StorPool
DISTRIBUTED STORAGE

# StorPool Storage

*#SFD18, @storpool*

# Introduction to StorPool

*Boyan Ivanov, co-founder & CEO*
*#SFD18, @storpool*

# Best-of-breed block storage software (1)

1. Software company scale-out, block storage software
    a. Primary, flash (SATA/NVMe)
2. Not your typical Valley startup
3. Doing this before SDS/SDN/SDDC & ***MARKETING-DEFINED STORAGE***
4. Delivered as a working storage solution on customer's hardware:
    a. Fully managed: software + 24/7/365 support, SLA, proactive monitoring
    b. Hardware Compatibility List (HCL) or
    c. A pre-integrated solution with partners
5. SDS 2.0 - feature rich shared storage system faster than local SSD

StorPool
DISTRIBUTED STORAGE

# Best-of-breed block storage software (2)

6. Developed from scratch:
    a. Own on-disk format, protocol, quorum, client, etc, etc.
    b. Fully distributed, scale-out, online changes of everything, etc.
    c. Running in production for 6+ years; numerous 1PB+ flash systems; 17 major releases; Global spread of customers
7. Target customers - companies building public & private clouds:
    a. Service providers & public clouds
    b. Enterprises & various private clouds
8. Use cases - anything block - DBs, VM disks, VDI, etc.
9. Replacing single-purpose SAN / AFA or other storage software

StorPool DISTRIBUTED STORAGE

# Best-of-breed block storage software - Example:

**NVMe shared storage system with:**

1. **Latency**: < 100 µs
2. **Throughput**: >1M IOPS per server (scale-out, 10 servers = 10M IOPS @ ~ 300µs)
3. **Feature rich**: API, end-to-end data integrity, self-healing, online everything, thin provisioning, snaps & clones, QoS, backup & DR, etc.
4. **Fully managed**: 24/7 support; SLA; proactive monitoring & issue resolution
5. **TCO over 3 years**: < $0.10 /GB provisioned /month; < $0.002 /IOPS/month

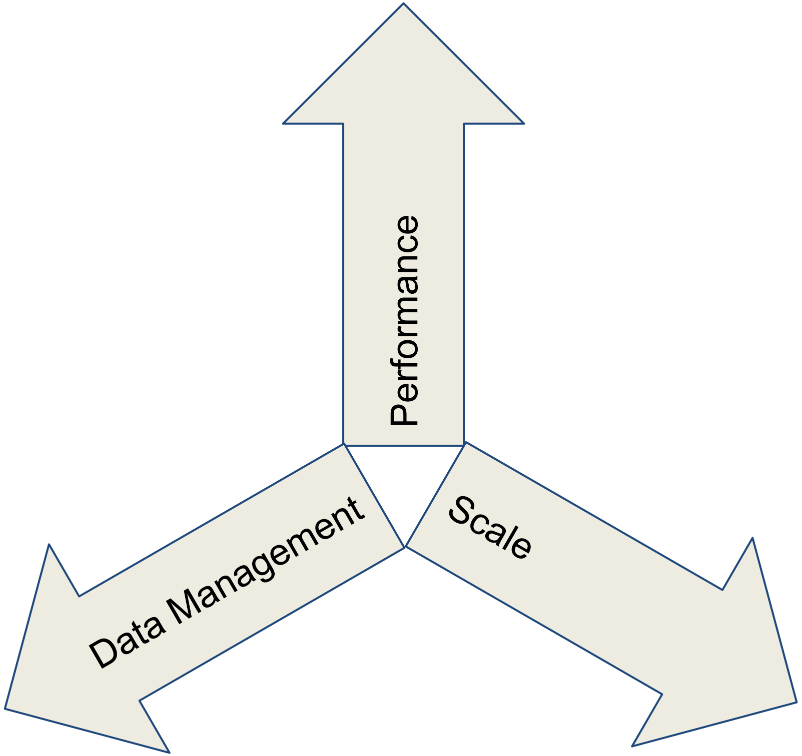
StorPool DISTRIBUTED STORAGE

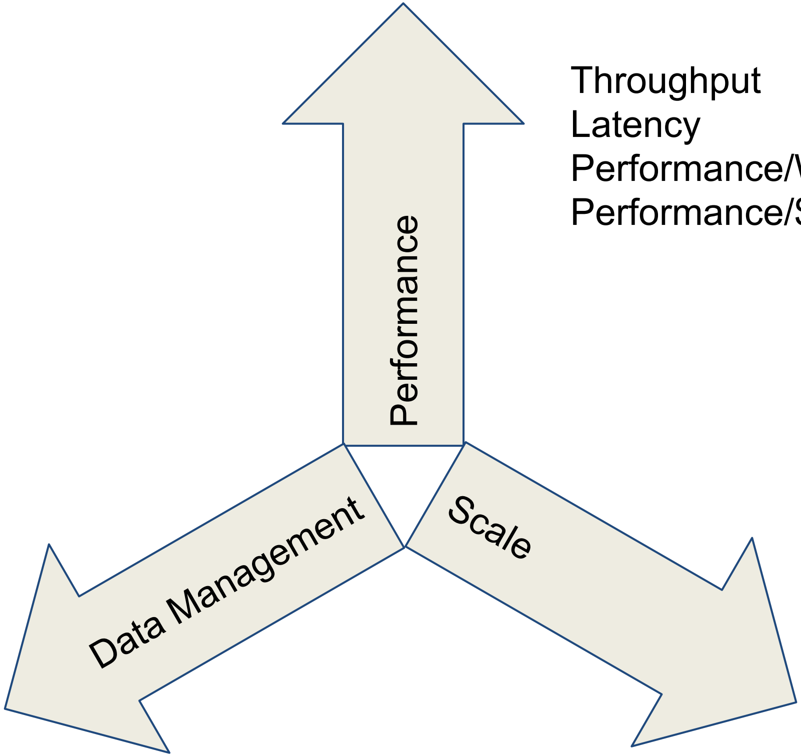
StorPool
DISTRIBUTED STORAGE

# You can't have your cake and eat it

or can you?

*Boyan Krosnov, Co-founder and CPO*
*#SFD18, @storpool*

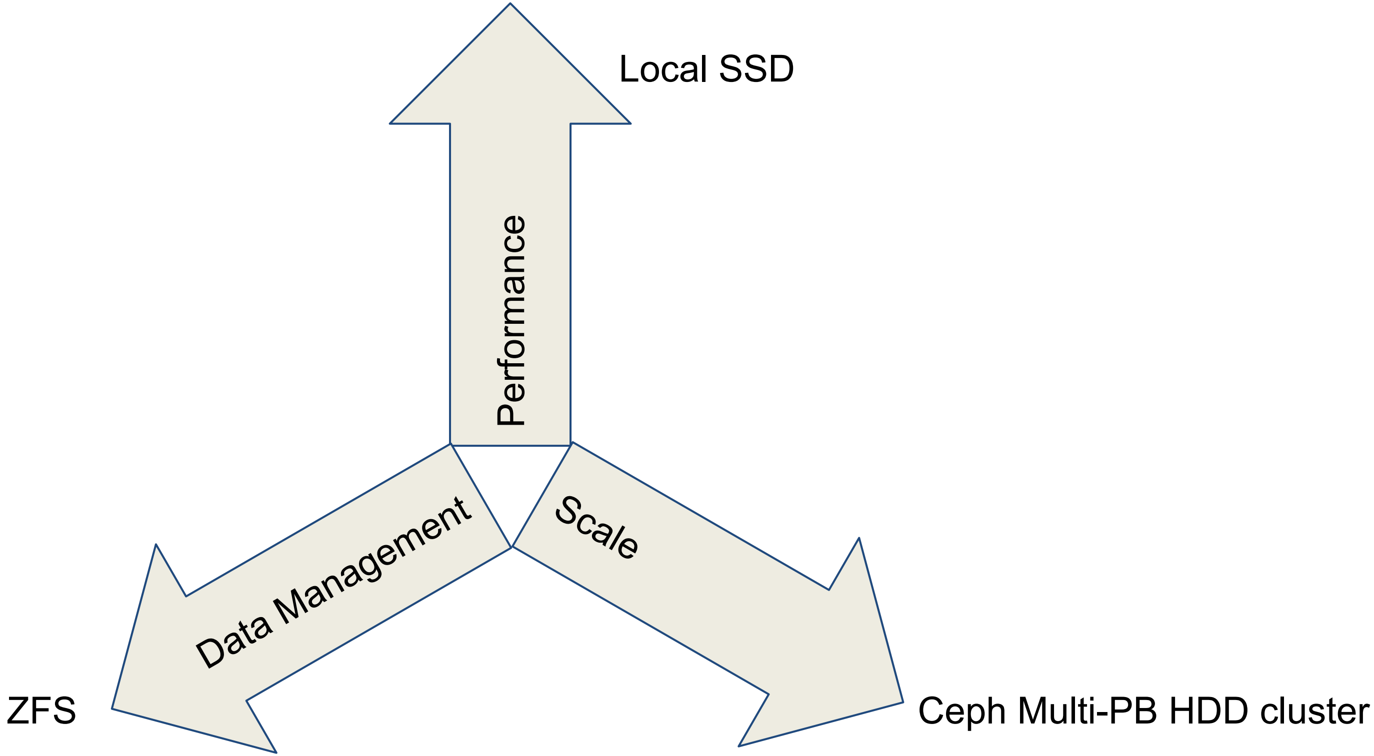

Performance
Data Management
Scale




Throughput
Latency
Performance/Watt,
Performance/$/GB

API-driven, integrations
Scale by adding nodes/drives
Pooling of capacity & performance
Metrics collection & Monitoring
Deployment automation

End-to-end data integrity guarantee
"LUN" per vDisk
CoW - Thin provisioning, Snapshots, Clones
Multi-site with efficient transport of changes
Fast recovery (changed block tracking)




StorPool
DISTRIBUTED STORAGE

# Why performance

Fast storage system = more work done per CPU

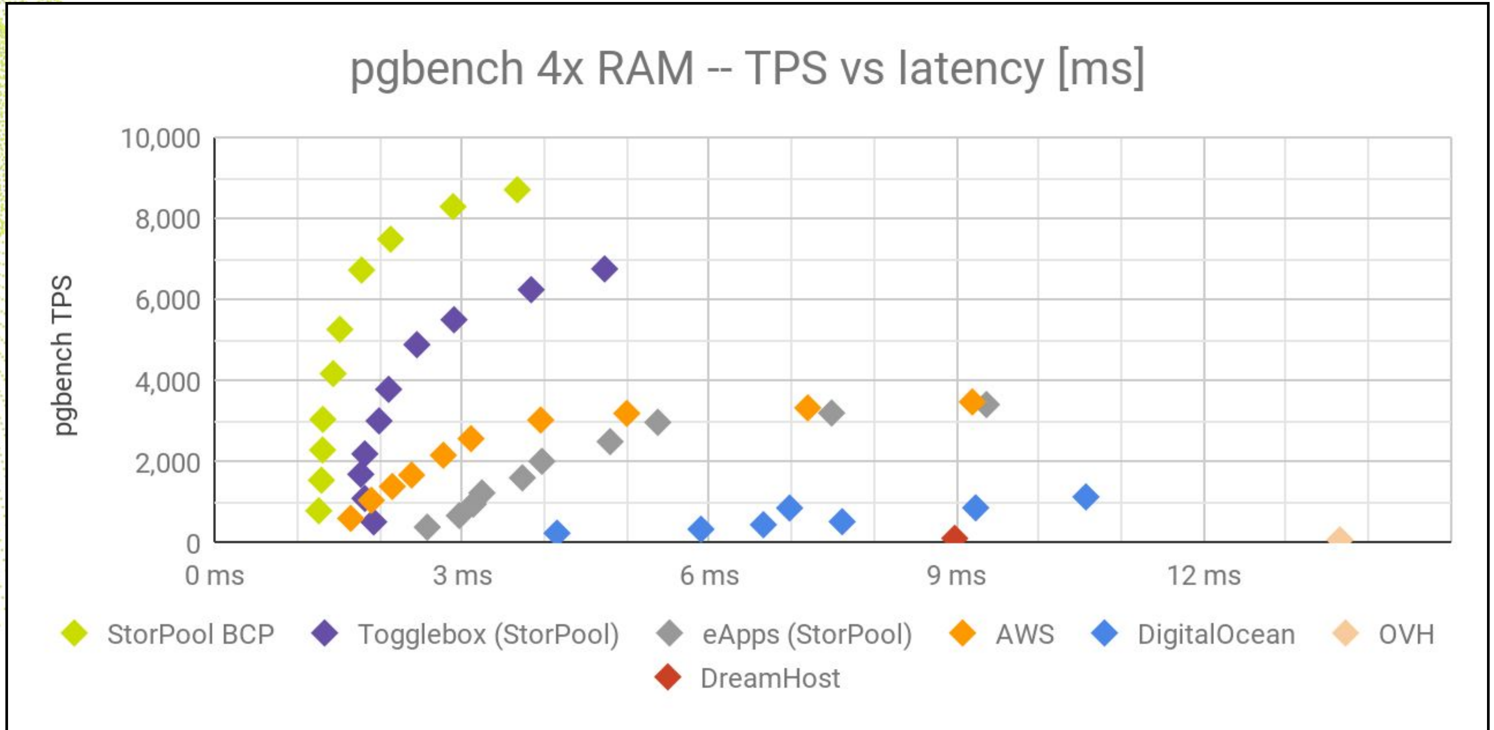
The virtualization & cloud promise:
- near-zero overhead
- vDisks as fast as a local SSD
  are they really?
- efficient consolidation of workloads

StorPool
DISTRIBUTED STORAGE

# Why performance

Fast storage system = more work done per CPU

The virtualization & cloud promise:




StorPool
DISTRIBUTED STORAGE

# Performance with StorPool

>1M IOPS per node
>250k IOPS per core (server)
>500k IOPS per core (client)

StorPool latency overhead on par with latency of NVMe devices.
End-to-end latency approx 2x local NVMe latency.

100k IOPS per NVMe drive with <150µs end-to-end latency
Writes at QD1 ~ 70 µs end-to-end

StorPool
DISTRIBUTED STORAGE

# Why Scale

Public and private cloud
Mobile and Web apps, SaaS
Containers & Microservices
DevOps, Infrastructure as code

What is scale:

- API-driven, integrations
- Scale by adding nodes/drives
- Pooling of capacity & performance
- Metrics collection & Monitoring
- Deployment automation

# Scale with StorPool

Scale-out architecture
>1PB usable All-SSD & Hybrid clusters in production for years
Some customers of StorPool have multiple clusters per location and multiple locations

API control and integrations with Kubernetes, OpenStack, OpenNebula, CloudStack & OnApp

Detailed metrics collection, monitoring.

StorPool DISTRIBUTED STORAGE

# Why Data Management

Assumed that every storage system has it

But many don't

What is data management:

- End-to-end data integrity guarantee
- CoW - Thin provisioning, Snapshots, Clones
- "LUN" per vDisk
- Multi-site with efficient transport of changes
- Fast recovery (changed block tracking)

StorPool
DISTRIBUTED STORAGE

# Data Management with StorPool

End-to-end data integrity protection

4k granularity
- thin provisioning / reclaim
- CoW snapshots, clones
- changed block tracking, incremental recovery and transfer

Multi-site
- connect 2 or more StorPool clusters over public Internet
- send snapshots between clusters for backup and DR
- commonly 100TB backup once per hour

StorPool
DISTRIBUTED STORAGE

Performance goal: vDisks as fast as Local SSD
Scalability goal: Better than Ceph
Data Management goal: Better than ZFS

In one system - StorPool

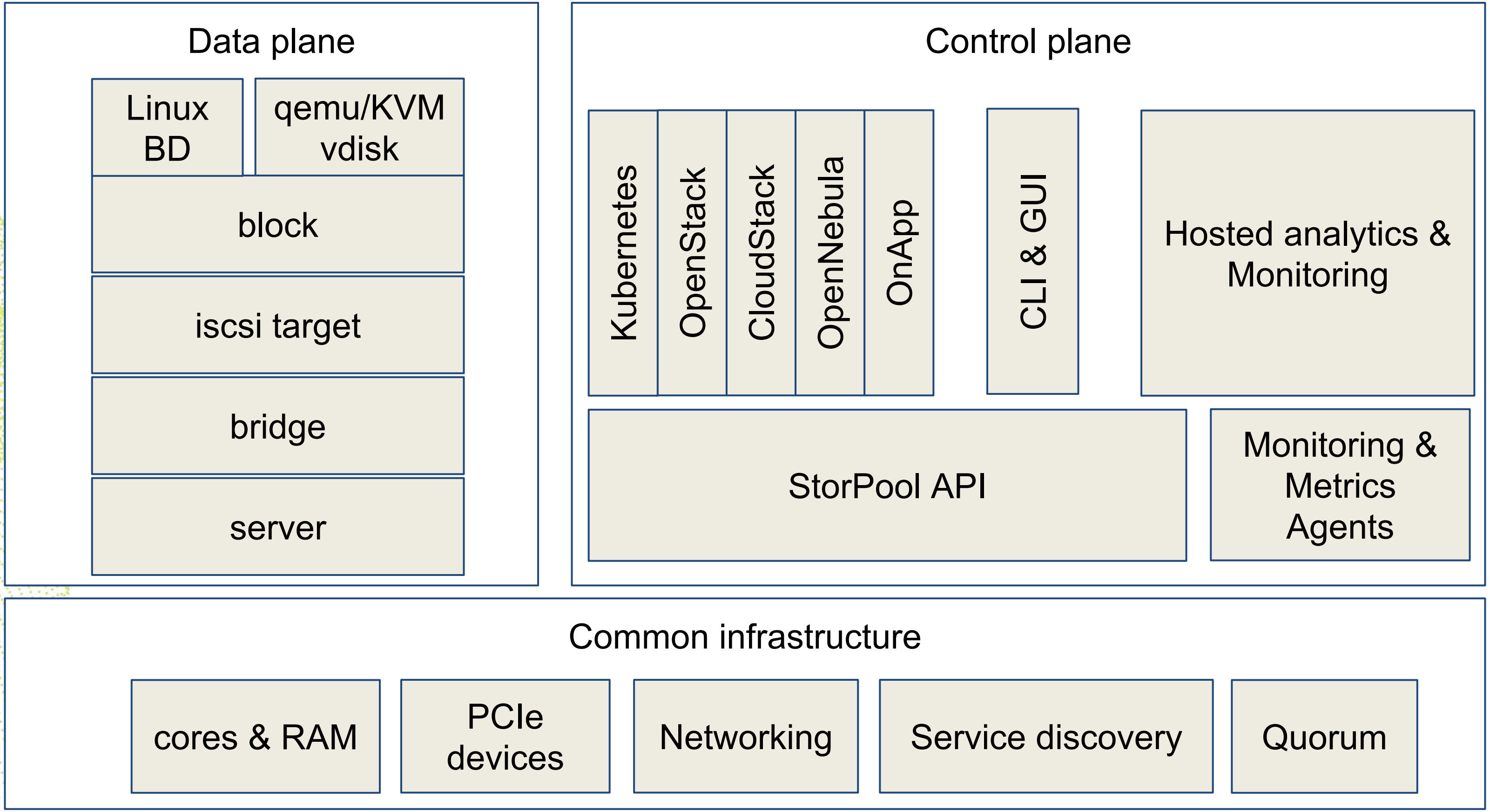
# And as conclusion...

# And as conclusion...

StorPool
DISTRIBUTED STORAGE

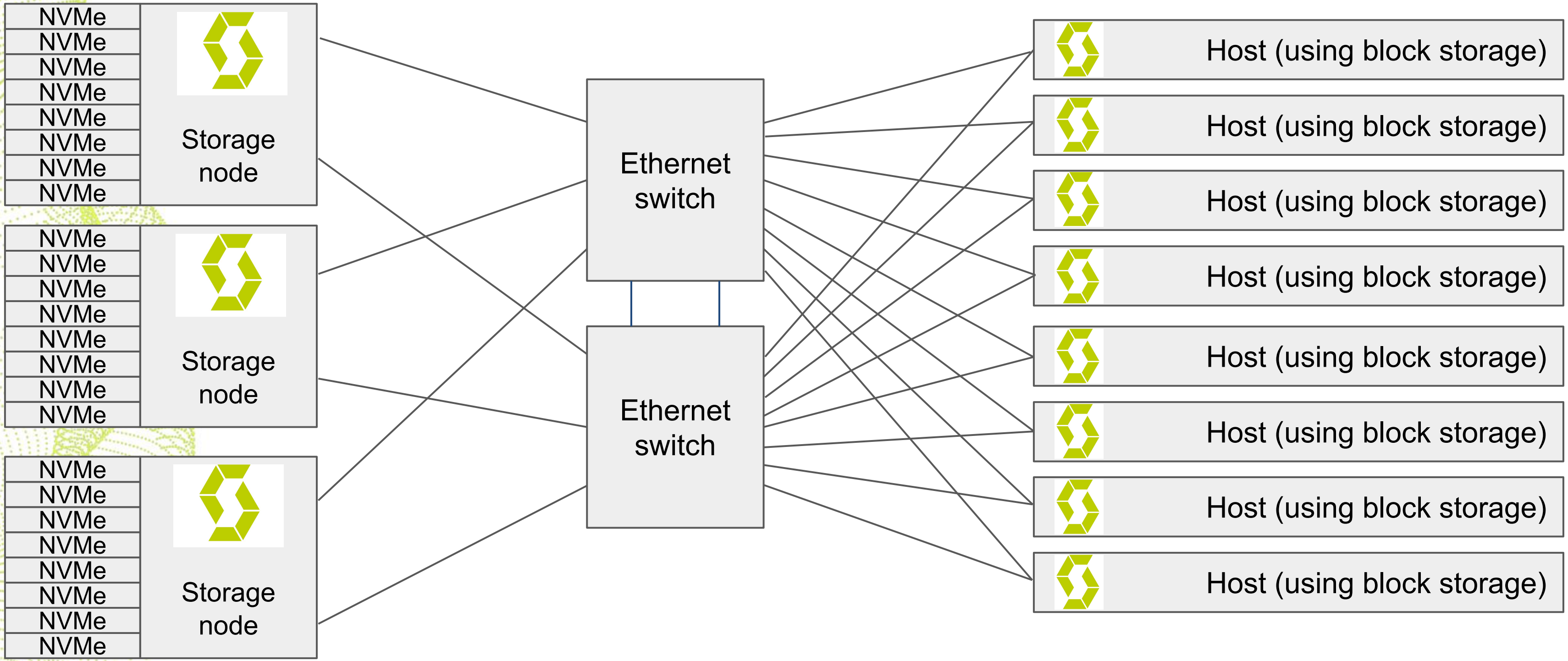
# Architecture and Demo

*Boyan Krosnov, Co-founder and CPO*
*#SFD18, @storpool*




StorPool DISTRIBUTED STORAGE

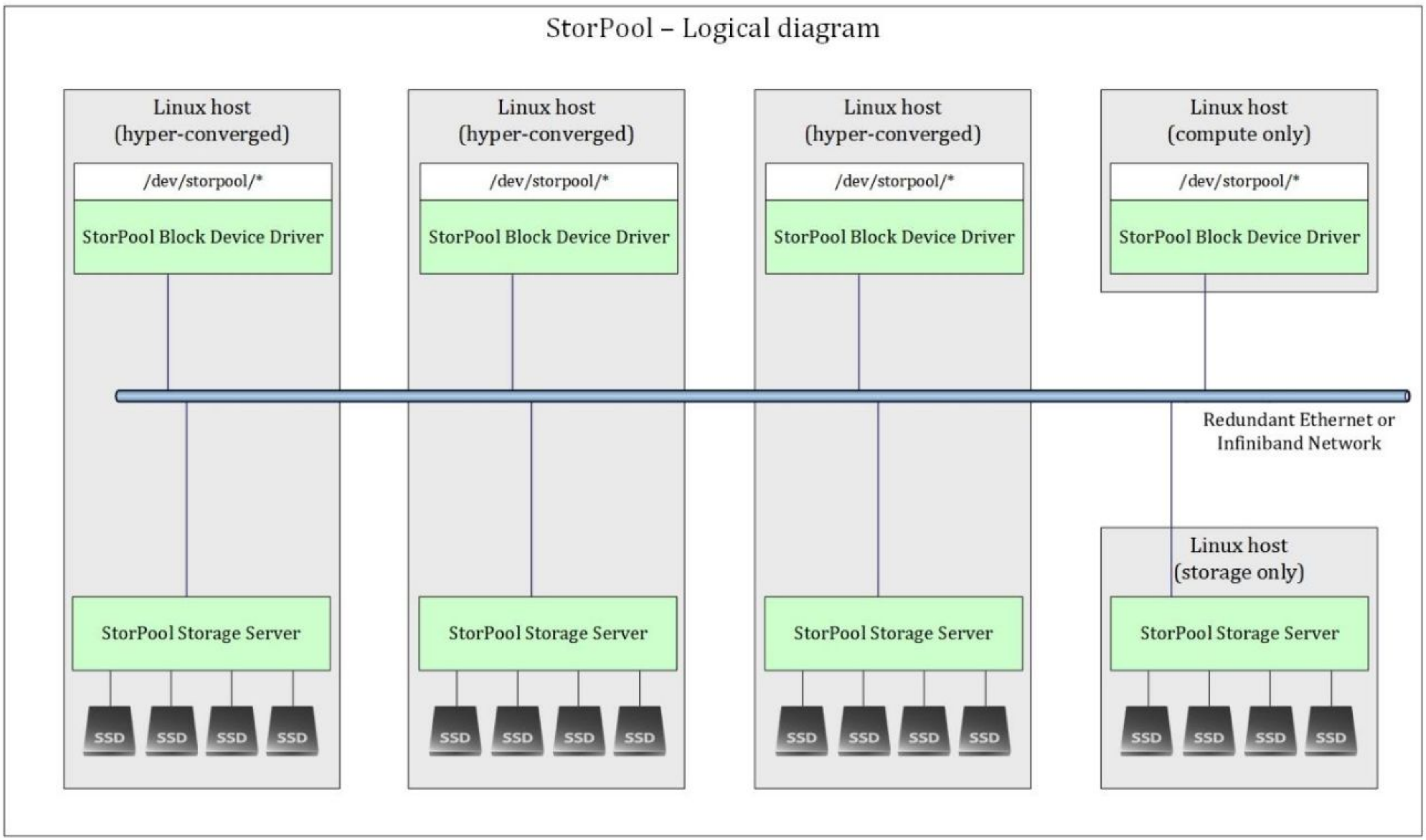
# Data plane

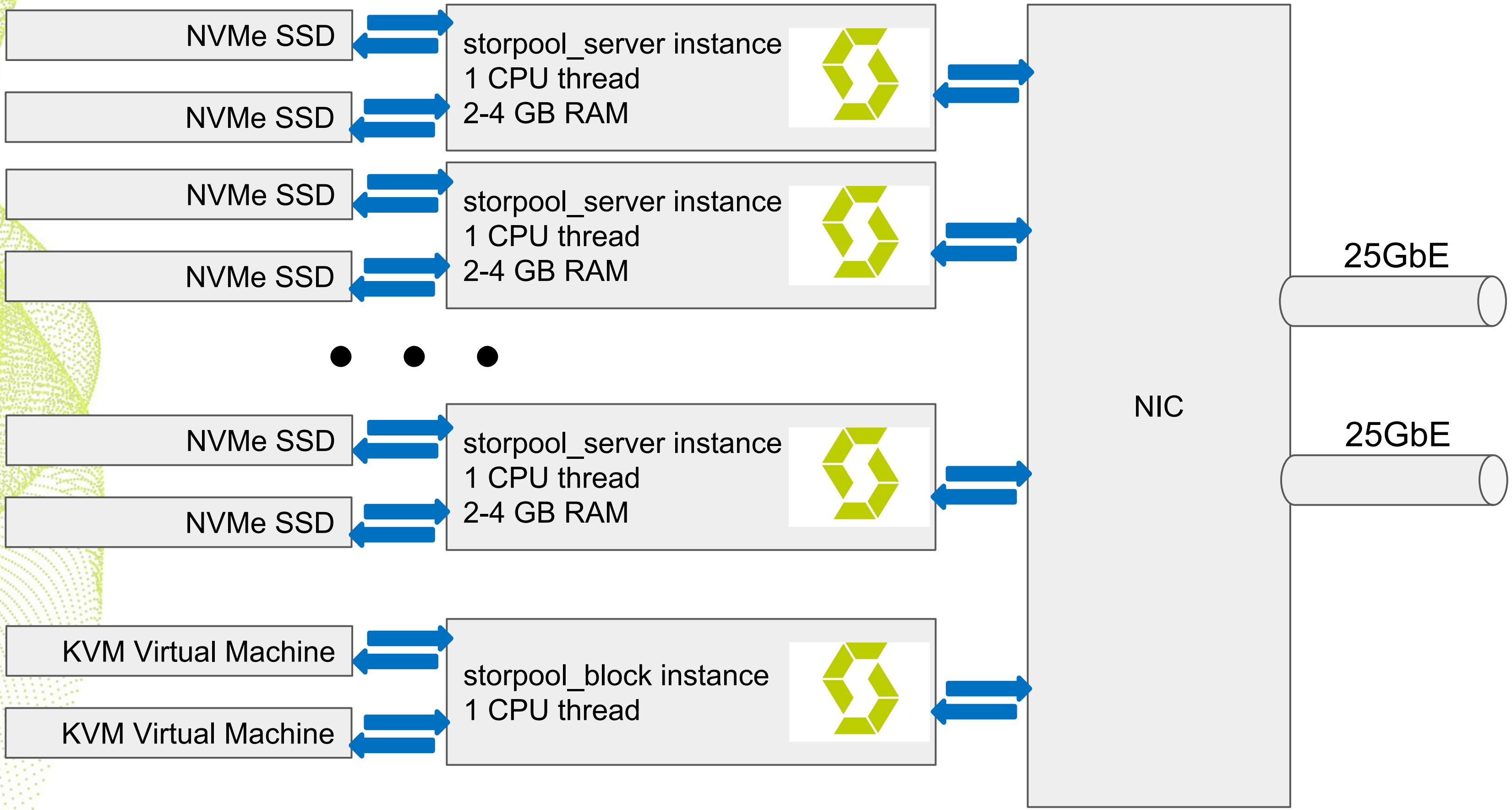



**3+ storage nodes**
**Scale-out ...**

**10s of servers using the storage system**

StorPool DISTRIBUTED STORAGE


StorPool – Logical diagram
Linux host
(hyper-converged)
/dev/storpool/*
StorPool Block Device Driver
StorPool Storage Server
SSD
Linux host
(compute only)
Linux host
(storage only)
Redundant Ethernet or
Infiniband Network




- **Highly scalable and efficient architecture**
- **Scales up in each storage node & out with multiple nodes**

StorPool DISTRIBUTED STORAGE

Protection schemes:

- 3 copies on SATA SSDs or NVMe SSDs
- StorPool Hybrids - 1+2 or 2+1 - lower cost
- 3 copies on HDD
- Erasure coding - soon

Writes are sequentialized and coalesced. Under load 1 write above is less than 1 write below.

Writes may go through "journal" write-back devices before "pool".

Evolution:

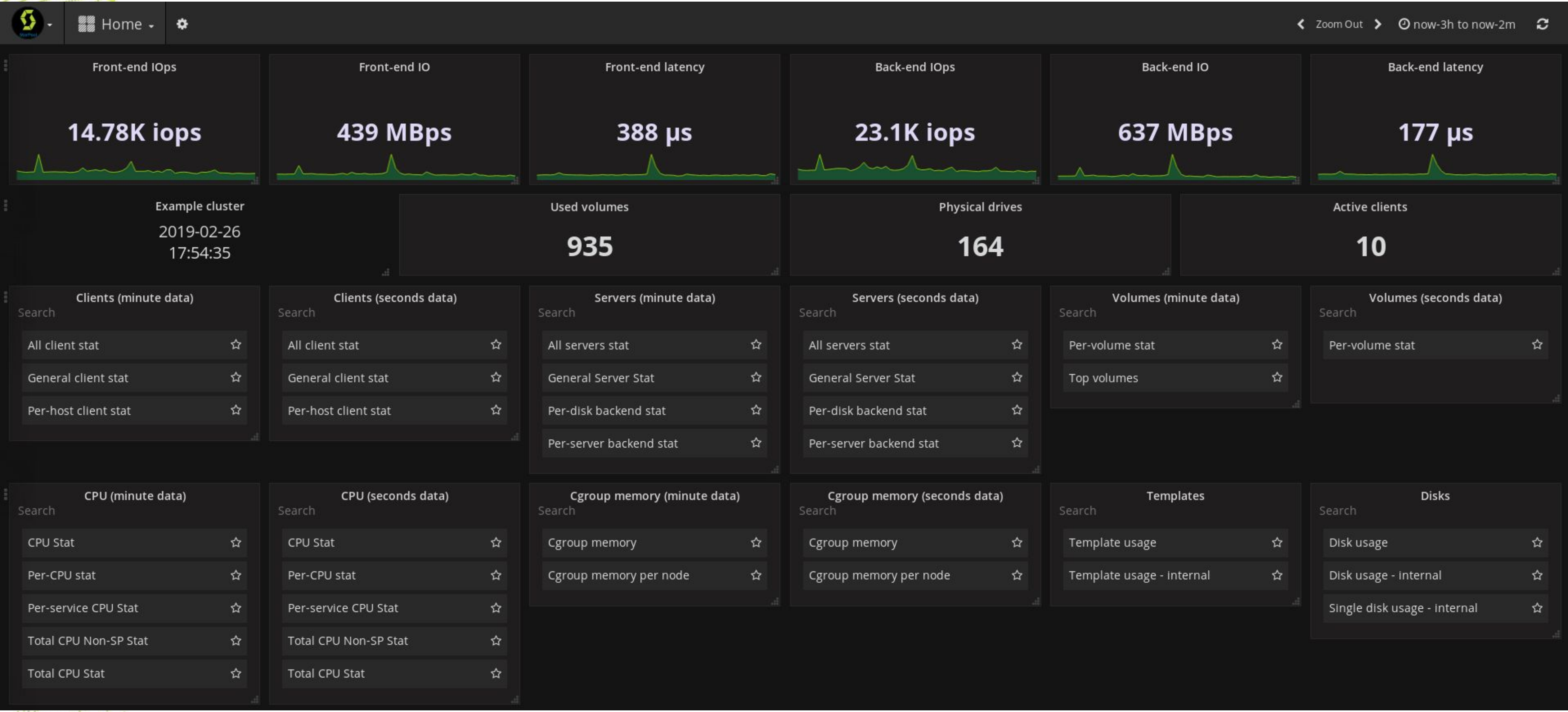
- 3108 RAID controller w/ CacheVault - legacy
- **fast SSD, including "pool" SSDs - current**
- **Intel Optane drive - current**
- NVDIMM / PM - future

StorPool DISTRIBUTED STORAGE

Demo:

- Private cloud use-case
- Basic CLI
- Analytics
- UI dashboard

# Analytics / metrics collection



# Analytics / metrics collection

Top volumes · Zoom Out · Last 6 hours

Number of volumes: 5 · Skip volumes: 5bf50f01f11731.01393943_context

**Top 5 volumes by read bytes**

| total | bps | volume |
| --- | --- | --- |
| 944.61 GiB | 46.96 MBps | 5bf7e61075ed01.01242219_root |
| 437.41 GiB | 21.74 MBps | 5c0533bfdee630.68819684_root |
| 380.50 GiB | 18.92 MBps | 5c177c4c52de63.97691854_root |
| 307.22 GiB | 15.27 MBps | 5c13ded49ccbe1.46303538_root |
| 265.24 GiB | 13.18 MBps | shared_server21_home |

**Top 5 volumes by reads**

| total | rps | volume |
| --- | --- | --- |
| 20.74 Mil | 960.15 iops | 5bf7e61075ed01.01242219_root |
| 10.32 Mil | 477.70 iops | shared_hertz_home |
| 8.56 Mil | 396.57 iops | shared_server28_home |
| 8.53 Mil | 394.84 iops | shared_server25_home |
| 8.37 Mil | 387.51 iops | 5c0533bfdee630.68819684_root |

**Top 5 volumes by written bytes**

| total | bps | volume |
| --- | --- | --- |
| 187.87 GiB | 9.35 MBps | shared_server29_home |
| 81.31 GiB | 4.04 MBps | shared_server24_home |
| 68.48 GiB | 3.40 MBps | shared_server28_localbkp |
| 64.48 GiB | 3.21 MBps | mailvps_home |
| 63.39 GiB | 3.15 MBps | shared_server23_localbkp |

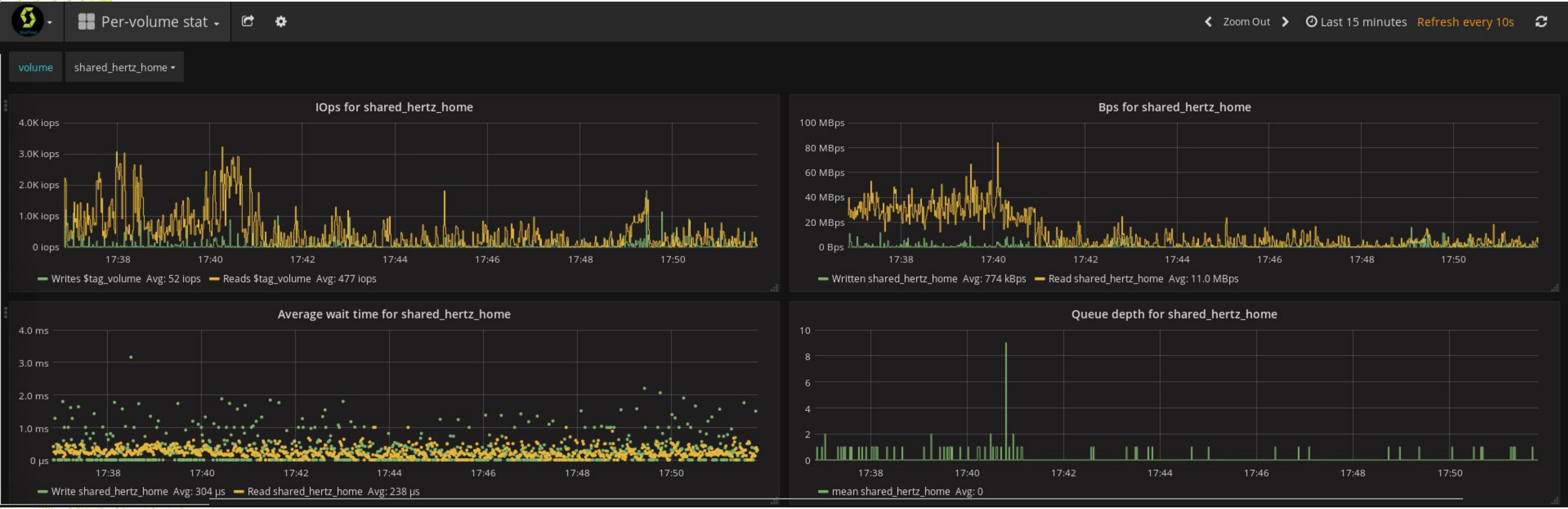
**Top 5 volumes by writes**

| total | wps | volume |
| --- | --- | --- |
| 2.94 Mil | 135.99 iops | 5c46e605e57d55.22476738_ssd |
| 2.38 Mil | 110.12 iops | shared_hertz_var |
| 1.83 Mil | 84.52 iops | shared_server29_home |
| 1.53 Mil | 70.63 iops | shared_hopkins_var |
| 1.29 Mil | 59.76 iops | shared_dio_var |

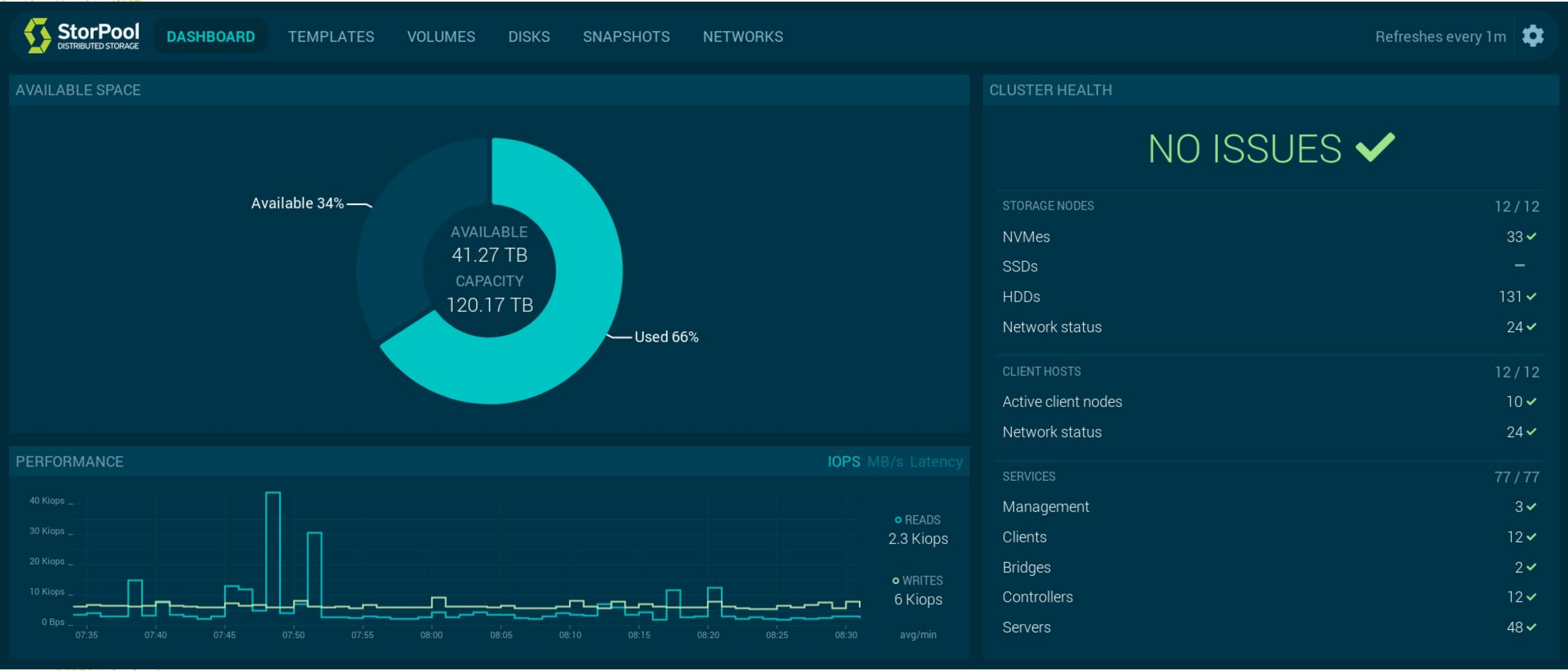
+ ADD ROW

StorPool DISTRIBUTED STORAGE

# Analytics / metrics collection




StorPool
DISTRIBUTED STORAGE




StorPool
DISTRIBUTED STORAGE

StorPool
DISTRIBUTED STORAGE

# Performance demo

*Boyan Krosnov, Co-founder and CPO*
*#SFD18, @storpool*

Intel® Data Center Builders

Intel® Builders Construction Zone

"The new HCI industry record: 13.7 million IOPS with Windows Server 2019 and Intel® Optane™ DC persistent memory"

https://blogs.technet.microsoft.com/filecab/2018/10/30/windows-server-2019-and-intel-optane-dc-persistent-memory/

StorPool
DISTRIBUTED STORAGE

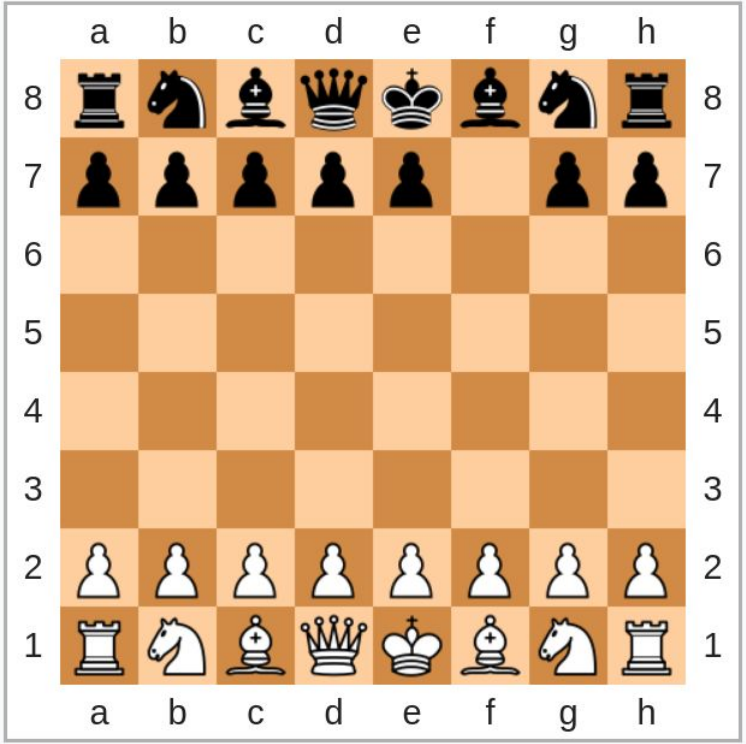
# Microsoft's HCI setup

- 12 nodes, each with:
  - 384 GiB (12 x 32 GiB) DDR4 2666 memory
  - 2 x Intel® Xeon® Scalable processor
  - 1.5 TB Intel® Optane™ DC persistent memory as cache
  - 32 TB NVMe (4 x 8 TB Intel® DC P4510) as capacity
  - 2 x Mellanox ConnectX-4 25 Gbps w/RDMA

- For the best performance, every VM runs on the server node that owns the volume where its VHDX file is stored.
- S2D, Hyper-V, Windows Server 2019

**Handicaps** (or "**odds**") in chess are variant ways to enable a weaker player to have a chance of winning against a stronger one. There are a variety of such handicaps, such as material odds (the stronger player surrenders a certain piece or pieces), extra moves (the weaker player has an agreed number of moves at the beginning of the game), extra time on the chess clock, and special conditions (such as requiring the odds-giver to deliver checkmate with a specified piece or pawn). Various permutations of these, such as "pawn and two moves", are also possible.

Handicaps were quite popular in the 18th and




Initial setup for *pawn and move*: Black starts without the f-pawn; the weaker player (White) moves first

# Microsoft's HCI setup

- 12 nodes, each with:
  - 2 x Intel® Xeon® Scalable processor
  - 384 GiB (12 x 32 GiB) DDR4 2666 memory
  - 1.5 TB Intel® Optane™ DC persistent memory as cache
  - 32 TB NVMe (4 x 8 TB Intel® DC P4510) as capacity
  - 2 x Mellanox ConnectX-4 25 Gbps w/RDMA

- For the best performance, every VM runs on the server node that owns the volume where its VHDX file is stored.
- S2D, Hyper-V, Windows Server 2019

- 12 nodes, each with:
  - 2 x Intel® Xeon® Scalable processor
  - 384 GiB (12 x 32 GiB) DDR4 2666 memory
  - 1.5 TB Intel® Optane™ DC persistent memory as cache
  - 32 TB NVMe (4 x 8 TB Intel® DC P4510) as capacity
  - 2 x Mellanox ConnectX-4 25 Gbps w/RDMA

- For the best performance, every VM runs on the server node that owns the volume where its VHDX file is stored.
- StorPool, KVM, CentOS 7

StorPool DISTRIBUTED STORAGE

- 12 nodes, each with:
  - 2 x Intel® Xeon® Scalable processor
  - 384 GiB (12 x 32 GiB) DDR4 2666 memory
  - 1.5 TB Intel® Optane™ DC persistent memory as cache - REMOVE
  - 32 TB NVMe (4 x 8 TB Intel® DC P4510) as capacity
  - 2 x Mellanox ConnectX-4 25 Gbps w/RDMA

- For the best performance, every VM runs on the server node that owns the volume where its VHDX file is stored.
- StorPool, KVM, CentOS 7

StorPool
DISTRIBUTED STORAGE

- 12 nodes, each with:
  - 2 x Intel® Xeon® Scalable processor
  - 384 GiB (12 x 32 GiB) DDR4 2666 memory
  - 1.5 TB Intel® Optane™ DC persistent memory as cache - REMOVE
  - 32 TB NVMe (4 x 8 TB Intel® DC P4510) as capacity
  - 2 x Mellanox ConnectX-4 25 Gbps w/RDMA - REMOVE
  - ADD: Intel XXV710-DA2 dual-port 25 Gbps
- For the best performance, every VM runs on the server node that owns the volume where its VHDX file is stored.
- StorPool, KVM, CentOS 7

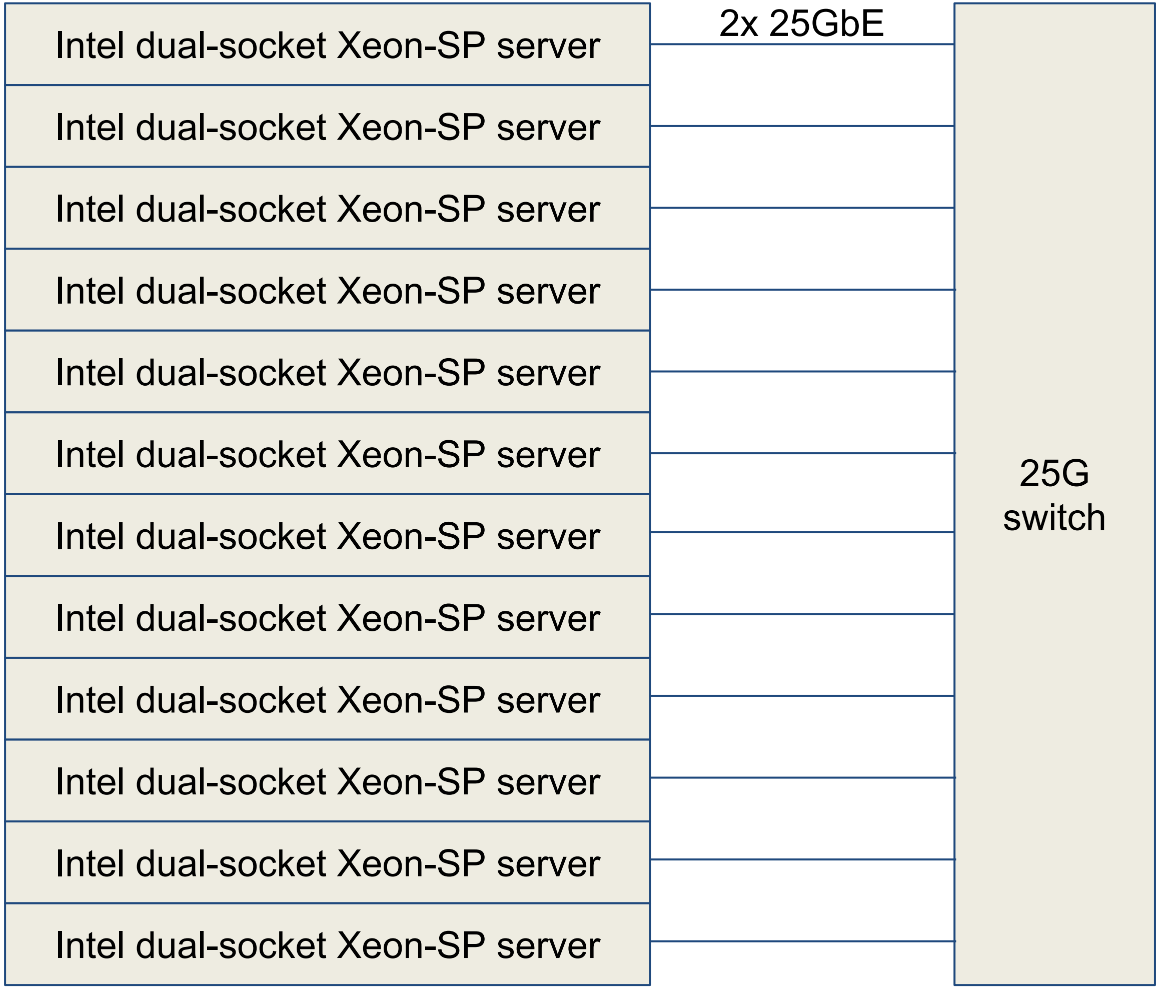
StorPool
DISTRIBUTED STORAGE

# StorPool's HCI setup

- 12 nodes, each with:
  - 2 x Intel® Xeon® Scalable processor
  - 384 GiB (12 x 32 GiB) DDR4 2666 memory
  - 1.5 TB Intel® Optane™ DC persistent memory as cache - REMOVE
  - 32 TB NVMe (4 x 8 TB Intel® DC P4510) as capacity
  - 2 x Mellanox ConnectX-4 25 Gbps w/RDMA - REMOVE
  - ADD: Intel XXV710-DA2 dual-port 25 Gbps

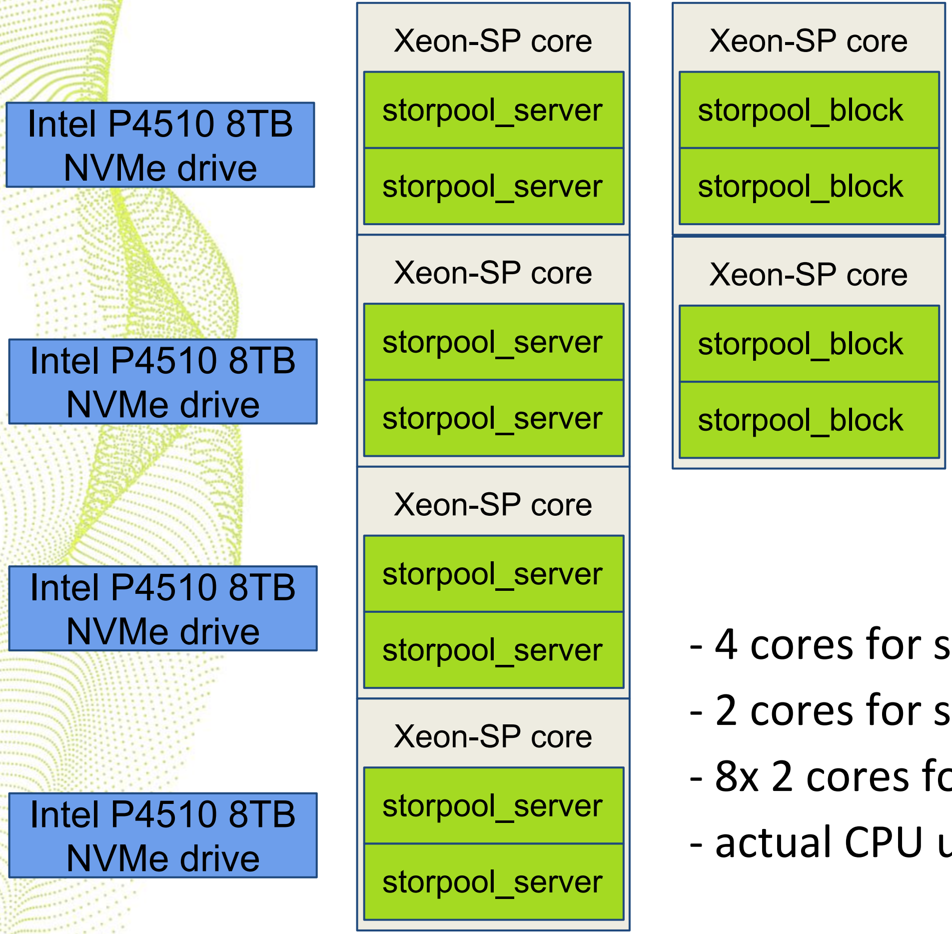
- For the best performance, every VM runs on the server node that owns the volume where its VHDX file is stored. No! - 100% remote.
- StorPool, KVM, CentOS 7

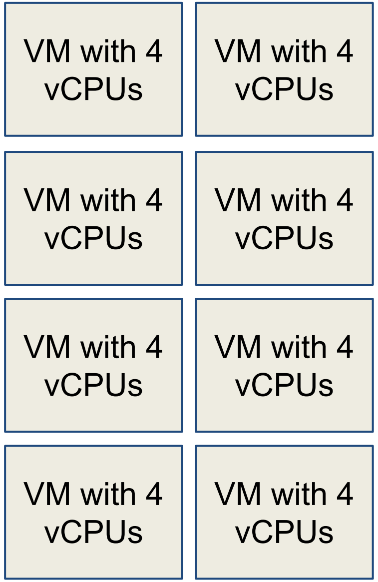
StorPool
DISTRIBUTED STORAGE




StorPool DISTRIBUTED STORAGE

# Resource usage in each node




- 4 cores for storpool_server, 25 GB RAM
- 2 cores for storpool_block
- 8x 2 cores for load generator VMs
- actual CPU usage at full load = ~14 cores

# Summary of results

| result | | parameters | comment |
|---|---|---|---|
| **13.8 M IOPS** | **Random read** | **4k qd 96x64** | **1.15M IOPS per node** |
| 5.5 M IOPS | Random R/W | 70/30 4k qd 96x40 | 183k read/writes /s per drive |
| 2.5 M IOPS | Random write | 4k qd 96x40 | 156k writes /s per drive |
| | | | |
| | | | |
| 64.6 GB/s | Sequential read | bs 128k qd 96x16 | |
| 20.8 GB/s | Sequential write | bs 128k qd 96x16 | |
| | | | |
| 70 µs | Write latency | bs 4k qd 1 | |

StorPool
DISTRIBUTED STORAGE

# Active set discussion

Active set in Microsoft / Hyper-V / S2D
312 VMs * 10 GiB each = 2.9 TB
(15% of 19.8 TB Optane DC persistent memory)

Active set in StorPool / KVM
96 VMs * 500 GiB each = 44.7 TB
(38% of 116 TB system capacity on P4510 NVMes)

storpool_server memory
12 servers * 8 instances * 3.1 GiB = 277 GB

StorPool
DISTRIBUTED STORAGE

# Case studies

*Boyan Ivanov, co-founder & CEO*
*#SFD18, @storpool*

# Case study 1: NVMe-powered VDI cloud - requirement

The need:

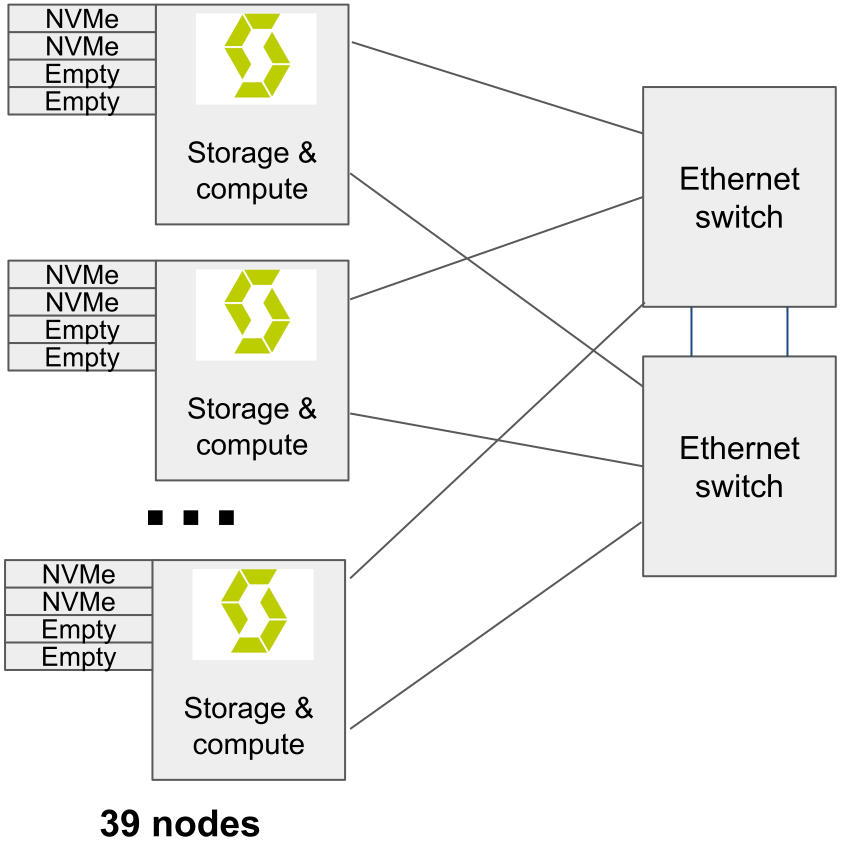
- Fast & cost efficient VDI as a service
- Minimal CPU footprint of software layers
- Latency: as low as possible

The solution:

- First stage: 39 servers, running Hyper-converged (Compute+storage)
- KVM + StorPool + CloudStack
- 2 CPU cores for StorPool (server & client)
- Just 2 NVMe per server, ~90 TB usable
- 2 x 25 GbE Ethernet

StorPool
DISTRIBUTED STORAGE

# Case study 1: NVMe-powered VDI cloud - diagram



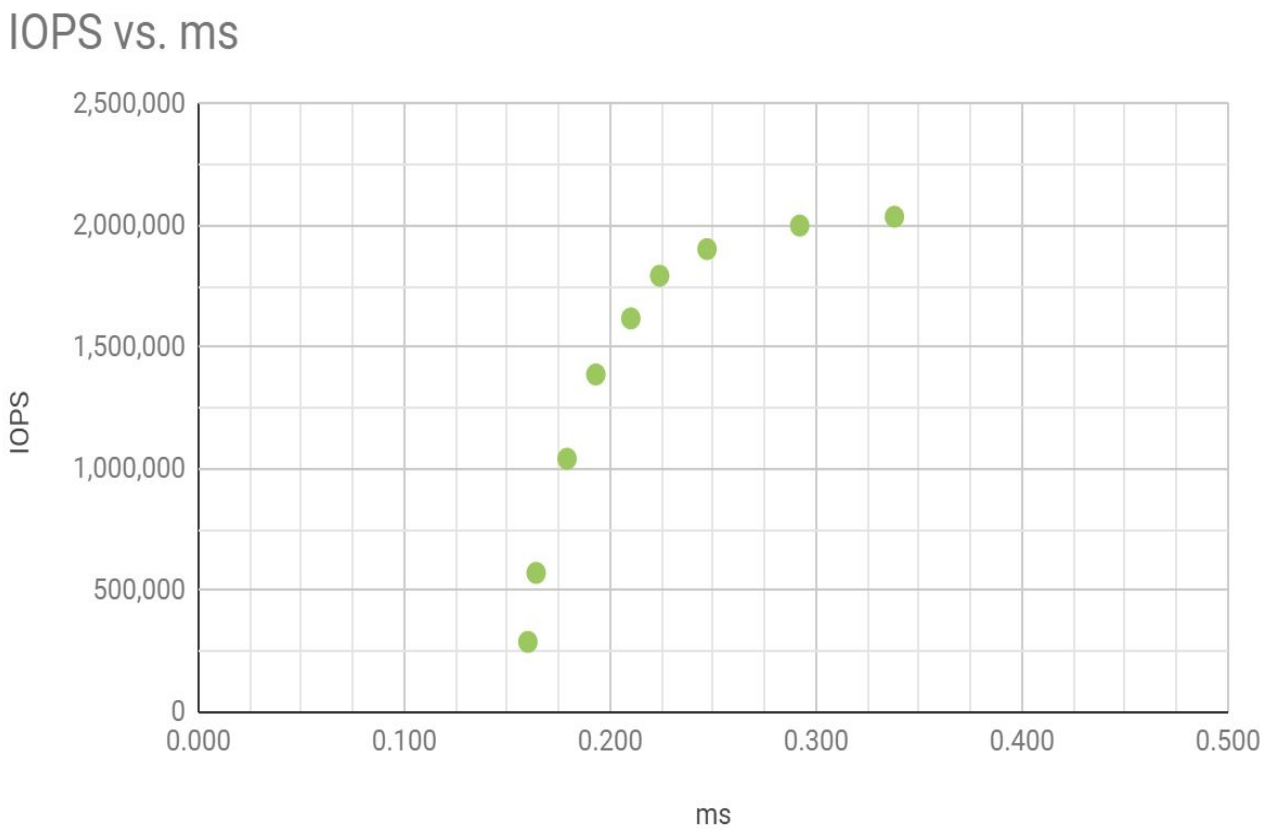

StorPool DISTRIBUTED STORAGE

# Case study 1: NVMe-powered VDI cloud - metrics

| Test | Block size | Queue Depth | Result | Unit |
|---|---|---|---|---|
| Latency read | 4k | 1 | 0.147 | ms |
| Latency write | 4k | 1 | 0.111 | ms |
| | | | | |
| Random read | 4k | 64 | 6,829,218 | IOPS |
| Random read/write | 4k | 64 | 1,904,408 | IOPS |
| Random write | 4k | 64 | 980,619 | IOPS |
| | | | | |
| Sequential read | 1M | 64 | 72,072 | MB/s |
| Sequential write | 1M | 64 | 20,716 | MB/s |

StorPool DISTRIBUTED STORAGE

# Case study 1: NVMe-powered VDI cloud - IOPS vs. latency




StorPool
DISTRIBUTED STORAGE

# Case study 2: Imperia Online - requirement

The need:

- Massively multiplayer online real-time strategy game (MMORTS)
- 40 million users
- Game responsiveness & latency is most important
- Zero downtime target
-

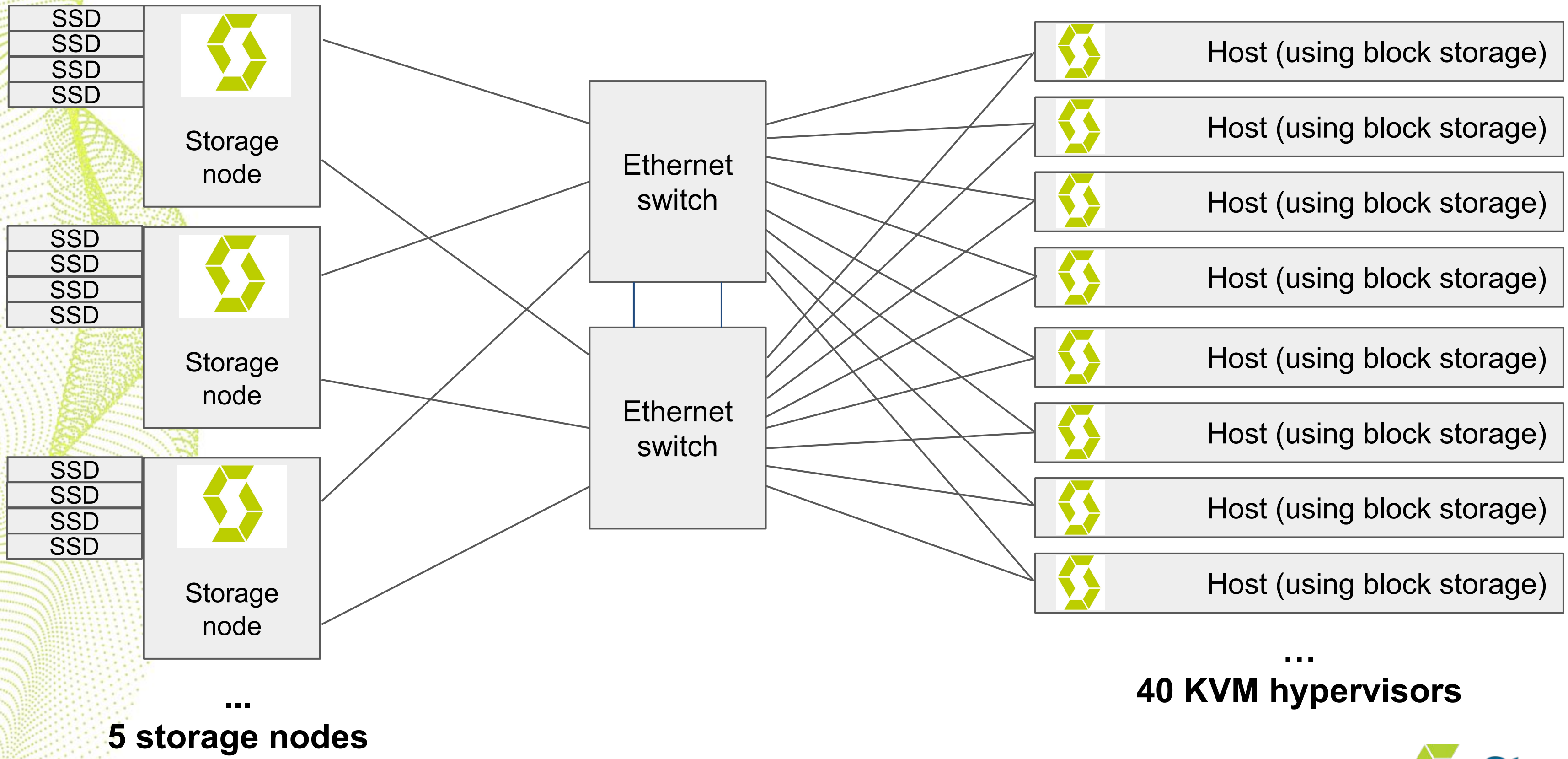
The solution:

- KVM + StorPool + OpenNebula
- Multiple StorPool clusters
  - Each cluster: 5 storage nodes & 40 hypervisors
- Each storage node: 4x SATA SSDs
- 2 x 10 GbE Ethernet

StorPool DISTRIBUTED STORAGE

# Case study 2: Imperia Online

# Case study 2: Imperia Online - diagram




StorPool
DISTRIBUTED STORAGE

# Case study 2: Imperia Online - Metrics

- 100% uptime
- Constant < 1ms real life latency
- Page loading time was reduced from 200-300 milliseconds per page to 75-100 milliseconds.
  - This made Imperia Online 4 times faster than its biggest competitor.
- Financial Times study on page load time (28 days measurement):
  - 1 second slower: -4.57% in USD revenue
  - 3 seconds slower: -7.89% in USD revenue

StorPool
DISTRIBUTED STORAGE

StorPool
DISTRIBUTED STORAGE

# Thank you!

Boyan I. / Boyan K.

StorPool Storage
www.storpool.com
info@storpool.com
@storpool