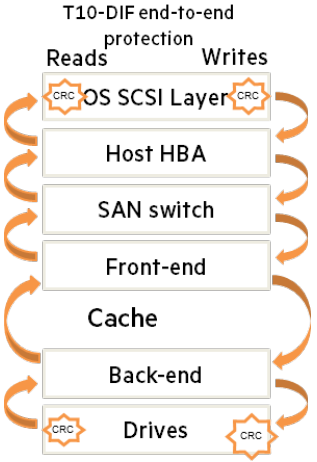
Hardened and Optimized technologies – stable software stacks and hardware/FW



DCBx/pfc + ECN for lossless operation? Software stacks are fluid. RoCEv2 or TCP? DAS–replacement. Remember FCoE?



We tried lossless iSCSI DCBx/pfc Latency is quite varied due to TCP Scale is difficult to manage



FC-NVMe/FC-NVMe-2 Update

FC-NVMe Standard was T11 ratified 08/10/2017

- Limited link level error recovery included
- Protocol level error recovery was limited due to SCSI and NVMe having their own procedures
- FC transport support is available in Unified Host/Target SPDK
- MPIO optimizations continue to be evolved by the Linux community

Why the need for FC-NVMe-2 ?

1. Bit errors happen – sometimes just due to optics temperature, cosmic particles, laser wearout, hardware fails, and software fails - basically resulting in frame loss.
2. NVMe recovery procedures were further defined and documented in later revisions of ratified TPs.
3. New FC Link Services were defined for faster error recovery without protocol layer notification.
4. Improvements were made in FC-NVMe and also SCSI FCP-5

FC-NVMe-2 focused on enhanced/refined error recovery – also documented in FC-FS-6 as a result FC-NVMe-2 has industry leading error detection and recovery, centralized fabric control, side-by-side SCSI FCP and robust hardened discovery and name service, zoning and security.

1. Detected error recovery at the transport layer i.e., Sequence Level Error Recovery (SLER)
2. Eliminate as much as possible the dependency on protocol layer for error recovery
3. Sequence Level Error Recovery (SLER), which allows error recovery by using sequence re-transmissions
4. Accomplished via new commands (FLUSH and Responder Error Detected – RED)
5. Support added for confirmed completion – a mechanism so that the target and initiator ports can use messages to determine successful completion of all sequences within a given exchange

Beyond B2B Credit Zero Detection

SAN congestion is a reduction in the throughput of the SAN

- Known to affects most SANs without administrators knowing – not easy to detect without SAN analytics
- SAN congestion is a term covering at least four behaviors of an FC SAN :
  - Slow drain devices - devices withholding the ability of the FC SAN to send traffic
  - Speed mismatches between host and target
  - Excessive Temporal Proximate I/O (micro bursts on a target port)
  - B2B credit return; stalling of R_RDY transmissions (by targets and initiators) which back pressure the SAN fabric, and possibly an ISL, which may impact many devices

FC flow control attempts to minimize the chance of dropped frames by transmitting when the receiver has a buffer

- For each frame sent an R_RDY (B2B Credit) should be returned
- R_RDYs are returned once an occupied receive buffer location has been handled
- R_RDYs are not sent reliably – they can be corrupted/lost or just withheld i.e., stalled
- B2B credit automated non-disruptive recovery Link Credit Reset is a default in today's FC fabrics
- Each side informs the other side of the number of buffer credits it has
- When at 0 Tx credits, no frames can be sent! If at 0 Tx credits long enough – begin automatic recovery.

B2B credits are not negotiated – they are agreed to during login exchange of parameters

- ✓ Dynamic B2B Credit Recovery is part of this capability exchange (see FC-FS-4) – B2B recovery works!

Fabric Notifications – Help from the Ends

Extend Exchange Diagnostic Capabilities ELS support of Register Diagnostic Functions (RDF) and Fabric Performance Impact Notification (FPIN). End nodes can be taught to help with fabric congestion problems!

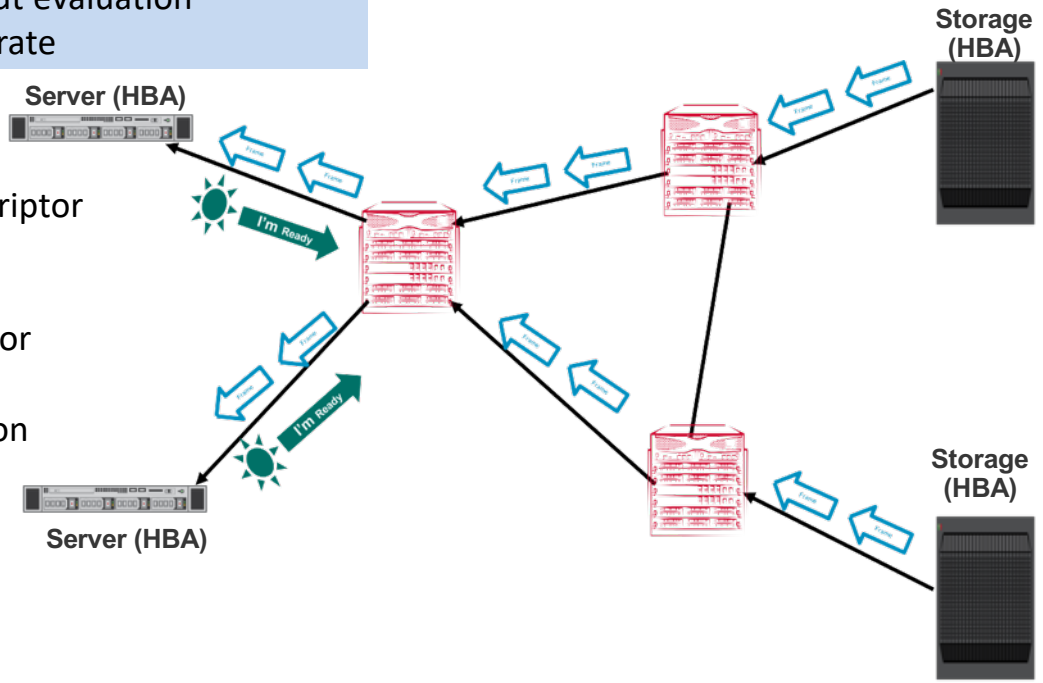
Example: Distinguish between Credit Stall and Oversubscription

- Credit Stall requires an internal evaluation
  - Buffer credits are not returned at line rate
- Oversubscription requires a throughput evaluation
  - Buffer credits are returned at line rate

FC-LS-5 and FC-FS-6 modifications for congestion notifications, add descriptors:

- Congestion Detection Capability descriptor
- Link Integrity Notification descriptor
- Delivery Notification descriptor
- Peer Congestion Notification descriptor
- Congestion Notification descriptor
- Fabric Performance Impact Notification (FPIN) Registration descriptor

Add link primitives as notifications:
ARB(F1) Warning Congestion Signal
ARB(F7) Alarm Congestion Signal



Q & A

After this Webcast

- Please rate this event – we value your feedback
- We will post a Q&A blog at http://fibrechannel.org/ with answers to the questions we received today
- Follow us on Twitter @FCIAnews for updates on future FCIA webcasts
- Visit our library of FCIA on-demand webcasts at http://fibrechannel.org/webcasts/ to learn about:
  - Fibre Channel Fundamentals
  - FC-NVMe
  - Long Distance Fibre Channel
  - Fibre Channel Speedmap
  - FCIP (Extension): Data Protection and Business Continuity
  - Fibre Channel Performance
  - FICON
  - Fibre Channel Cabling
  - 64GFC
  - FC Zoning Basics

Our Next FCIA Webcast:

The Making of Standards

Follow us @FCIANews

for date and time

Thank You