IDF15
INTEL DEVELOPER FORUM

Next Generation Storage Architecture: Microsoft* Storage Spaces Direct and Intel® SSD Data Center Family for NVM Express™

Hamesh Patel – Senior Software Engineer, Intel Corporation
Spencer Shepler – Principal Architect, Microsoft

SFTS015

intel
experience
what's inside™

Agenda

- Current Trends and Impact to Storage
- Intel and Microsoft* Joint Collaboration
- What is Microsoft Storage Spaces Direct?
- The Need For NVM Express™
- Methodology and Performance Results
- Details of Demo in NVM Express Community
- Summary and Next Steps/Q&A

Current Trends and Impact to Storage

Current Trends and Impact to Storage

- Rise of content rich and content driven enterprise
- Hyper-scale public cloud
- SDS is evolving
- Non Volatile Memory

Moore's Law and Intel's Innovation Engine is Key!

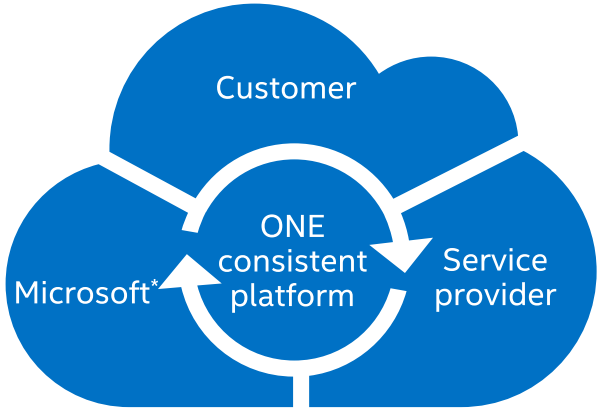
Intel and Microsoft* Joint Collaboration

IDF15
INTEL DEVELOPER FORUM

Intel and Microsoft* are Collaborating on a Storage Solution

- Joint design of a reference architecture
  - Integrated solution that is performance optimized with the latest Intel hardware and Microsoft* software
  - Cost optimized to meet broad market demands
  - Feature rich to compete in the market

Customer Choices



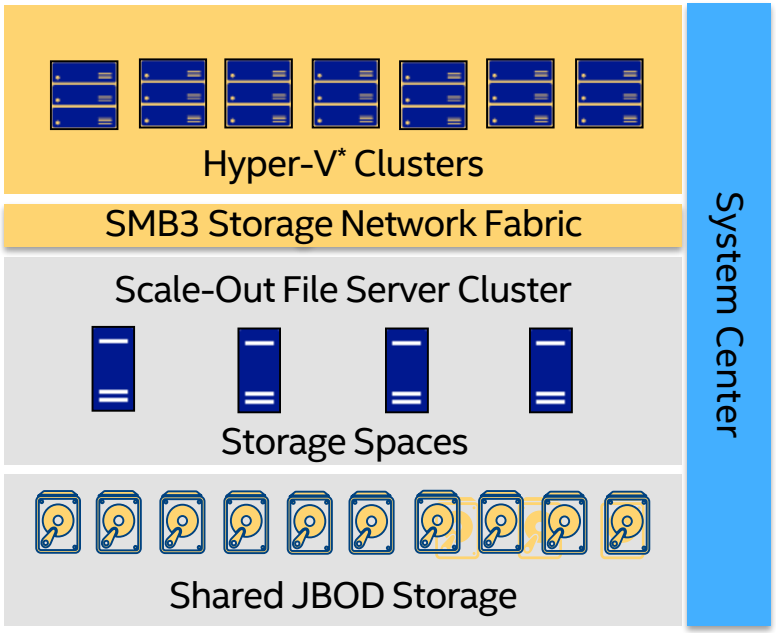
Private cloud with traditional storage	Private cloud with Microsoft SDS	Hybrid cloud Storage	Public cloud Storage
SAN and NAS storage	Microsoft Azure* Stack Storage	StorSimple* with Azure storage	Azure storage

What is Microsoft* Storage Spaces Direct?

Windows® Server* 2012 R2 Software Defined Storage Recap

Primary application data storage

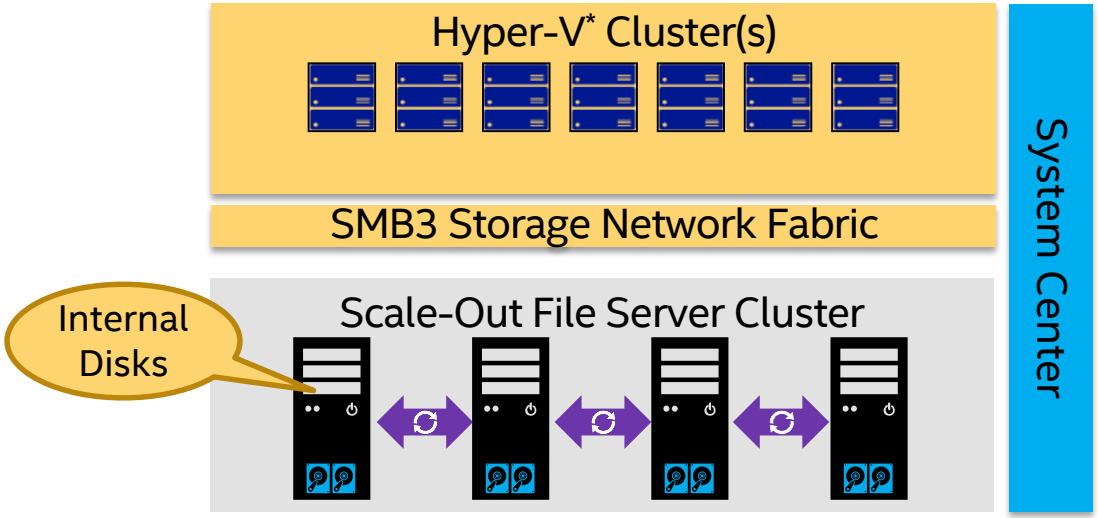
- Cost effective with standard volume hardware
- Continuously available
- Scale-out and elasticity
- Reliability
- Unified management



Introducing Microsoft* Storage Spaces Direct

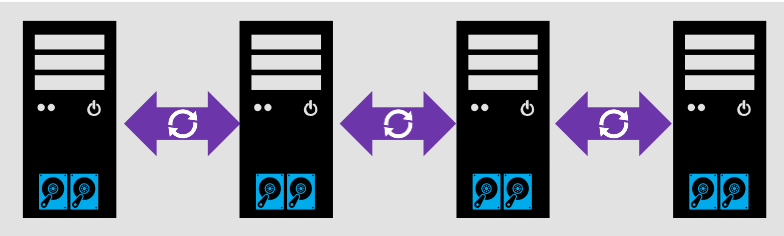
What is Storage Spaces Direct?

- Private cloud storage
- Servers with local storage
- Highly available and scalable
- Enabling new scenarios
- Evolution of Storage Spaces



Why Microsoft* Storage Spaces Direct?

- Embrace new disk device types
  - Lower cost flash storage with SATA* SSDs
  - Better flash performance with NVM Express™ SSDs
- Simple deployment
  - Network fabric instead of SAS fabric
  - External enclosures not needed
- Seamless expansion
  - Add more nodes
  - Storage rebalancing
  - Increase scalability

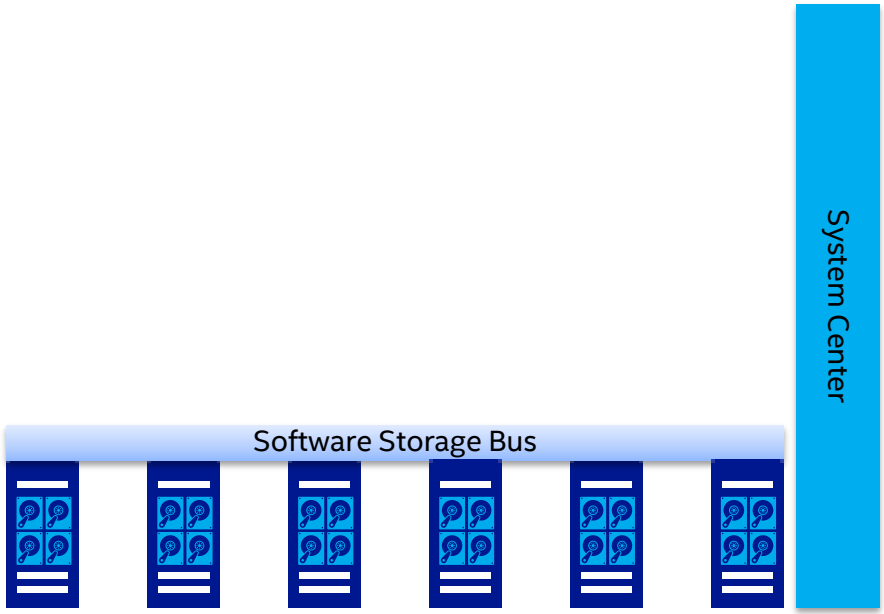


Under the Hood

- Servers with local disks
  - SATA*, NVM Express™, SAS

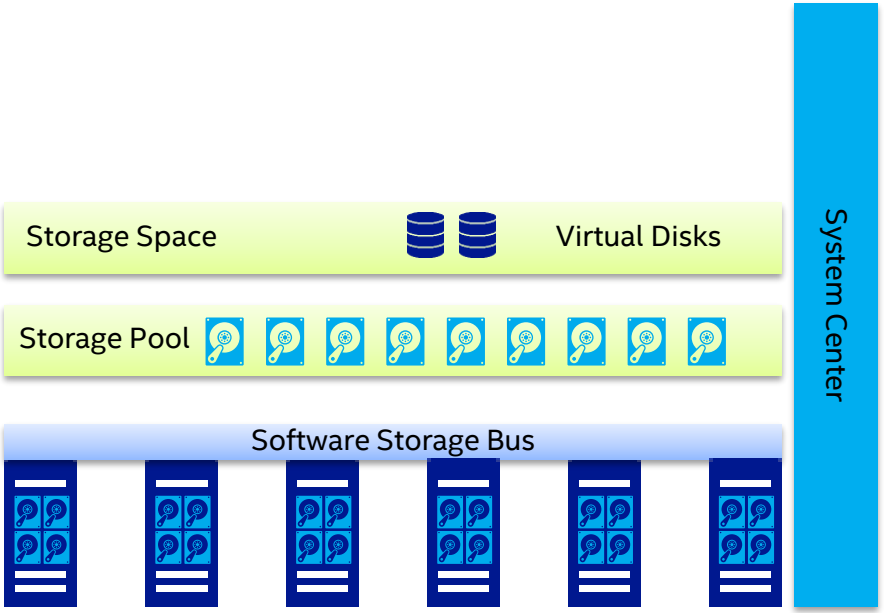
Under the Hood

- Software Storage Bus
  - Spans entire cluster, Leverages SMB3/SMBDirect
- Servers with local disks
  - SATA*, NVM Express™, SAS



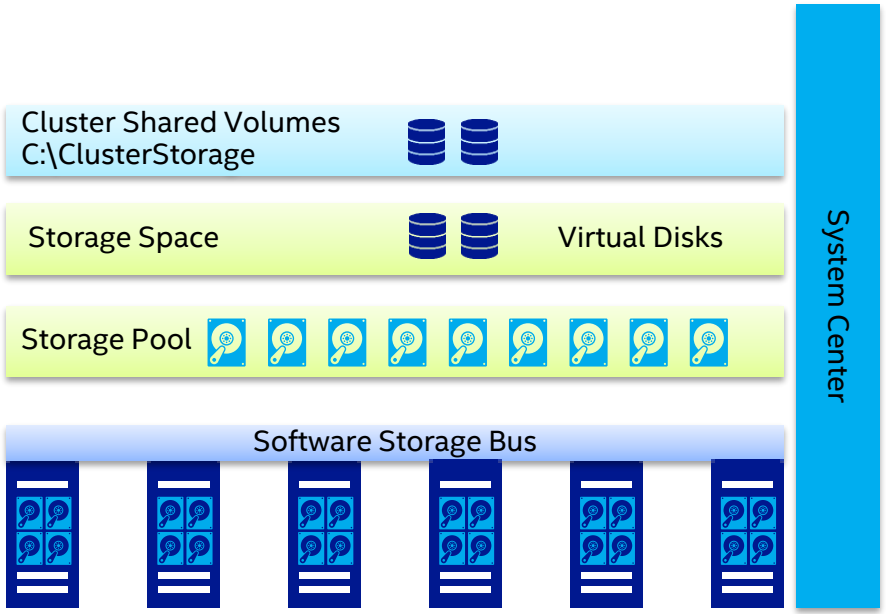
Under the Hood

- Storage Spaces
  - Scalable pool with all disk devices
  - Resilient virtual disk
- Software Storage Bus
  - Spans entire cluster, Leverages SMB3/SMBDirect
- Servers with local disks
  - SATA*, NVM Express™, SAS



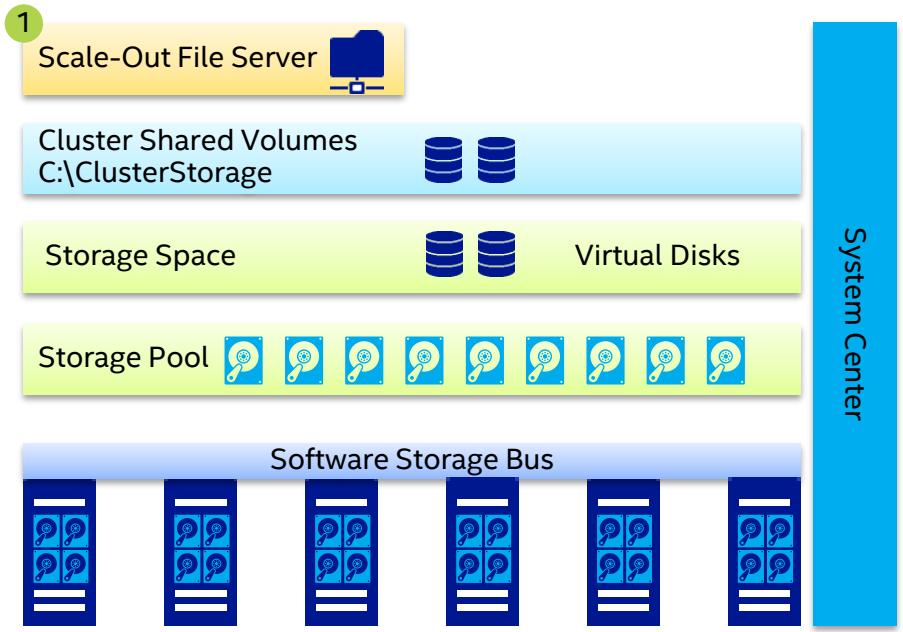
Under the Hood

- File System (CSVFS with ReFS)
  - ReFS is the primary Cluster-wide file system
  - Fast VHDX creation, expansion and checkpoints
- Storage Spaces
  - Scalable pool with all disk devices
  - Resilient virtual disk
- Software Storage Bus
  - Spans entire cluster, Leverages SMB3/SMBDirect
- Servers with local disks
  - SATA*, NVM Express™, SAS



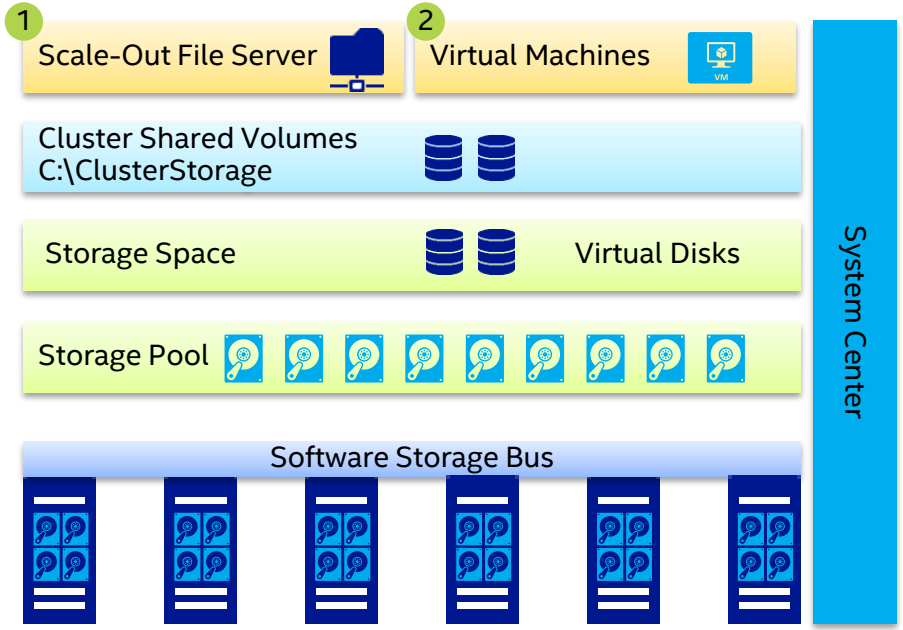
Under the Hood

- Deployment Modes
  - Remote data Access using Scale-out File Server
- File System (CSVFS with ReFS)
  - ReFS is the primary Cluster-wide file system
  - Fast VHDX creation, expansion and checkpoints
- Storage Spaces
  - Scalable pool with all disk devices
  - Resilient virtual disk
- Software Storage Bus
  - Spans entire cluster, Leverages SMB3/SMBDirect
- Servers with local disks
  - SATA*, NVM Express™, SAS



Under the Hood

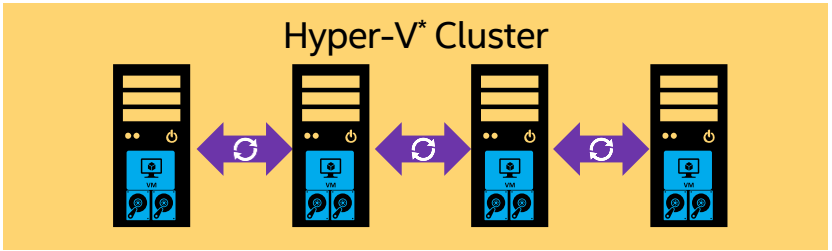
- Deployment Modes
  - Remote data Access using Scale-out File Server
  - Hyper-Converged
- File System (CSVFS with ReFS)
  - ReFS is the primary Cluster-wide file system
  - Fast VHDX creation, expansion and checkpoints
- Storage Spaces
  - Scalable pool with all disk devices
  - Resilient virtual disk
- Software Storage Bus
  - Spans entire cluster, Leverages SMB3/SMBDirect
- Servers with local disks
  - SATA*, NVM Express™, SAS



Microsoft* Storage Spaces Direct – Deployment Choice

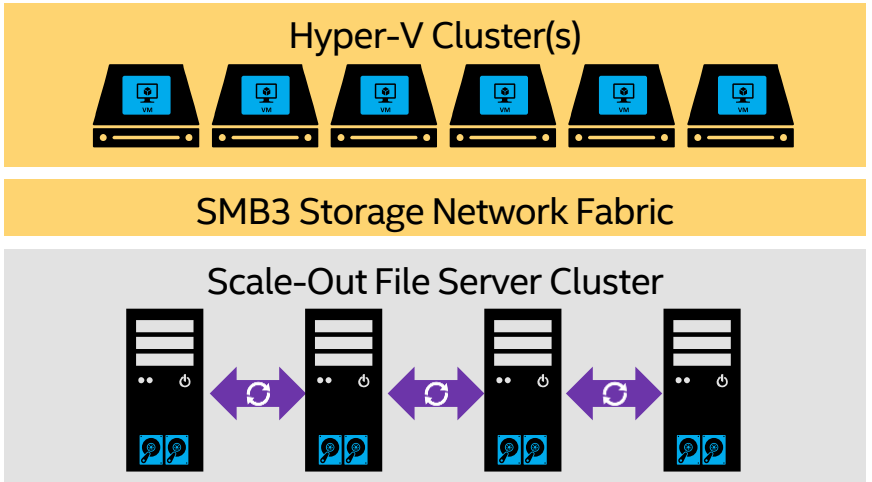
Hyper-converged

- Compute and Storage resources together
- Compute and Storage scale and are managed together
- Typically small to medium sized scale-out deployments



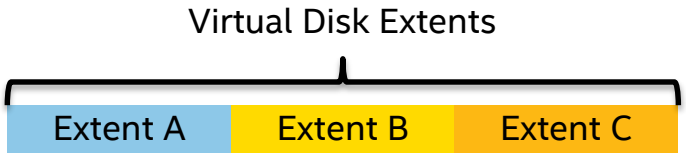
Converged (Disaggregated)

- Compute and Storage resources separate
- Compute and Storage scale and are managed independently
- Typically larger scale-out deployments



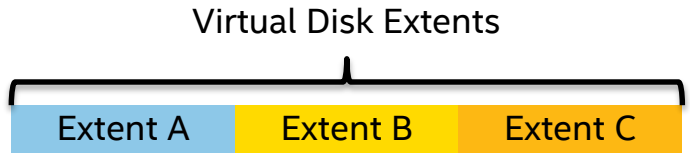
Microsoft* Storage Spaces Direct – Data Placement

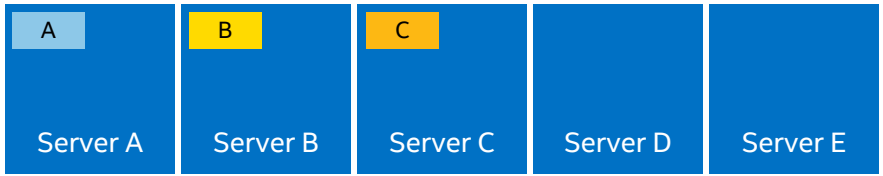
- Virtual Disks
  - Virtual disks consists of extents
  - Extents are 1GB, so a 100GB virtual disks has 100 extents



Microsoft* Storage Spaces Direct - Data Placement

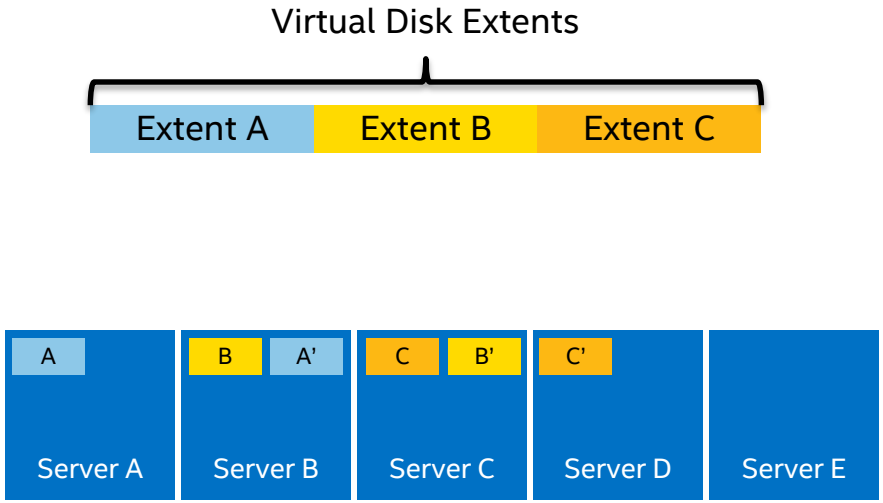
- Virtual Disks
  - Virtual disks consists of extents
  - Extents are 1GB, so a 100GB virtual disks has 100 extents
- Scale-Out
  - Extents are distributed across disks and servers





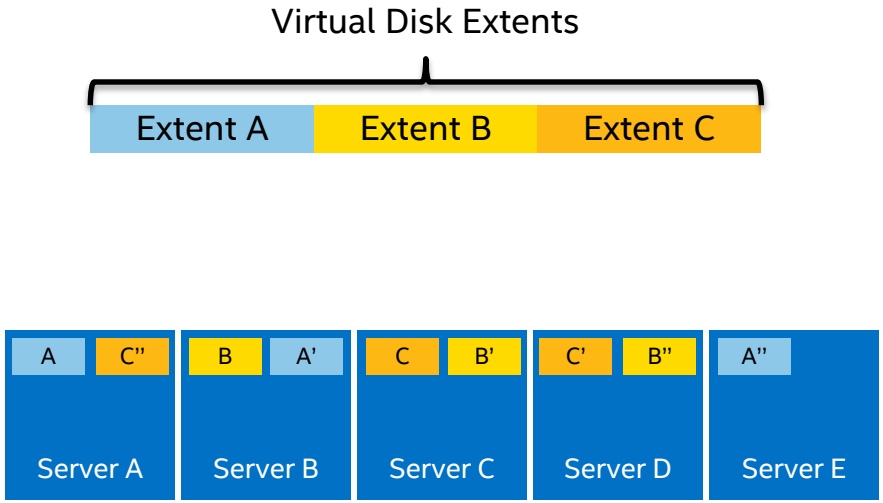
Microsoft* Storage Spaces Direct - Data Placement

- Virtual Disks
  - Virtual disks consists of extents
  - Extents are 1GB, so a 100GB virtual disks has 100 extents
- Scale-Out
  - Extents are distributed across disks and servers
- Resiliency
  - A 2nd copy of an extent is placed on a different server



Microsoft* Storage Spaces Direct - Data Placement

- Virtual Disks
  - Virtual disks consists of extents
  - Extents are 1GB, so a 100GB virtual disks has 100 extents
- Scale-Out
  - Extents are distributed across disks and servers
- Resiliency
  - A 2nd copy of an extent is placed on a different server
  - A 3rd copy of an extent is placed on yet another server



Microsoft* Storage Spaces Direct – Scalability

- Scalable pools
  - Scaling to large pools

Microsoft* Storage Spaces Direct – Scalability

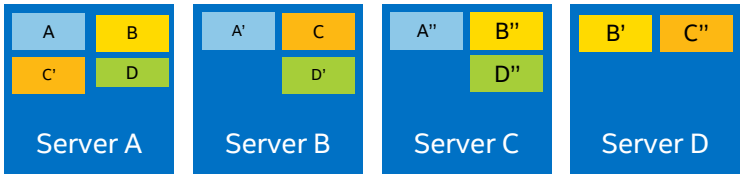
- Scalable pools
  - Scaling to large pools
- Fast interconnect
  - Utilizes SMB3 and SMB Direct
  - RDMA interconnect for low latency\CPU usage

Microsoft* Storage Spaces Direct – Scalability

- Scalable pools
  - Scaling to large pools
- Fast interconnect
  - Utilizes SMB3 and SMB Direct
  - RDMA interconnect for low latency\CPU usage
- Simple expansion
  - Add node(s) and expand pool
  - Rebalance for capacity

Microsoft* Storage Spaces Direct - Scalability

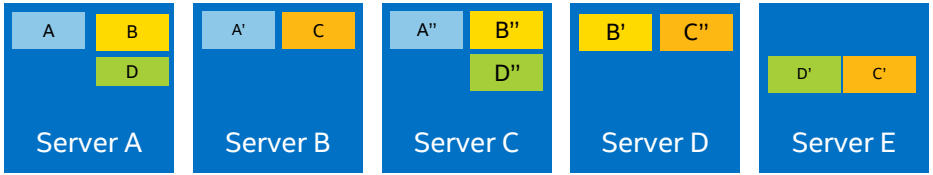
- Scalable pools
  - Scaling to large pools
- Fast interconnect
  - Utilizes SMB3 and SMB Direct
  - RDMA interconnect for low latency\CPU usage
- Simple expansion
  - Add node(s) and expand pool
  - Rebalance for capacity



3-way mirror

Microsoft* Storage Spaces Direct - Scalability

- Scalable pools
  - Scaling to large pools
- Fast interconnect
  - Utilizes SMB3 and SMB Direct
  - RDMA interconnect for low latency\CPU usage
- Simple expansion
  - Add node(s) and expand pool
  - Rebalance for capacity
  - Min – 4 servers, Max - 12 servers
  - Maximum of 240 disk devices in a single pool



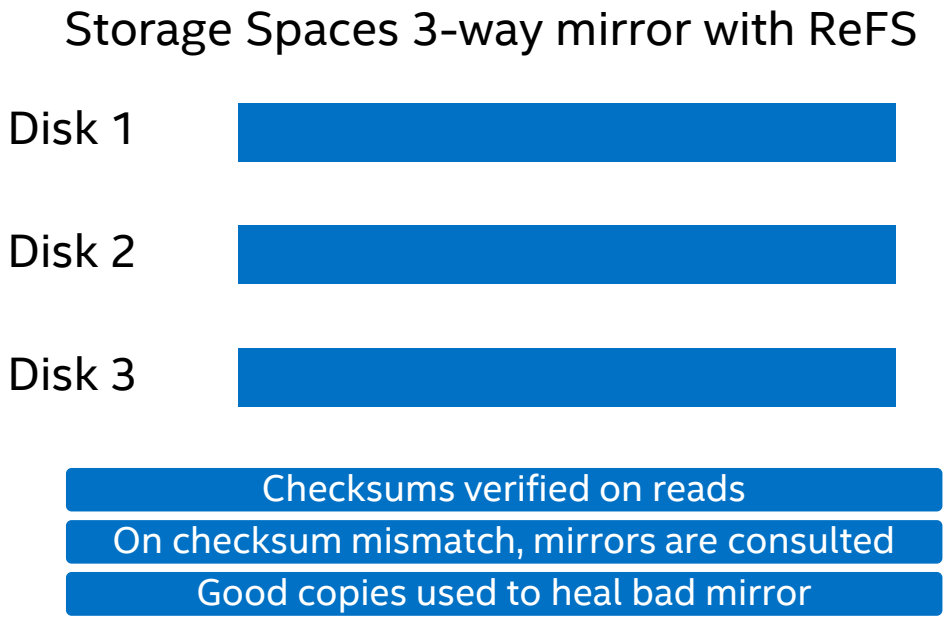
3-way mirror

ReFS – Data Integrity

- Metadata Checksums
  - Checksums protect all filesystem metadata
- User Data Checksums
  - Optional checksums protect file data
- Checksum Verification
  - Occurs on every read of checksum-protected data
  - During periodic background scrubbing

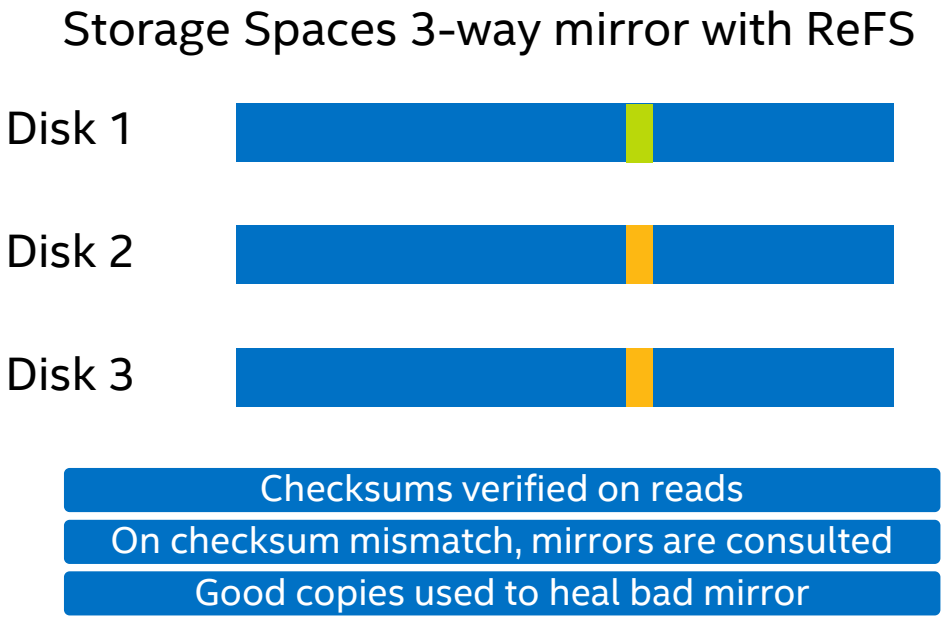
ReFS – Data Integrity

- Metadata Checksums
  - Checksums protect all filesystem metadata
- User Data Checksums
  - Optional checksums protect file data
- Checksum Verification
  - Occurs on every read of checksum-protected data
  - During periodic background scrubbing



ReFS – Data Integrity

- Metadata Checksums
  - Checksums protect all filesystem metadata
- User Data Checksums
  - Optional checksums protect file data
- Checksum Verification
  - Occurs on every read of checksum-protected data
  - During periodic background scrubbing



ReFS – Resiliency and Availability

ReFS – Resiliency and Availability

- Availability
  - Designed to stay online
  - Keep data accessible when all else fails
- Online Repair
  - Repairs performed without taking the volume offline
  - No downtime due to repair operations!
- On-Volume Backups of Critical Metadata
  - Automatically maintain backup of critical metadata
  - Online repair consults backup if checksum-based repair fails
  - Provides additional protection for volume-critical metadata

ReFS – Speed and Efficiency

-
-
-

ReFS – Speed and Efficiency

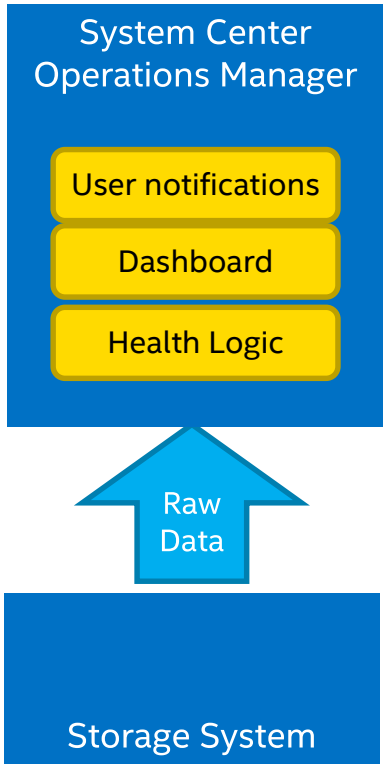
- Efficient VM checkpoints and backup
  - Checkpoints cleaned up without physical data copies
  - Data migrated as a ReFS metadata operation
    - Significant reduction of I/O to disk
    - Almost instantaneous
    - Minimal impact to workloads

ReFS – Speed and Efficiency

- Efficient VM checkpoints and backup
  - Checkpoints cleaned up without physical data copies
  - Data migrated as a ReFS metadata operation
    - Significant reduction of I/O to disk
    - Almost instantaneous
    - Minimal impact to workloads
- Accelerated Fixed VHD(X) Creation
  - VHD(X) zeroed as a metadata operation
  - No impact to workloads
  - Decreased VM deployment time
  - Also applicable to dynamic VHD(X) expansion

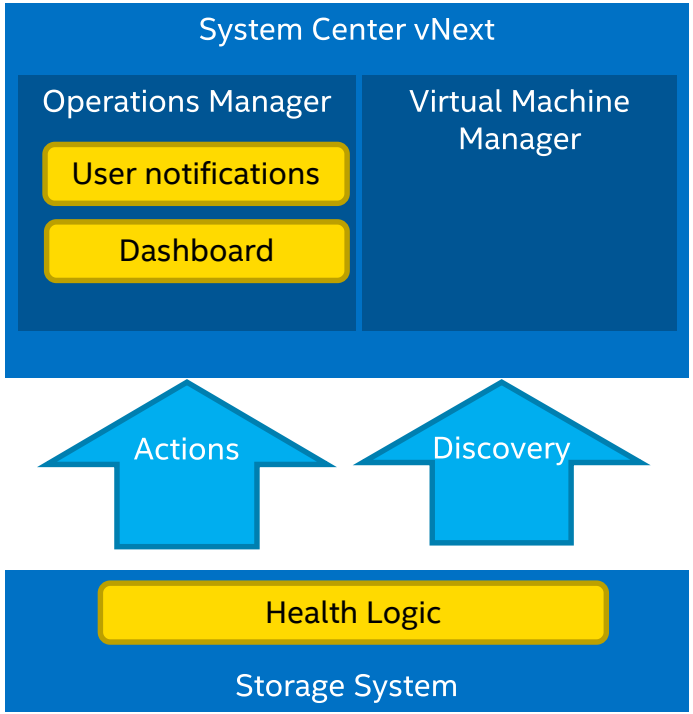
Traditional SCOM Storage Monitoring

- System Center Operations Manager determines health
  - Collect health indicators, state computation and rollup
  - Management pack (MP) tightly coupled to the topology, version and storage technology
- Requires highly knowledgeable MP authors
  - Storage subsystem model replicated by MP author
    - Complex health model
    - Need to handle different subsystems, topologies, versions, etc.
- Difficult to represent system health accurately
  - Monitoring likely to be incomplete, erroneous and noisy
  - Unable to determine actual impact on pools/shares
- No easy extensibility
  - Everything encoded in management packs
  - Write new/updated monitoring for each new variant



SCOM Storage Monitoring Reimagined

- Storage subsystem determines health
  - Focus on relevant objects
    - Storage subsystem, volumes and file shares
    - Automatic remediation when possible
- Actionable alerts
  - Alert specifies urgency
  - Alert specifies remediation action
  - Alerts automatically resolve when issue is addressed
  - Easy to determine affected objects
- Extensibility
  - Consumable through PowerShell and 3rd parties
  - SMAPI

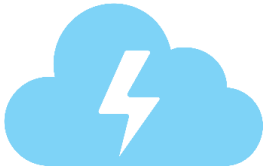


The Need For NVM Express™

Your Stuff Works Better with Intel® SSDs with NVM Express™

Virtualization

NVM Express™ SSDs lower enterprise IT TCO by enabling increased Virtual Machine scalability and optimizing platform utilization

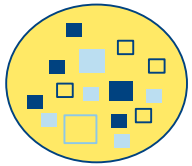


Private Cloud

Software Defined Infrastructure or hyper-convergence is made affordable with high performance SSDs

Database

Consistent, low latency, high bandwidth performance of NVM Express shines in traditional relational databases



Big Data

Analytics and NoSQL databases fully utilize NVM Express performance to provide near real time results

HPC

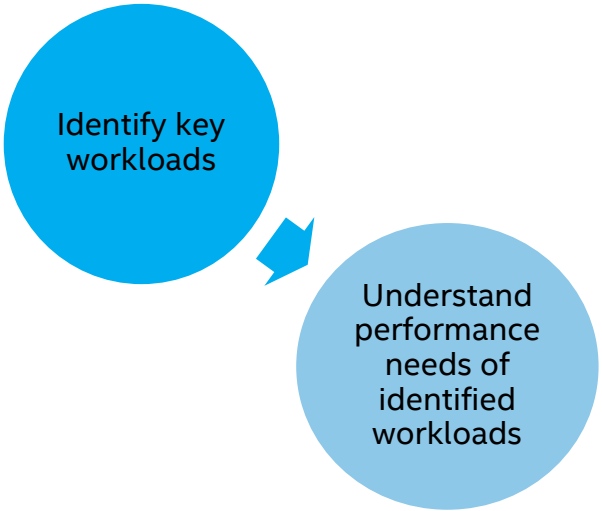
NVM Express keeps up with high bandwidth demands of HPC designed to speed up overall workflow times

Methodology and Performance Results

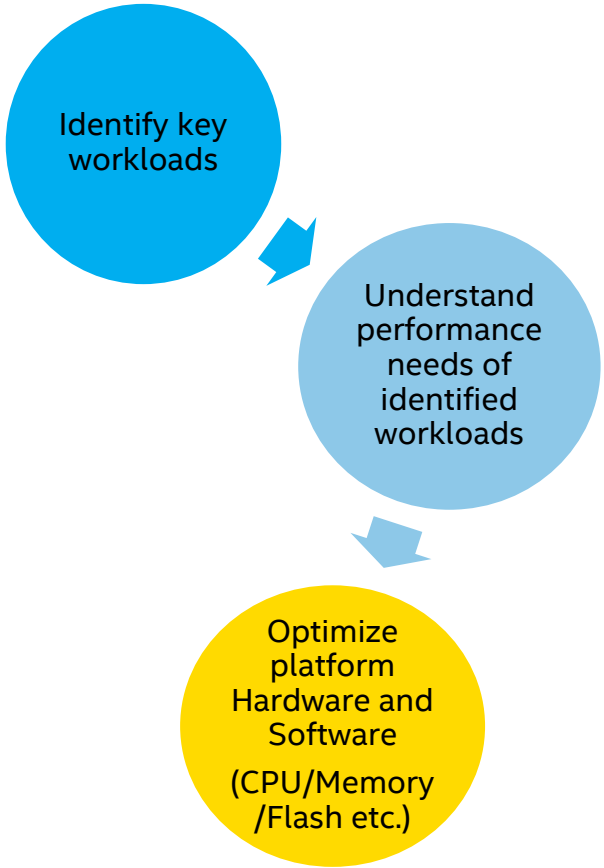
Methodology

Methodology

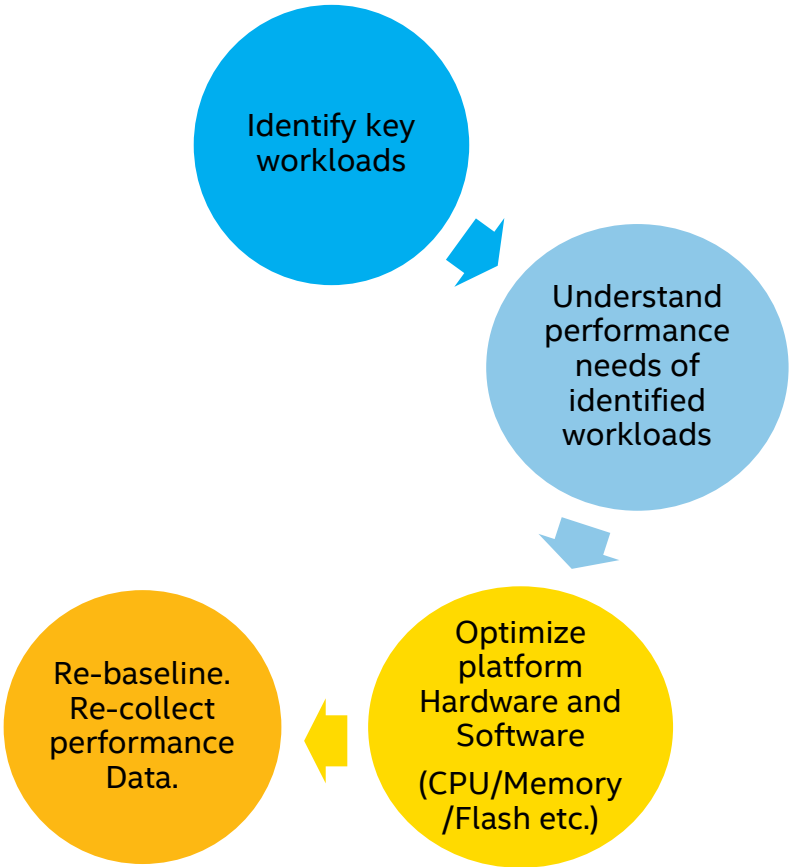
Methodology



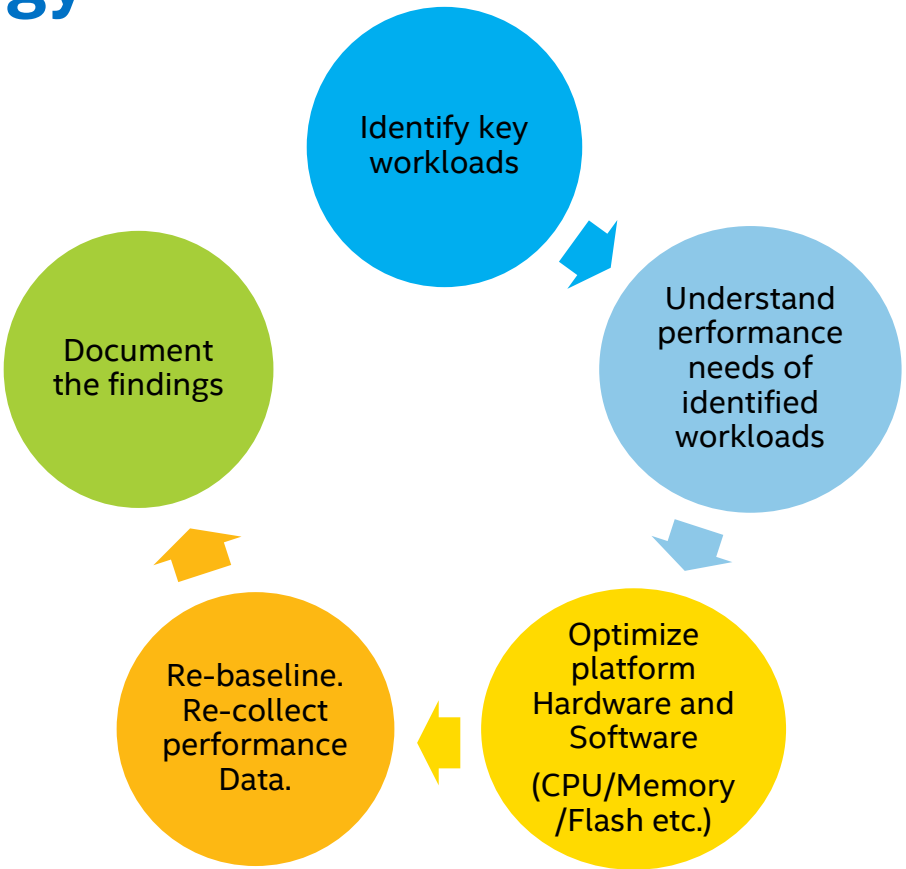
Methodology



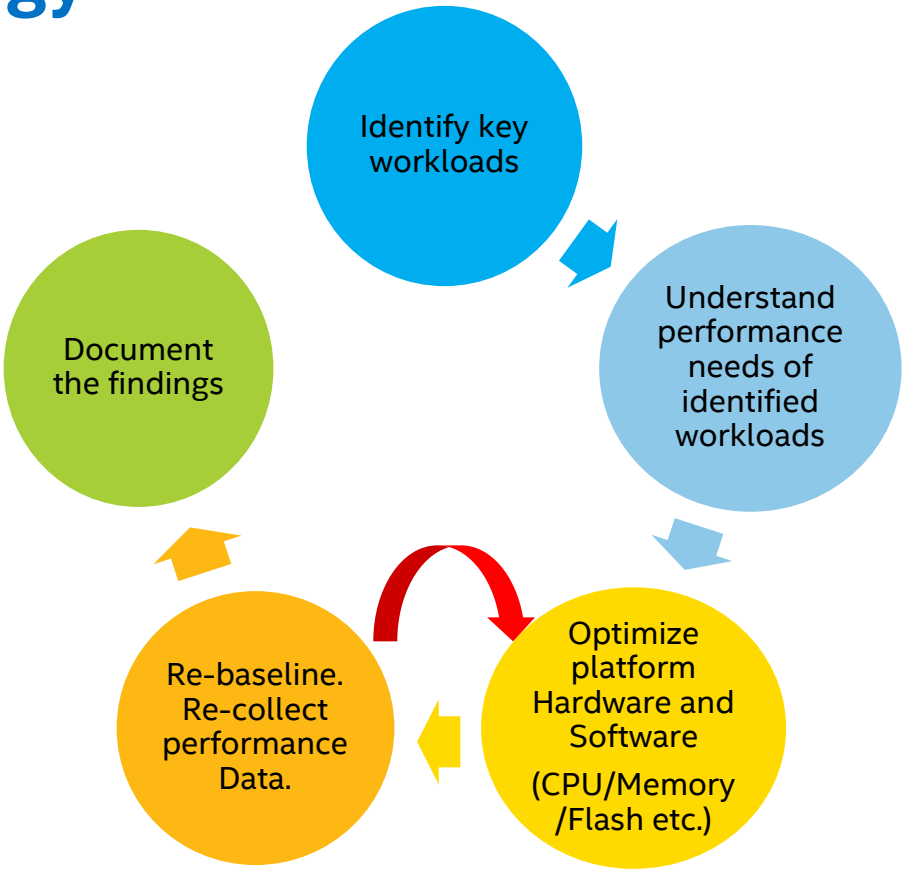
Methodology



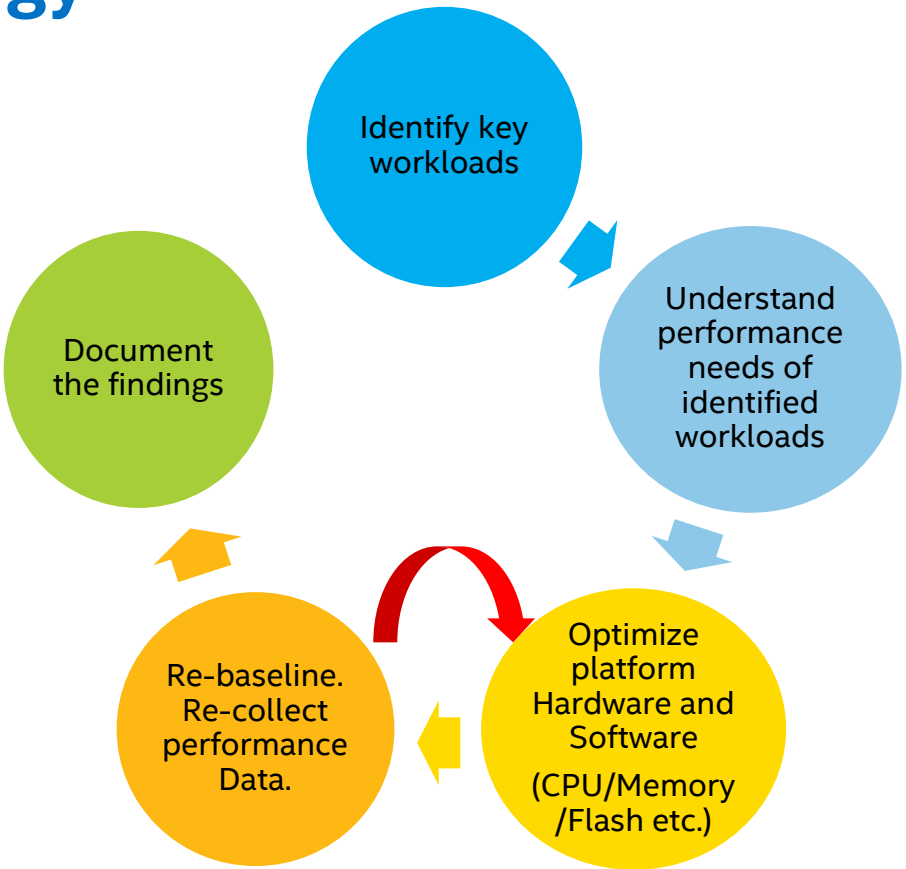
Methodology



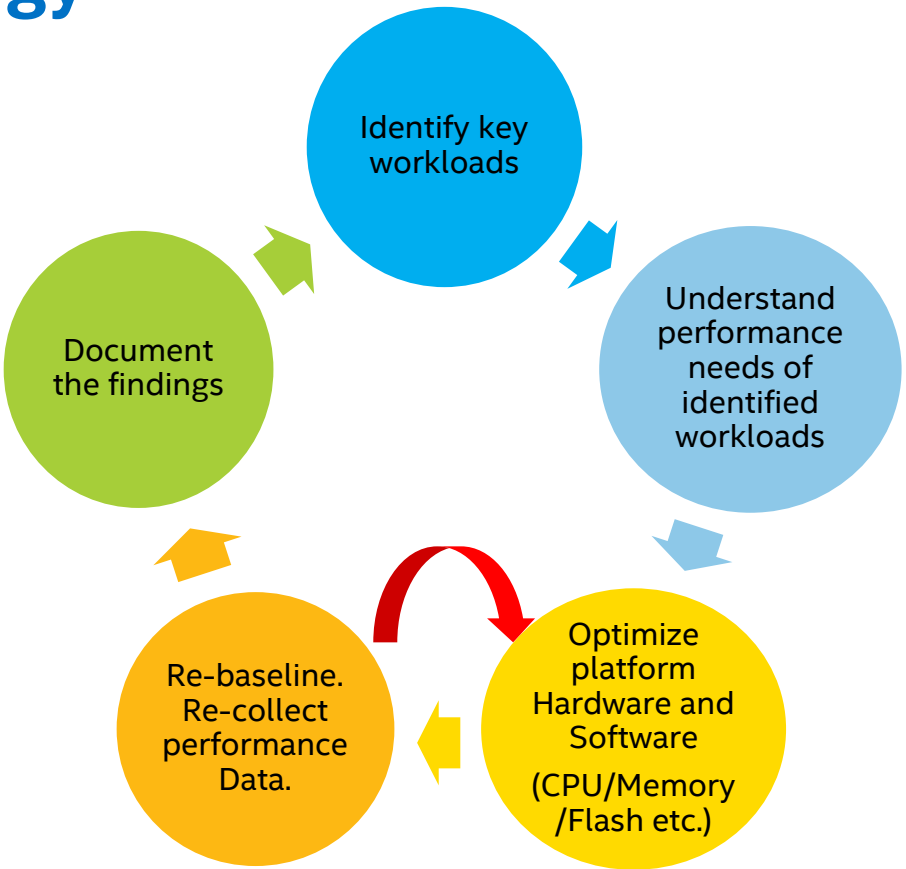
Methodology



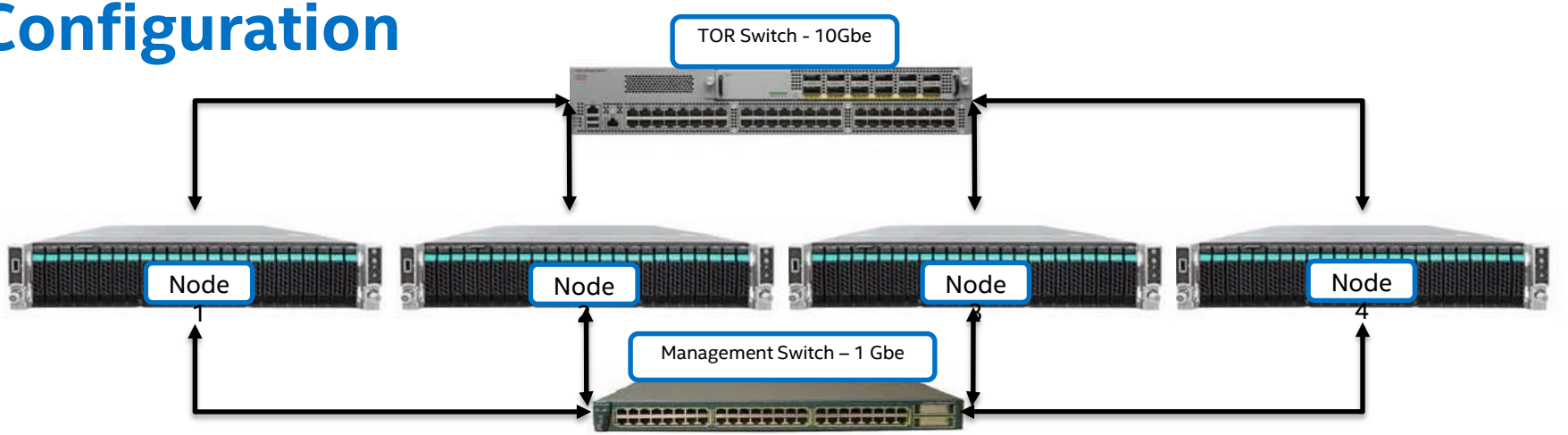
Methodology



Methodology



Configuration



Hardware Configuration

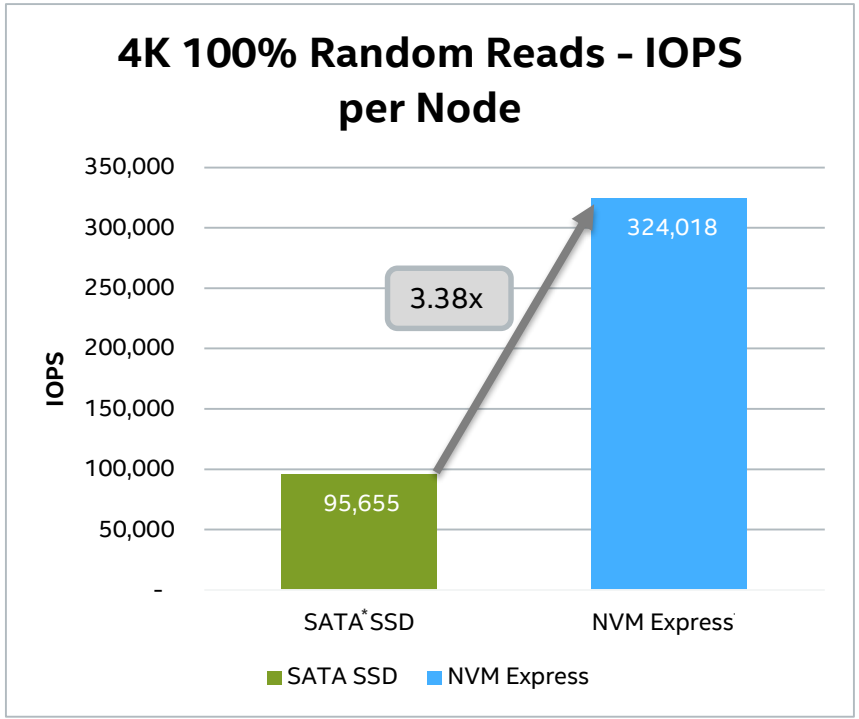
- Intel® Xeon® processor E5-2699 v3 based Server 2U 24x2.5" [R2224WTTYS-IDD]
- 2x10Gbe Chelsio* T520-CR Ethernet Adapter
- Cache:
  - NVM Express™ Config-1x Intel® SSD DC P3700 – 2TB
  - SATA* Config – 2xIntel SSD S3700 – 800GB
- Storage: 5x1TB Seagate* Constellation 2
- TOR Switch-Cisco* Nexus N9K-C93128TX- 10Gbe

Software Configuration

- Windows® Server* 2016 Technical Preview 3†
- Diskspd Version 2.0.15 running on 4x10GB files per share per node
- Storage Spaces Direct with 3-Way mirrored configuration

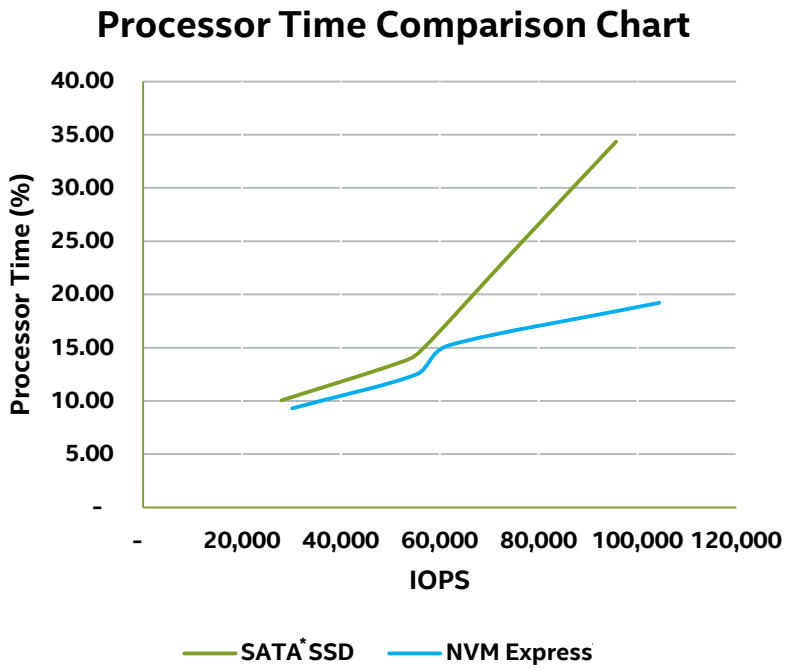
† This technology demonstration uses pre-release software; features and functionality may differ in the final release.

Results –SATA SSD Vs NVM Express™ Max Performance (IOPS, Latency)



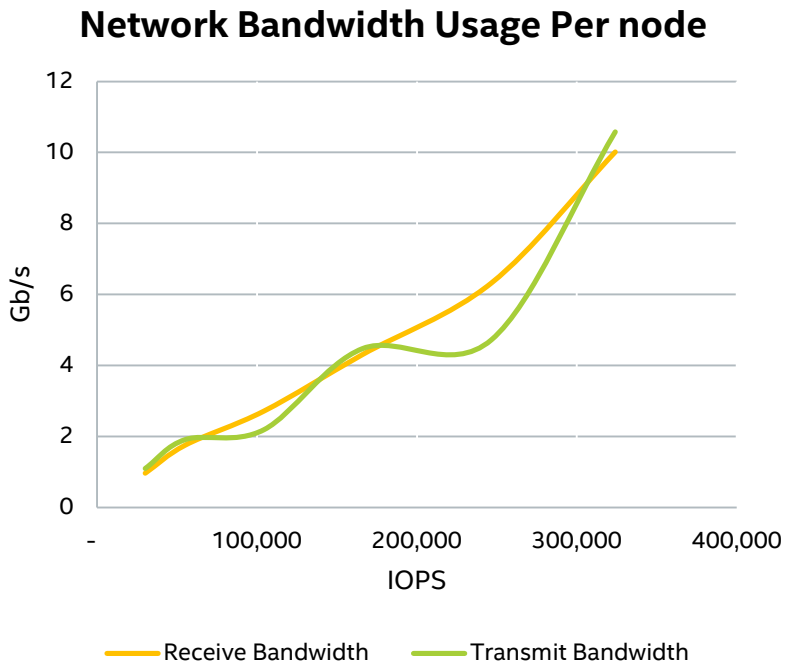
- NVM Express™ Configuration
  - Performance Boost of 3.38x per node
  - Each node delivers similar IOPs (3 Way Mirrored Configuration)
  - Cumulatively 4 nodes deliver 1.25M IOPS!

Per Node Average Processor Time



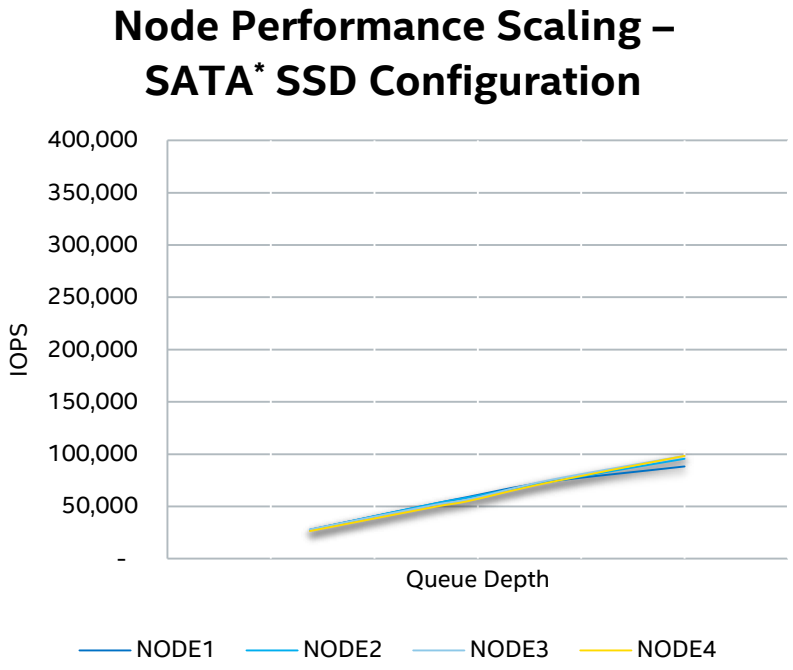
- NVM Express™ Configuration
  - Consumes less CPU resources at the same IOPs (10%-40%)
  - Reflective of reduced stack execution using NVM Express
  - Compute given back to VMs\user

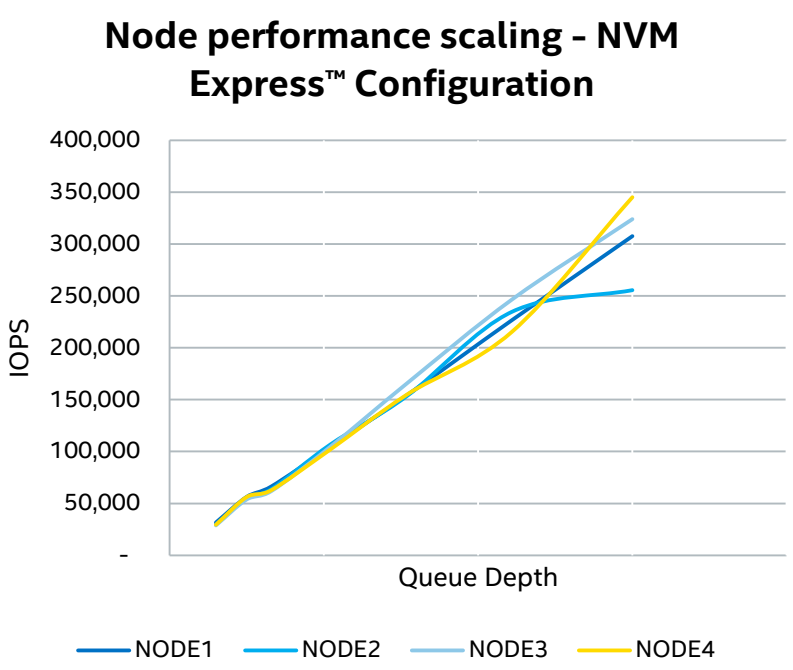
Per Node Network Utilization



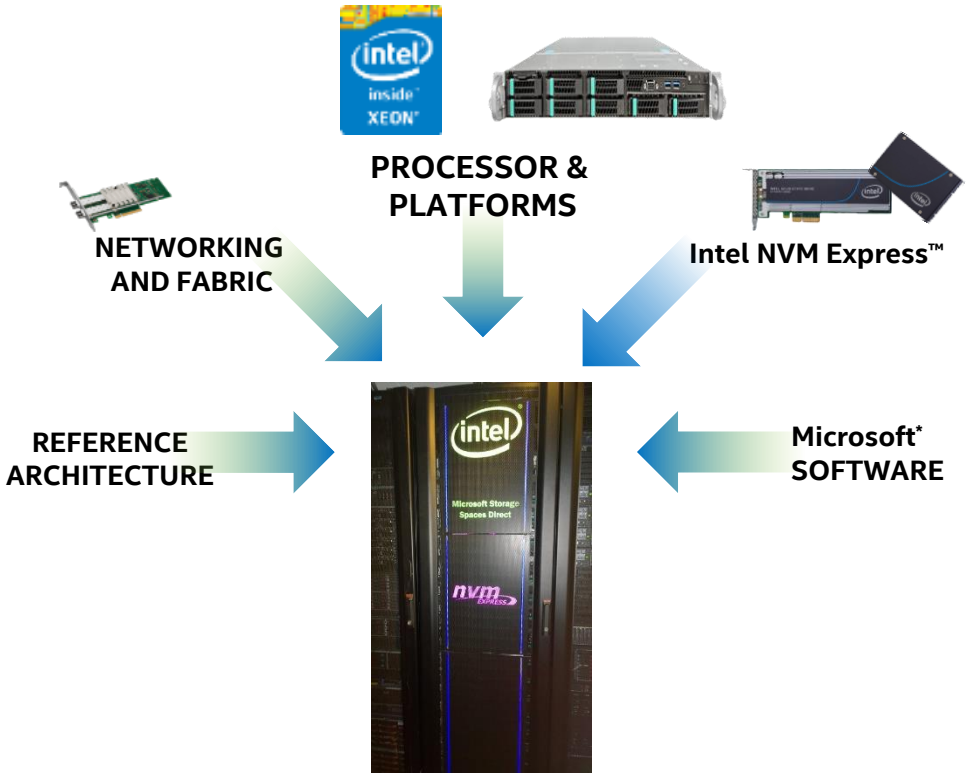
- NVM Express™ Configuration
  - Network Utilization scales with performance
  - No network bottleneck perceived as we scale IOPS
  - We continue to optimize for greater efficiencies

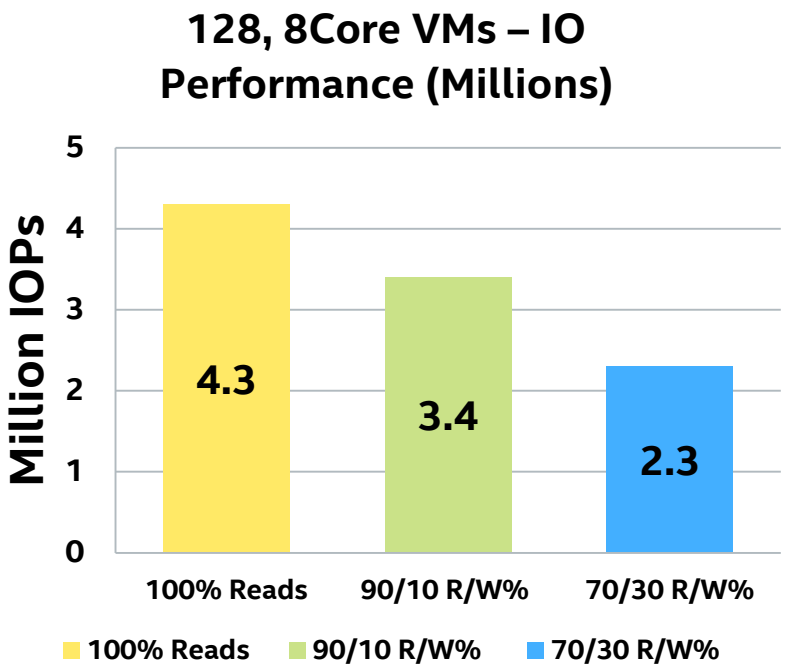
Per Node Performance Scaling





Intel® and Microsoft* Storage Spaces Direct Hyper-Converged Demo





Summary and Next Steps

Summary and Next Steps

- Our optimized NVM Express™ configurations show great results!
  - Greater than 3X performance boost compared to SATA* (similar capacity)
  - Lower CPU time for the same IO rate compared to SATA
- Critical inflection points are impacting storage now more than EVER!
- Non Volatile Memory is MAINSTREAM
- CLOUD usage will continue to drive efficiencies in storage technologies

Additional Sources of Information

- A PDF of this presentation is available from our Technical Session Catalog: www.intel.com/idfsessionsSF. This URL is also printed on the top of Session Agenda Pages in the Pocket Guide.

- More web based information:

http://www.microsoft.com/en-ca/server-cloud/solutions/storage.aspx

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, Xeon, Core, and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the second quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "plans," "believes," "seeks," "estimates," "may," "will," "should" and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel's actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be important factors that could cause actual results to differ materially from the company's expectations. Demand for Intel's products is highly variable and could differ from expectations due to factors including changes in business and economic conditions; consumer confidence or income levels; the introduction, availability and market acceptance of Intel's products, products used together with Intel products and competitors' products; competitive and pricing pressures, including actions taken by competitors; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel's gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; and product manufacturing quality/yields. Variations in gross margin may also be caused by the timing of Intel product introductions and related expenses, including marketing expenses, and Intel's ability to respond quickly to technological developments and to introduce new products or incorporate new features into existing products, which may result in restructuring and asset impairment charges. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Results may also be affected by the formal or informal imposition by countries of new or revised export and/or import and doing-business regulations, which could be changed without prior notice. Intel operates in highly competitive industries and its operations have high costs that are either fixed or difficult to reduce in the short term. The amount, timing and execution of Intel's stock repurchase program could be affected by changes in Intel's priorities for the use of cash, such as operational spending, capital spending, acquisitions, and as a result of changes to Intel's cash flows or changes in tax laws. Product defects or errata (deviations from published specifications) may adversely impact our expenses, revenues and reputation. Intel's results could be affected by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. Intel's results may be affected by the timing of closing of acquisitions, divestitures and other significant transactions. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.

Rev. 4/14/15

Backup

4 Node Configuration Details

- Hardware Configuration:
  - Intel® Xeon® processor E5-2699 v3 based Server 2U 24x2.5" [R2224WTTYS-IDD]
  - 16GB Micron* DIMMs
  - 2x10Gbe Chelsio* Ethernet Adapter
  - Cache:
    - 1xIntel® SSD DC P3700 – 2TB
    - 2xIntel SSD S3700 – 800GB
  - Storage: 1TB Seagate* Constellation 2 ST91000640NS
  - Boot: SSD DC 3710 - 200GB
  - Cisco* Nexus Switch N9K-C93128TX
  - Cisco Catalyst Switch – 3548G – 1Gb
  - 1U Intel Xeon Utility Server

4 Node Configuration Details Continued

- Software Configuration
  - Windows® Server* 2016 Technical Preview 3†
  - Hyper Converged 4 node Configuration
  - Diskspd Version 2.0.15 running on 4x10GB files per share per node
  - Spaces Direct - 3-Way mirrored configuration

† This technology demonstration uses pre-release software; features and functionality may differ in the final release.

Intel and Microsoft* Storage Demo Configuration

- Hyper-Converged using Storage Spaces Direct
- 16 Intel® Server Systems S2600WT
  - Intel Server System R2224WTTYS-IDD (2U)
  - Dual Intel® Xeon® processor E5-2699 v3 Processors
  - 128GB Memory (16GB DDR4-2133 1.2V DR x4 RDIMM)
- Total Raw Capacity for Hyper-V* Cluster: **51.2 TB**
- Data network
  - Chelsio* 10GbE RDMA Card (CHELT520CRG1P10)
- Top of Rack Switch- Cisco* Nexus Switch N9K-C93128TX
- Per Server
  - 4 - Intel® SSD DC P3700 Series (800 GB, 2.5” SFF)
  - 2U NVM Express™ e kit for Wildcat Pass A2U8X25PCIDSPP
  - Boot Drive: Intel SSD DC S3710 Series (200 GB, 2.5” SFF)

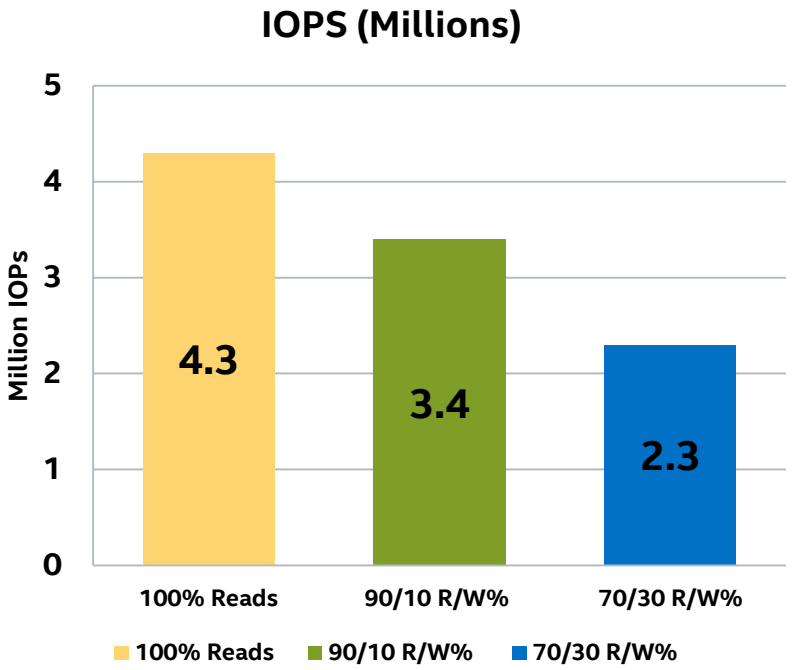
Intel and Microsoft* Storage Demo Results

Spaces Configuration

- Storage Spaces Direct
- Single Storage Pool: 51.2 TB Formatted
- 3-way Mirrored Virtual Disks
- Allocated to Virtual Disks: 44.3 TB
- Free for Rebuild: 6.8 TB
- 16 Virtual Disks for Workload
  - 919 GB each, Total 14.7 TB

Load Generators

- 8 Virtual Machines per Server (128 Total)
  - 8 Virtual Cores, 7.5 GB Memory per VM
  - Equivalent to Azure* A4 sizing
- DISKSPD for load generation
  - 8 threads per DISKSPD instance
  - Queue Depth of 20 per thread

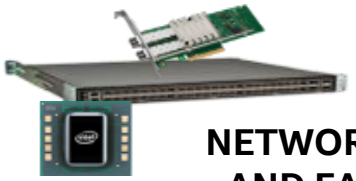


What are Intel's Storage Assets?

Intel® QuickAssist Technology

Intel® Storage Acceleration Library (ISA-L)

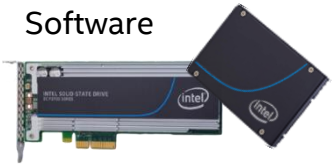
PROCESSOR PLATFORMS



NETWORKING AND FABRIC

SSD'S AND NVM

Intel® Cache Acceleration Software



REFERENCE ARCHITECTURES

Intel® Architecture Storage Solutions

SOFTWARE ECOSYSTEM

Intel® SSD DC P3700 Series

Intel SSD DC P3600 Series

Intel SSD DC P3500 Series

Capacity

Intel® SSD DC P3700 Series	Intel SSD DC P3600 Series	Intel SSD DC P3500 Series
400 GB	400 GB	400 GB
800 GB	800 GB	1.2TB
1.6TB	1.2TB	2TB
2TB	1.6TB	
	2TB	

Endurance

Intel® SSD DC P3700 Series	Intel SSD DC P3600 Series	Intel SSD DC P3500 Series
10 DWPD	3 DWPD	0.3 DWPD
High Endurance Technology	Mixed use	Read Intensive

Performance

	Intel® SSD DC P3700 Series	Intel SSD DC P3600 Series	Intel SSD DC P3500 Series
Random 4k Read	450k IOPS	450k IOPS	450k IOPS
Random 4k Write	175k IOPS	56k IOPS	35k IOPS
Random 4k 70/30 R/W	265k IOPS	160k IOPS	85k IOPS
Sequential Read	2800 MB/s	2600 MB/s	2500 MB/s
Sequential Write	2000 MB/s	1700 MB/s	1700 MB/s

Sequential latency of 20µs

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Configurations: Intel® Core™ i7-3770K CPU @ 3.50GHz, 8GB of system memory, Windows® Server 2012, IOMeter*. Random performance is collected with 4 workers each with 32 QD

Recommendations

- Buy as much NVM Express™ storage as your working set
- Build your configuration for worst-case N/W and storage limits
  - Test with basic disk workloads to gauge sizing (Small IO–> IOPS, Large IO -> Bandwidth
  - N/W - To account for optimal cross-node traffic
- Different Resiliency themes will deliver different results
- Leverage new Microsoft* Storage Spaces Direct features for optimal results
  - Monitoring
  - ReFS