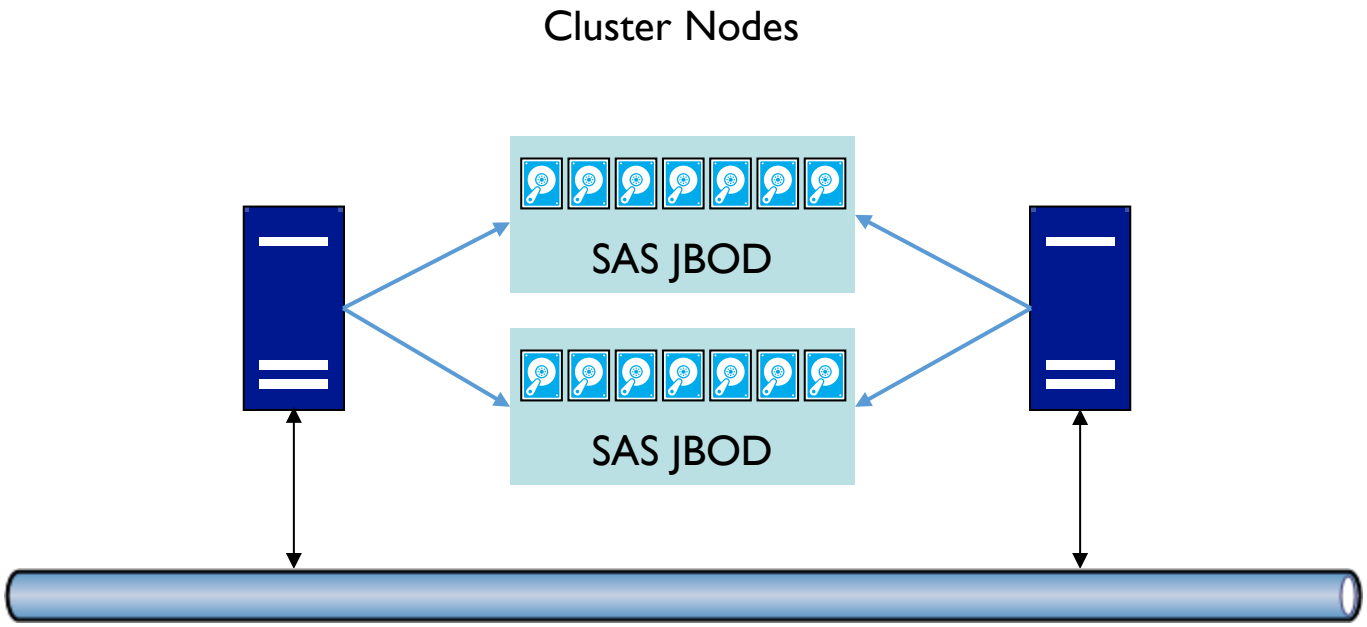
SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2015

Software Defined Storage Based on Direct Attached Storage

Slava Kuznetsov
Microsoft

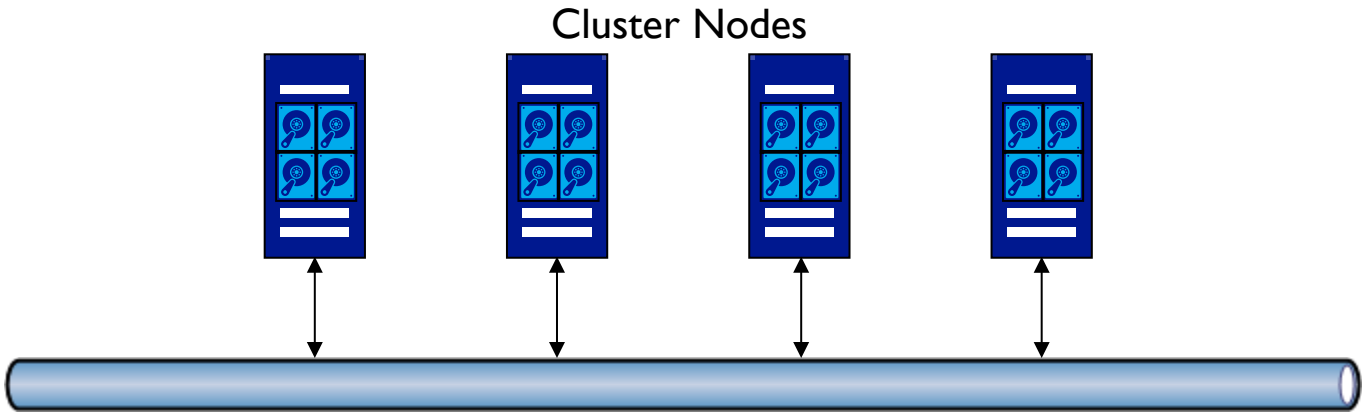
Storage Spaces in Windows Server 2012 R2 with Shared SAS



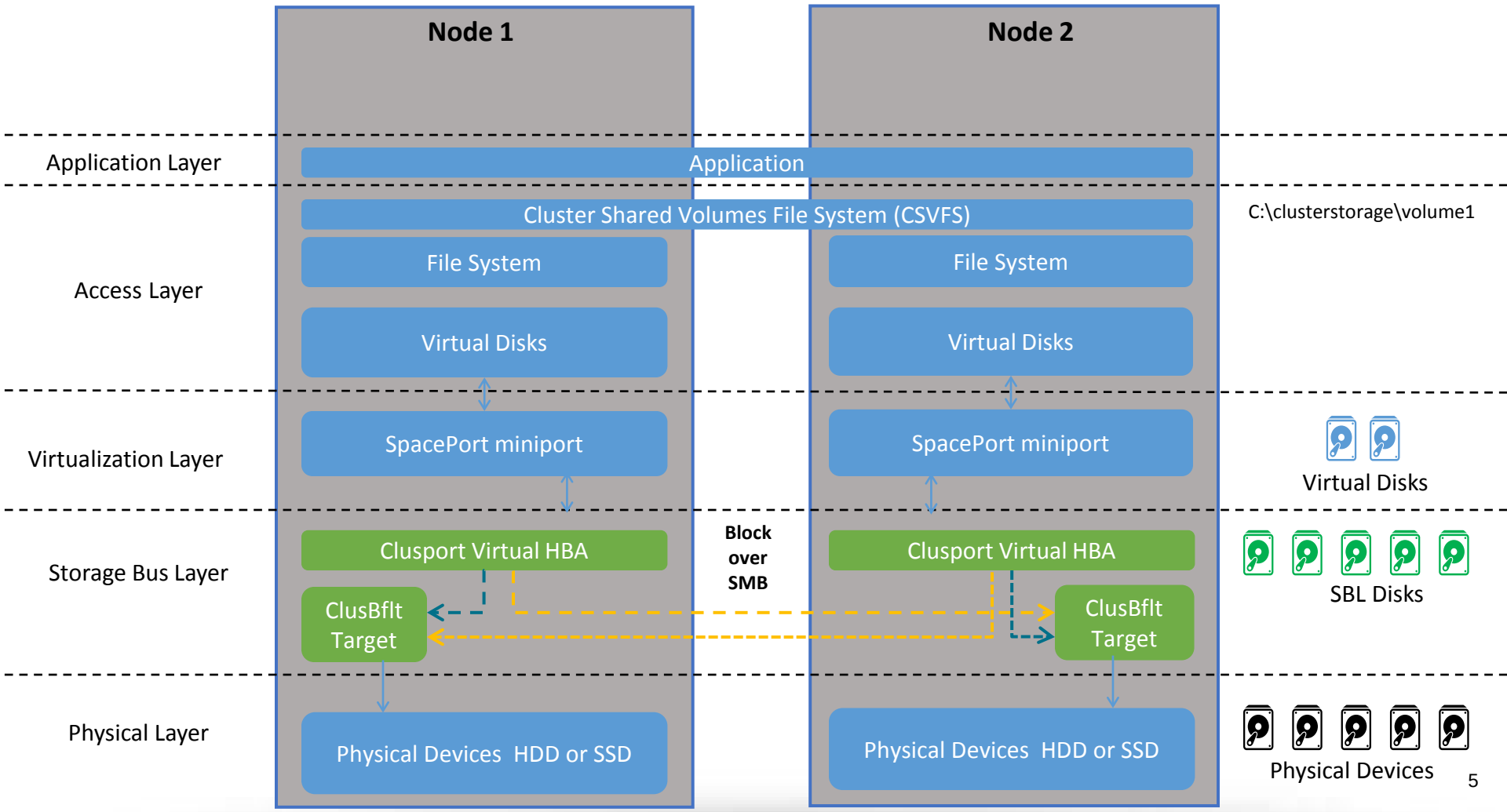
Design Goals for Windows Server 2016

- Lower cost
- Easy to maintain
- Better scale out
- Support wider range of vendors, devices and technologies
- High IOPS, low latency, small CPU overhead

Storage Spaces Direct in Windows Server 2016 with Directly Attached Storage



Storage Bus Architecture



Storage Devices

- Shared
  - SAS HDDs and SSDs with Persistent Reservations (PR) support
- Storage Spaces Direct
  - SAS HDDs and SSDs, PRs not required
  - Lower cost SATA HDDs and SSDs
  - NVMe SSDs for high perf and low latency
  - Forward looking for NVDIMM support

Connectivity

- Shared
  - Requires SAS cabling and shared JBODs, may require SAS switches
- Storage Spaces Direct
  - Uses network, shared JBODs not required
  - Utilizes SMB3 and SMB Direct. TCP or RDMA interconnect for low latency and CPU usage
  - Supports multi-rack storage, not constrained by SAS cabling distance limitations

Storage Bus Bandwidth Management

- Shared
  - Not supported
- Storage Spaces Direct
  - Predictable performance
  - Better performance

Storage Bus Bandwidth Management

- Predictable performance
  - IOs may have different priorities
  - Application IOs have to coexist with system IOs
  - Nodes should have equal access to storage devices
  - SAS and SATA HDDs are optimized for different usage patterns

Storage Bus Bandwidth Management

- Each IO cost is estimated as a sum seek cost + setup cost + data transfer cost
- Supports fair priority scheduling, higher priority bucket leaves 5%..20% to the next pri buckets
- App and system IOs are throttled to 50% / 50%
- IOs from different nodes are throttled to ensure fair access from all nodes

Storage Bus Bandwidth Management

- Better performance
  - HDDs perform 100 times better with sequential IOs vs random 8K IOs
  - Extra seeks consume energy and reduce disk lifetime
  - Devices have limits on max number of pending IOs they can handle 32 for SATA, 64 for SAS

Storage Bus Bandwidth Management

- De-Randomization - discovers sequential/sparse streams and orders IOs to minimize seek and increase performance
- Holds other IOs until sequential stream drains of maximum debt for priority/category/node is reached

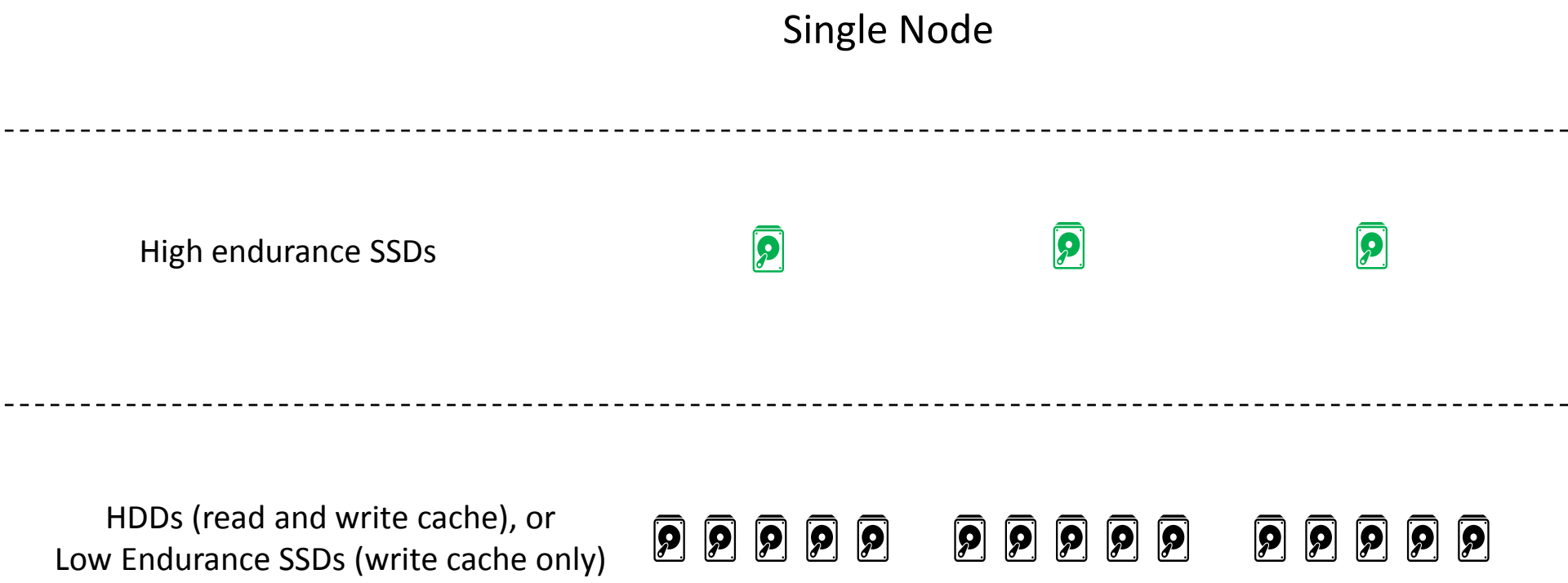
Tiering and caching

- Shared
  - Supports storage tiering at file system level
- Storage Spaces Direct
  - Supports storage bus cache – hybrid storage
    - Low cost per TB, high capacity
    - Number of Device or PCIe slots is limited
    - CPU & Network max out with 1..4 fast flash (NVMe) devices

Storage Bus Cache

- Increase performance of rotational storage by caching small reads and writes on SSDs or other storage devices (NVMEs) with no seek penalty
- Increase durability of low endurance flash devices
- Persistent Read/Write cache
- Keep storage cost low for archival (cold) data
- Adapt fast to changing workloads at high granularity for best cache efficiency (8K page)

Storage Bus Cache



Storage Bus Cache IO types

IOs are classified by size. 64KB or less go through cache, large IO bypass cache.

- Small read
  - Cache hit – read from SSD
  - Cache miss – read ahead HDD populate SSD
- Large read – read HDD, read dirty from SSD
- Small write – write to SSD
- Large Write – write to HDD, purge dirty data on SSD

Storage Bus Cache Destager

- Dirty pages are split into two main categories – Dirty Hot (50%) and Dirty Cold (50%). Only Dirty Cold pages are considered for destage
- Picks area on a HDD with highest density of dirty pages
- Coalesces IOs when possible
- Orders IOs by LBA to achieve maximum throughput

Storage Bus Cache Destager

- Variable destage priority based on % of dirty pages in cache
  - 0% .. 30% - 5% of HDD bandwidth
  - 30% .. 40% - 15% of HDD bandwidth
  - 40% .. 50% - 50% of HDD bandwidth

Storage Bus Cache Performance

- Read Cache Hit: same as SSD
- Read Cache Miss: HDD – 0..5% *
- Small Writes: SSD – 15%
- Destager: x3 .. x25 more IOPS than random
  - Destages mostly when HDDs are not busy
  - Leaves more bandwidth for read misses, large reads and large writes in HDDs
- Rich set of performance counters
  - Cluster Storage Cache Stores
  - Cluster Storage Hybrid Disks

Acknowledgments

- Cluster team
- SMB team
- Storage Spaces team
- File System team
- StorSimple team

Q & A

Storage Spaces Direct feature is available in Windows Server 2016 Technical Preview

Have a question later? Send me e-mail –

slavak@microsoft.com