HyperCore-Direct: NVMe Optimized Hyperconvergence

August 29, 2017

Table of Contents

Overview

Scale Computing's award winning HC3 solution has long been a leader in the hyperconverged infrastructure space. Now targeting even higher performing workloads, Scale Computing is announcing HyperCore-Direct, the first hyper-converged solution to provide software defined block storage utilizing NVMe over fabrics at near bare-metal performance. In this whitepaper, we will showcase the performance of a Scale HyperCore-Direct cluster which has been equipped with Intel® P3700 NVMe drives, as well as a single-node HyperCore-Direct system with Intel® Optane™ P4800X NVMe drives. Various workloads have been tested using off-the-shelf Linux and Windows virtual machine instances. The results show that HyperCore-Direct's new NVMe optimized version of SCRIBE, the same software-defined-storage powering every HC3 cluster in production today, is able to offer the lowest latency per IO delivered to virtual machines.

Specifications

Hardware

The benchmarking in this whitepaper was performed on two separate systems:

System A

- 4x Intel S2600TPR, each with:
- 2x Intel® Xeon® E5-2650v4 @ 2.2GHz 12 core (24 thread)
- 128GiB Memory (8x 16GiB DDR4 2400MHz, 1 DIMM / channel)
- Mellanox ConnectX-5 100GbE RNIC (single port connected)
- 2x Intel® SSD P3700 (2TB)

System B

- 1x Lenovo System x3650, with:
- 2x Intel® Xeon® E5-2690 v4 @ 2.6GHz 14 core (28 thread)
- 128GiB Memory (8x 16GiB DDR4 2400MHz)
- 2x Intel® Optane™ SSD P4800X (375GB)

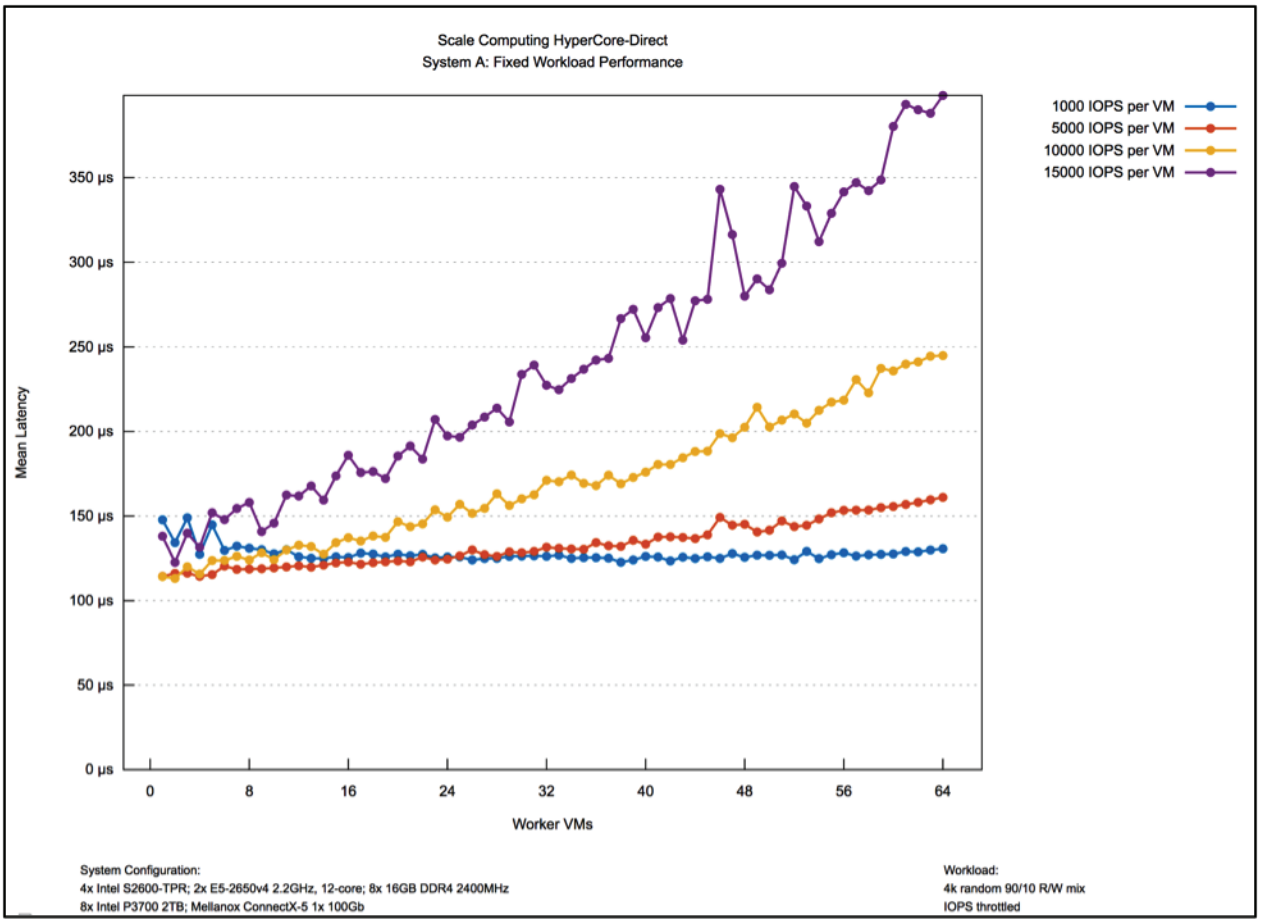
Software

Benchmark workloads were driven using Linux and Microsoft® Windows™ based virtual machines. Purely synthetic benchmark testing was done with fio v2.18 on CentOS 7.2, and IOMeter on Windows. All SCRIBE virtual disks were configured for mirroring with RF=2 (2 copies persisted on independent nodes / storage). System A was configured for a failure domain per node, while System B (a single-node system) was configured for a failure domain per SSD.

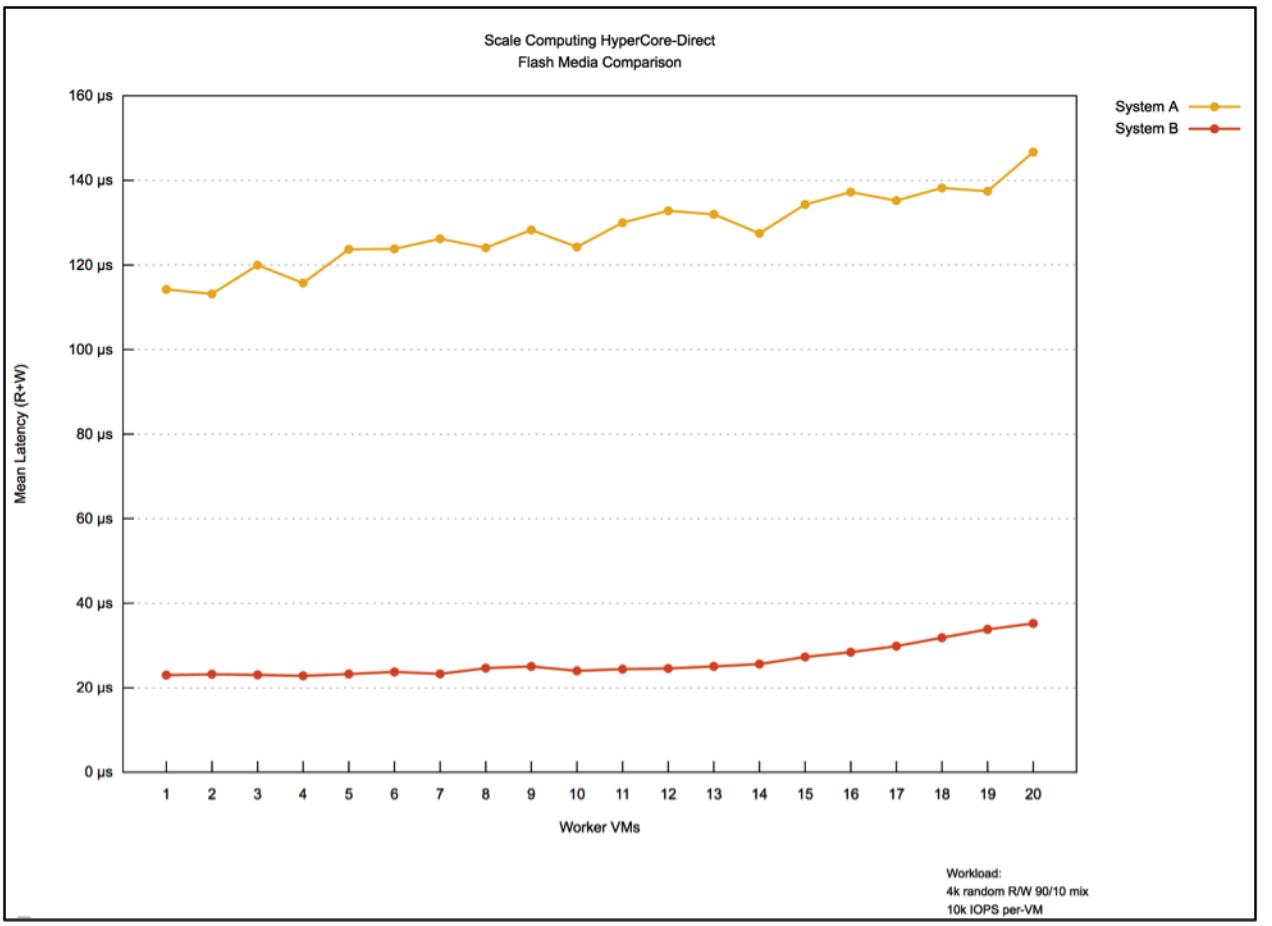
Benchmarks

Consistent Performance

HyperCore-Direct exhibits consistent performance on mixed read/write workloads as new virtual machines are added. For example, System A running 64 VMs, each VM performing 10,000 IOPS of 90 percent read, 10 percent write random 4k IO never exceeds 250µs (µs = microsecond) mean latency per I/O.



Comparing System A to System B, the difference in the speeds of the flash media (NAND vs 3D XPoint) becomes very apparent, and the ability of HyperCore-Direct to allow a virtualized workload to experience the low latency of the Intel® Optane™ SSD is demonstrated. The graph below compares just the 10K IOPS workload between the two types of media, with a VM count from 1 to 20:

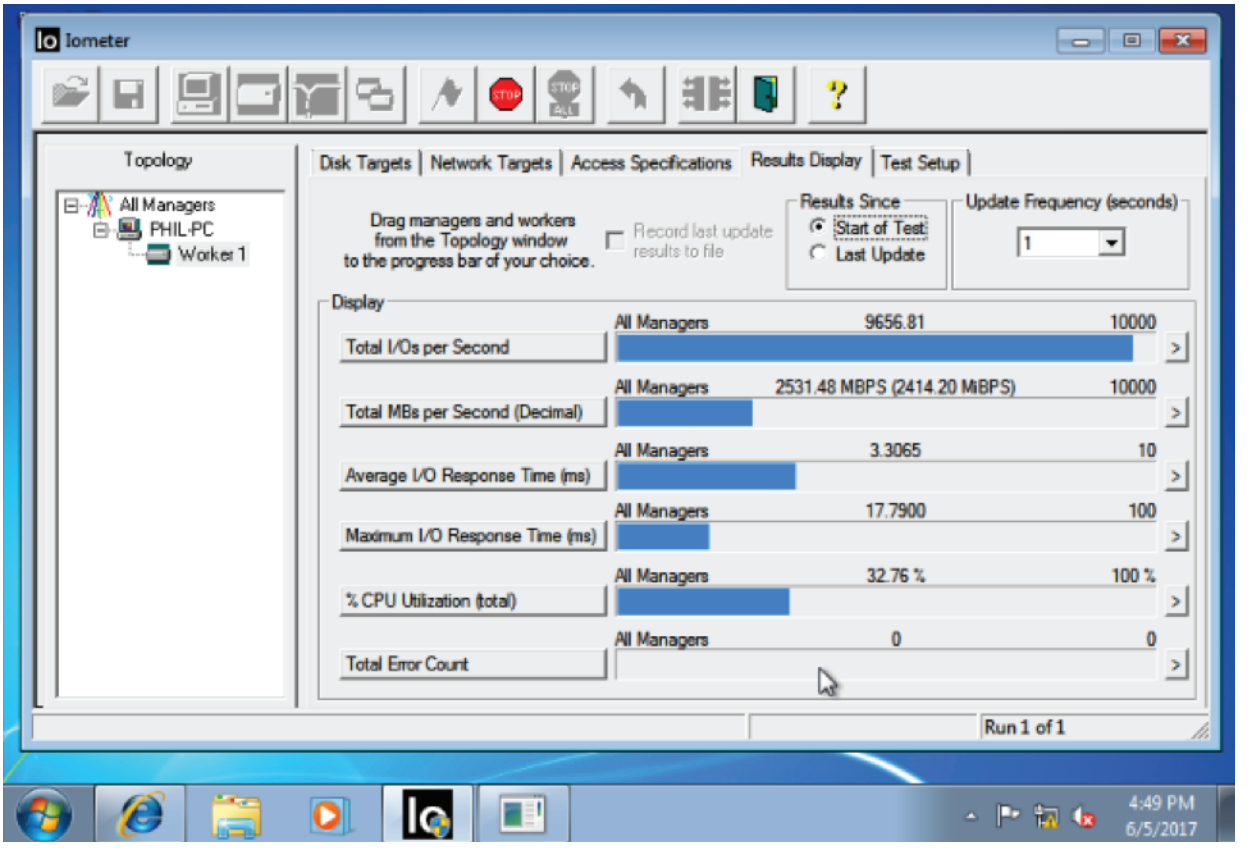


Mega-IOPS with Minimum Effort

For a hyperconverged solution employing resource sharing between the software-defined storage and virtualized workloads, CPU headroom under load is a defining factor. SCRIBE's data path is highly optimized allowing guest virtual machines to communicate directly with SCRIBE (using standard virtio compatible drivers) without involving a hypervisor in any way. This significantly reduces CPU bottlenecks, context switches, and system calls; all of which would add latency to each IO. At maximum IO rates (over 2.9 million IOPS on HyperCore-Direct system A), overall host CPU consumption is below 25%, with SCRIBE itself consuming only a portion of that.

Stunning Bandwidth From Within Virtual Machines

SCRIBE's zero-copy data path allows virtual machines to achieve high IO bandwidth without taxing the CPU. To demonstrate how even a legacy workload might take advantage of this, we have tested a Microsoft Windows 7 virtual machine running IOMeter 1.1. This virtual machine was deployed on HyperCore-Direct System A. With a random read workload and a block size of 256KiB, we obtained a bandwidth of 2.53GB/s.



Low Latency Pushes Workloads to New Speeds

Most organizations still utilize legacy software in many areas, software which is not always optimized for modern multi-core CPUs and multi-queue storage devices. Improving the performance of these workloads can be challenging. IO latency is often a key factor here, and the extreme low latency that SCRIBE and NVMe are able to provide is a huge advantage for these types of workloads. With legacy single queue depth workloads enjoying mean write latencies below 50µs, serialized single-threaded operations become blazing fast.

Here are some sample single queue depth workloads using fio in CentOS 7.2 Linux virtual machines:

System A: Single queue depth writes

```
[root@localhost ~]# fio --name=test --ioengine=libaio --direct=1 --rw=randwrite --io-
depth=1 --bs=4k --filename=/dev/vda --runtime=60 --time_based=1
test: (g=0): rw=randwrite, bs=4096B-4096B,4096B-4096B,4096B-4096B, ioengine=libaio,
iodepth=1
fio-2.18
Starting 1 process
Jobs: 1 (f=1): [w(1)][100.0%][r=0KiB/s,w=90.1MiB/s][r=0,w=23.3k IOPS][eta 00m:00s]
test: (groupid=0, jobs=1): err= 0: pid=21175: Tue Aug 29 05:52:14 2017
  write: IOPS=21.3k, BW=83.2MiB/s (87.2MB/s)(4986MiB/60001msec)
    slat (usec): min=2, max=39, avg= 3.50, stdev= 1.26
    clat (usec): min=25, max=12881, avg=41.50, stdev=24.95
     lat (usec): min=33, max=12884, avg=45.00, stdev=25.21
    clat percentiles (usec):
     |  1.00th=[   35],  5.00th=[   36], 10.00th=[   37], 20.00th=[   37],
     | 30.00th=[   38], 40.00th=[   38], 50.00th=[   39], 60.00th=[   39],
     | 70.00th=[   40], 80.00th=[   50], 90.00th=[   52], 95.00th=[   53],
     | 99.00th=[   56], 99.50th=[   58], 99.90th=[   68], 99.95th=[   78],
     | 99.99th=[  117]
    lat (usec) : 50=78.45%, 100=21.54%, 250=0.02%, 500=0.01%, 750=0.01%
    lat (msec) : 2=0.01%, 4=0.01%, 10=0.01%, 20=0.01%
  cpu          : usr=6.98%, sys=16.20%, ctx=1276533, majf=0, minf=34
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwt: total=0,1276531,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=83.2MiB/s (87.2MB/s), 83.2MiB/s-83.2MiB/s (87.2MB/s-87.2MB/s), io=4986MiB
(5229MB), run=60001-60001msec

Disk stats (read/write):
  vda: ios=116/1274172, merge=0/0, ticks=64/47529, in_queue=47294, util=78.89%
```

Notice the mean latency of 45µs, a P99 latency of 56µs, and P99.99 of 117µs. These numbers include writes to both replicas (RF=2). Remote IOs in HyperCore-Direct utilize NVMe over fabrics, taking full advantage of the low latency of RDMA via the Mellanox ConnectX®-5 RNIC.

When Intel® Optane™ media is involved, the latency is reduced significantly:

System B: Single queue depth writes

```
[root@localhost ~]# fio --name=test --ioengine=libaio --direct=1 --rw=randwrite --io-
depth=1 --bs=4k --filename=/dev/vda --runtime=60 --time_based=1
test: (g=0): rw=randwrite, bs=4096B-4096B,4096B-4096B,4096B-4096B, ioengine=libaio,
iodepth=1
fio-2.18
Starting 1 process
Jobs: 1 (f=1): [w(1)][100.0%][r=0KiB/s,w=143MiB/s][r=0,w=36.6k IOPS][eta 00m:00s]
test: (groupid=0, jobs=1): err= 0: pid=31934: Tue Aug 29 06:04:56 2017
  write: IOPS=36.8k, BW=144MiB/s (151MB/s)(8621MiB/60001msec)
    slat (usec): min=1, max=49, avg= 1.99, stdev= 0.39
    clat (usec): min=0, max=12639, avg=24.21, stdev= 9.35
     lat (usec): min=23, max=12641, avg=26.20, stdev= 9.36
    clat percentiles (usec):
     |  1.00th=[   22],  5.00th=[   23], 10.00th=[   23], 20.00th=[   23],
     | 30.00th=[   23], 40.00th=[   23], 50.00th=[   24], 60.00th=[   24],
     | 70.00th=[   24], 80.00th=[   24], 90.00th=[   25], 95.00th=[   26],
     | 99.00th=[   44], 99.50th=[   45], 99.90th=[   52], 99.95th=[   58],
     | 99.99th=[   72]
    lat (usec) : 2=0.01%, 20=0.02%, 50=99.84%, 100=0.13%, 250=0.01%
    lat (usec) : 500=0.01%, 750=0.01%
    lat (msec) : 2=0.01%, 20=0.01%
  cpu          : usr=9.69%, sys=12.65%, ctx=2206963, majf=0, minf=34
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwt: total=0,2206961,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=144MiB/s (151MB/s), 144MiB/s-144MiB/s (151MB/s-151MB/s), io=8621MiB
(9040MB), run=60001-60001msec

Disk stats (read/write):
  vda: ios=116/2203190, merge=0/0, ticks=60/48735, in_queue=48575, util=81.03%
```

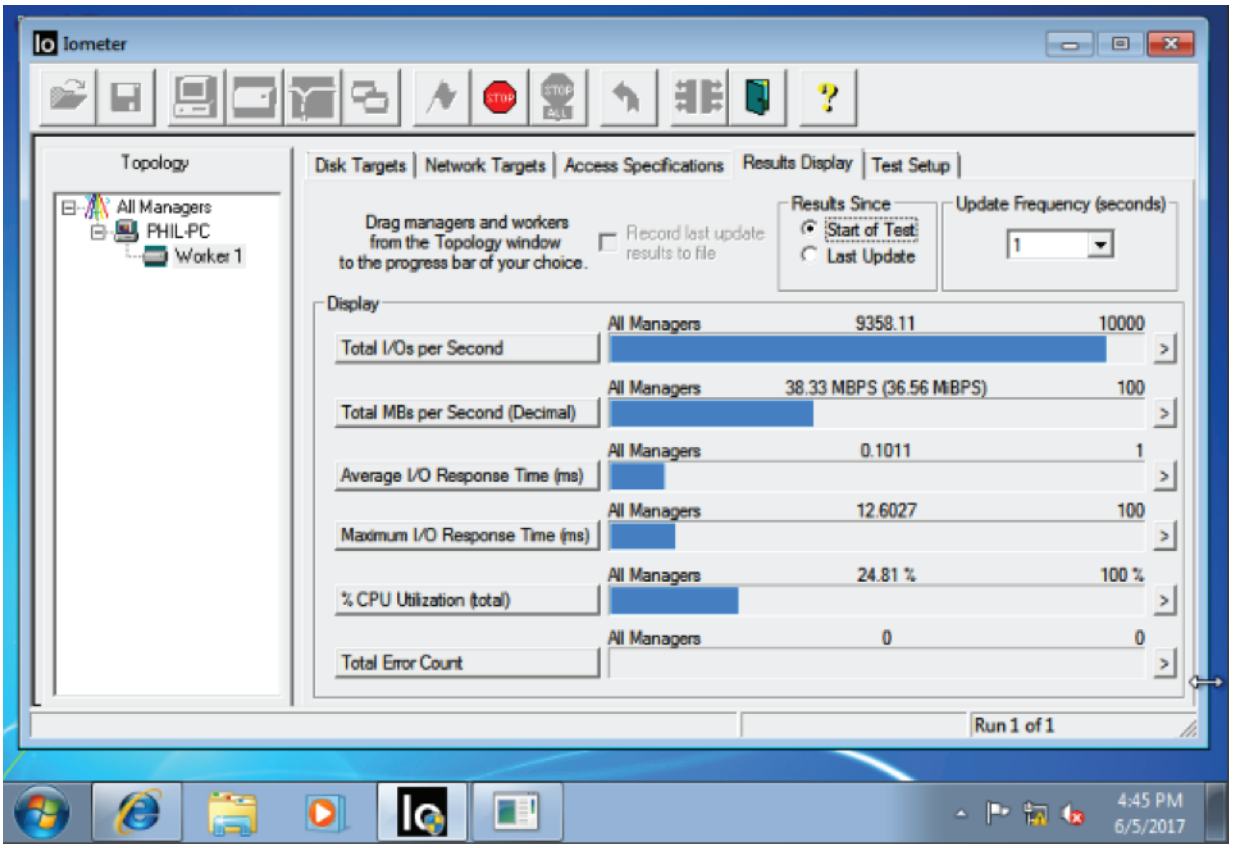
System B: Single queue depth reads

```
[root@localhost ~]# fio --name=test --ioengine=libaio --direct=1 --rw=randread --io-
depth=1 --bs=4k --filename=/dev/vda --runtime=60 --time_based=1
test: (g=0): rw=randread, bs=4096B-4096B,4096B-4096B,4096B-4096B, ioengine=libaio,
iodepth=1
fio-2.18
Starting 1 process
Jobs: 1 (f=1): [r(1)][100.0%][r=186MiB/s,w=0KiB/s][r=47.7k,w=0 IOPS][eta 00m:00s]
test: (groupid=0, jobs=1): err= 0: pid=31939: Tue Aug 29 06:08:37 2017
   read: IOPS=47.4k, BW=185MiB/s (194MB/s)(10.9GiB/60001msec)
    slat (usec): min=1, max=32, avg= 1.94, stdev= 0.45
    clat (usec): min=8, max=12625, avg=18.20, stdev=10.50
     lat (usec): min=18, max=12627, avg=20.14, stdev=10.51
    clat percentiles (usec):
     |  1.00th=[   16],  5.00th=[   17], 10.00th=[   17], 20.00th=[   18],
     | 30.00th=[   18], 40.00th=[   18], 50.00th=[   18], 60.00th=[   18],
     | 70.00th=[   18], 80.00th=[   18], 90.00th=[   19], 95.00th=[   19],
     | 99.00th=[   22], 99.50th=[   23], 99.90th=[   39], 99.95th=[   43],
     | 99.99th=[   62]
    lat (usec) : 10=0.01%, 20=96.71%, 50=3.27%, 100=0.02%, 250=0.01%
    lat (usec) : 500=0.01%
    lat (msec) : 2=0.01%, 20=0.01%
  cpu          : usr=12.28%, sys=16.27%, ctx=2839370, majf=0, minf=36
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwt: total=2839364,0,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
   READ: bw=185MiB/s (194MB/s), 185MiB/s-185MiB/s (194MB/s-194MB/s), io=10.9GiB
(11.7GB), run=60001-60001msec

Disk stats (read/write):
  vda: ios=2834583/0, merge=0/0, ticks=44509/0, in_queue=44252, util=73.90%
```

Legacy workloads may also benefit significantly from HyperCore-Direct. For example, Windows 7 running IOMeter on HyperCore-Direct System A is able to obtain average latencies of 101µs for a 4k random read workload with a queue depth of 1.



Conclusion

With performance approaching what is possible on non-redundant bare metal systems, combined with low CPU overhead, HyperCore-Direct allows efficient use of NVMe and NVMe over fabrics in a fully redundant, distributed hyperconverged solution. Linux and Windows virtual machines are able to easily consume the advanced SCRIBE data services provided by HyperCore-Direct without any software modifications.

The blazing fast speeds that NVMe can provide will enable IT shops with exceptional performance needs to implement easy-to-use hyperconverged infrastructure for even their most data intensive workloads with HyperCore-Direct. With HC3 continuing to meet the needs of standard IT infrastructures, HyperCore-Direct will fill the needs of select IT operations requiring speeds that only NVMe and optimized, near bare metal data pathing can provide.

Stay tuned to *www.scalecomputing.com/nvme* for additional whitepapers, demo videos, and future announcements.

SCALE COMPUTING

Corporate Headquarters
525 S. Meridian Street
Indianapolis, IN 46225
P. +1 317-856-9959

www.scalecomputing.com

West Coast Office
360 Ritch Street
Suite 300
San Francisco, CA 94107

1-877-SCALE-59 (877-722-5359)

EMEA Office
Saunders House
52-53 The Mall
London
W5 3TA
United Kingdom