# high performance software defined storage

*in 15 minutes or less*

Sam McLeod | smcleod.net | @s_mcleod

# Stop talking about $/GB

It's only relevant in *some* situations

# Start talking about $/IOPs

Unless you *really* don't need performance

# Why IOPs Matter

## Performance

Databases rely almost solely on small random reads and writes

## Data Safety

In flight data is a risk

The time it takes your storage to recover with RAID rebuild, (re)provisioning - is the time that your data or performance is at risk.

**Amazon found every 100ms of latency cost them 1% in sales**

# Existing Solution

## HP P4530 Lefthand SAN Units

$20,000/node

- 12x 15K 600GB Disks in RAID6
- 2-4K IOP/s
- 250-280W Idle
- Java (swing) app on Windows
- ILO requires IE6/8
- No access to underlying OS or Tools
- No additional per-node capacity

**$50,000 / 10K IOP/s**

Summary | ⚠ Upgrades

Click here to stay at current software version...

⚠ Upgrades Available

Centralized Management Console

Available Systems

s1-san-mg

HP StoreVirtual VSS Provider

HP StoreVirtual DSM for MPIO

HP StoreVirtual Command-Line Interface (CLI)

**Shell script to update your SAN?**

**Also, WTF**

Getting the hostname from 's1-san4 (10.51.40.41)'

**Upgrade Progress**

Installation process started on: 28/03/15 5:40 PM

Percent Complete:

66%

Hide Details

Upgrade Details

.51.40.31:11120: Received an UpgradeScriptStatus message from the install server running on the node it wa
s : C king storage node ...

/10.51.40.31:1 Received an UpgradeScriptStatus message from the install server running on the node it wa
s : begin MAIN install_saniq.sh: Preparing for upgrade ...

/10.51.40.31:11120: Received an UpgradeScriptStatus message from the install server running on the node it wa
s : Installing new software ...

Export Details...

Abort Unfinished Installations

Questions in SAM (1776325)

Please enter questions b

| ID | Me Too | Submitted By | Status |
|---|---|---|---|
| 1 | 1 | Sam McLeod | New |
| 2 | 1 | Sam McLeod | New |
| 3 | 1 | Sam McLeod | New |

ID :

Response 0 of 0

Add Response

People: 2

Presenters: 2

Sam McLeod

Participants: 0

[redacted] - Chat Room

Information

[redacted]

Sam McLeod

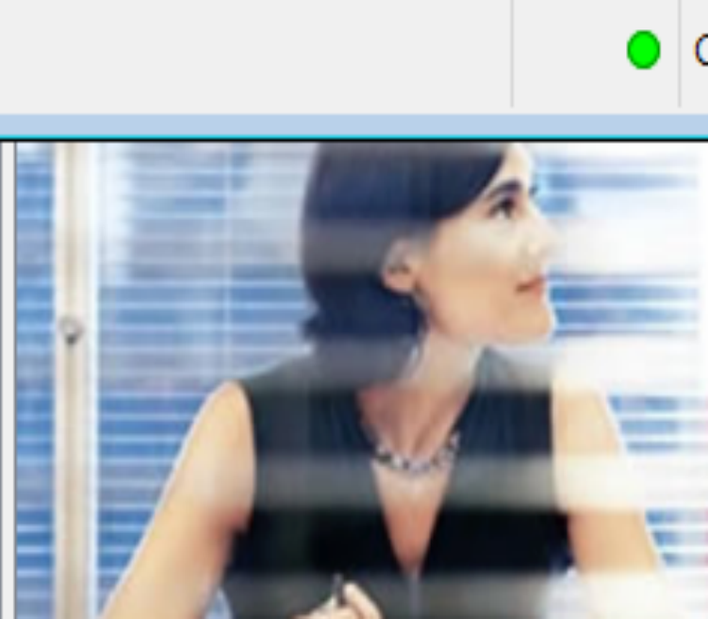
2

Connected

Sam McLeod: [redacted]are you there
Sam McLeod: you and someone else were just looking at our SAN storage
Sam McLeod: and rebooted one of our arrays
Sam McLeod: a lot of our production servers lost their storage and froze
[redacted]: oh no

Submit

Connection: Good



Welcome to
HP Virtual Room

Java Setup

Java™ ORACLE

## We recommend installing the FREE Browser Add-on from Ask


### Get the best of the Web delivered to you!

Receive Facebook status updates directly in your browser, listen to thousands of top radio stations, and get easy access to search, YouTube videos, local weather, and news.

- [x] **Install the Ask Toolbar in Google Chrome**
- [x] **Set and keep Ask as my default search provider in Google Chrome**

By installing this application you agree to the End User License Agreement and Privacy Policy.
The Ask Toolbar is a product of APN, LLC. You can remove this application easily at any time.

Cancel Next >

We don't work for the tools*

So make the tools work for us

**Pun slightly intended*

Standard tools with modern technology

Open source the design

Break the rules

But not at the expense of safety

# Modernise Landscape

Automated provisioning

Vendor neutral

Lower total cost of ownership

Standard components & software

Modular scalability

# Break the $100/10K IOP/s Barrier

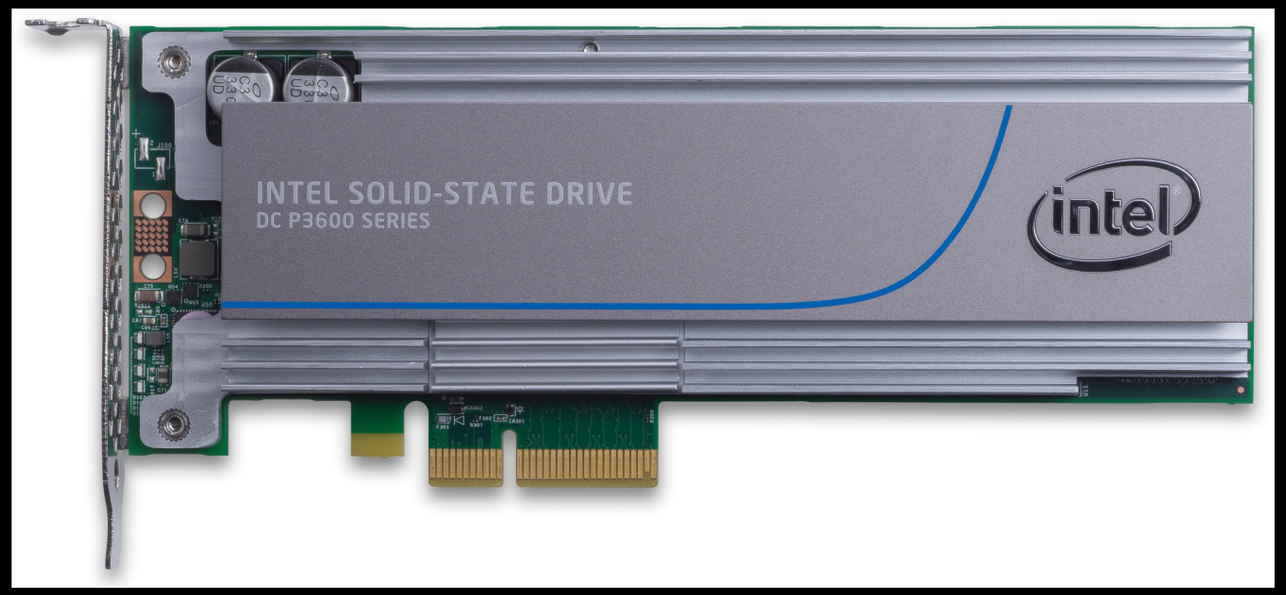
Must be resilient, reliable

Aim for 1M IOP/s per node

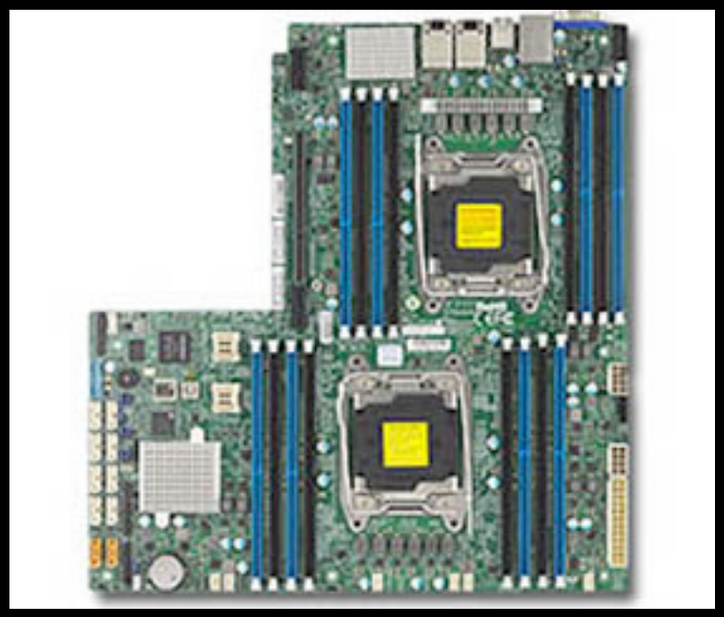
Reduce data latency and 'time in flight'

Shake up perception of 'enterprise' storage

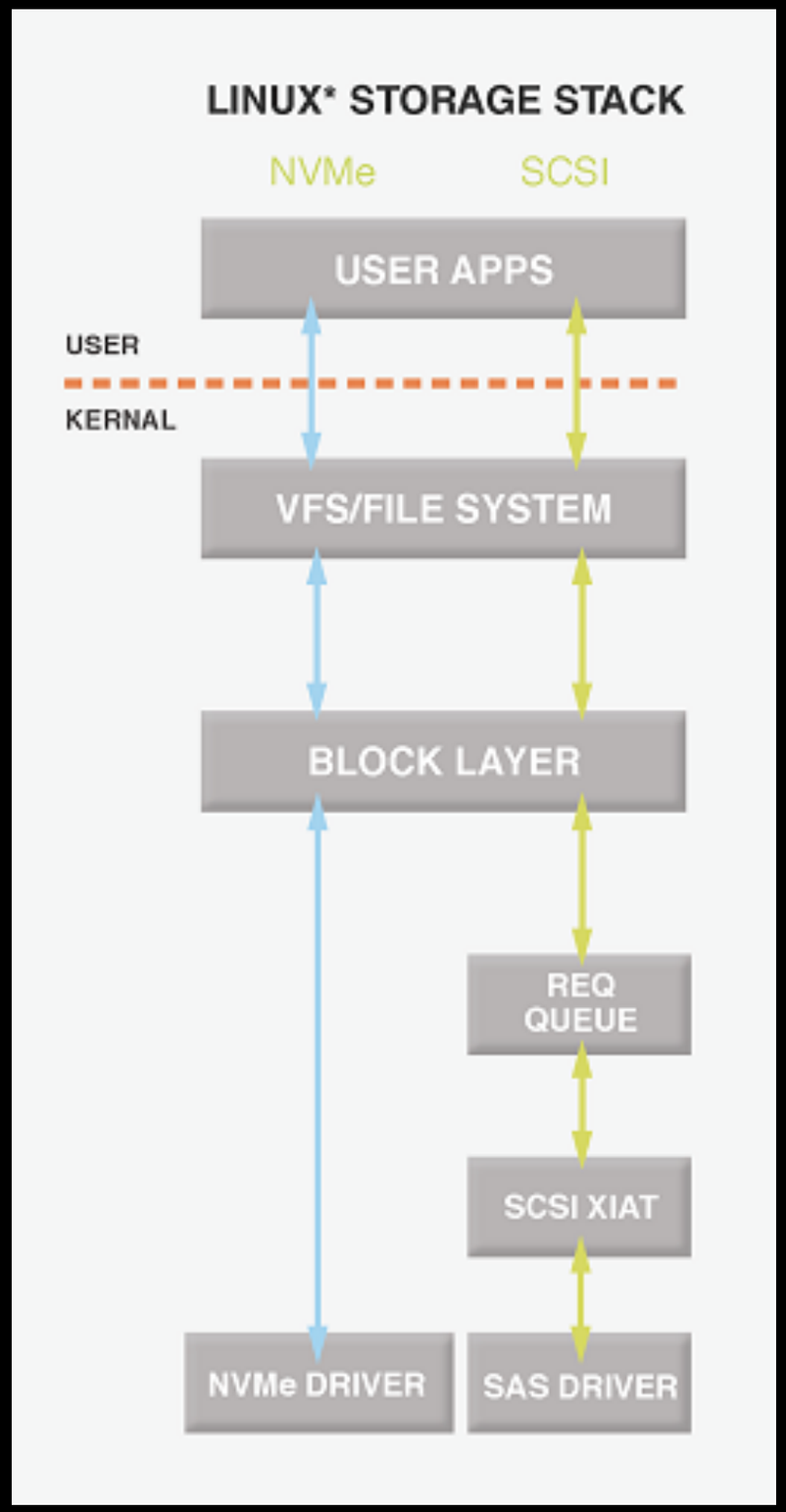
# Hardware

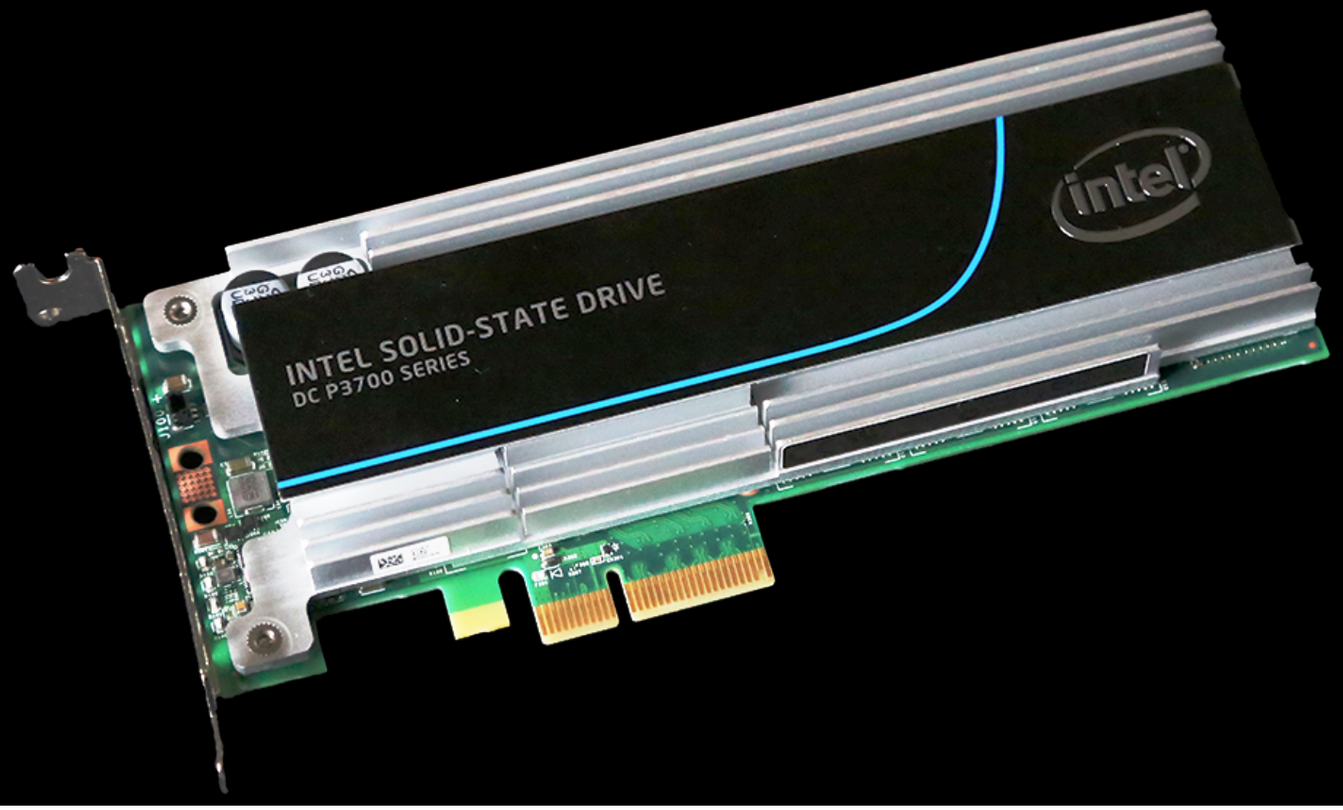

INTEL SOLID-STATE DRIVE
DC P3600 SERIES
intel


SanDisk
Extreme PRO



# NVMe




**High-level comparison of AHCI and NVMe**[3]

| | AHCI | NVMe |
|---|---|---|
| **Maximum queue depth** | 1 command queue;<br>32 commands per queue | 65536 queues;<br>65536 commands per queue |
| **Uncacheable register accesses (2000 cycles each)** | 6 per non-queued command;<br>9 per queued command | 2 per command |
| **MSI-X and interrupt steering** | single interrupt;<br>no steering | 2048 MSI-X interrupts |
| **Parallelism and multiple threads** | requires synchronization lock<br>to issue a command | no locking |
| **Efficiency for 4 KB commands** | command parameters require<br>two serialized host DRAM fetches | gets command parameters<br>in one 64 Bytes fetch |

*Dave Landsman. "AHCI and NVMe as Interfaces for SATA Express Devices – Overview"*

# Intel DC3600 NVMe PCIe 1.2TB

- Mean Time Between Failures = 2,000,000 Hrs (228.311 Years)
- Endurance = >8.6 PB written
- Uncorrectable Bit Error Rate = less than 1 sector per 1e17 bits read (12.500 petabytes)
- 5 Year Warranty

# SanDisk Extreme Pro SATA 480GB

- Mean Time Between Failures = 2,000,000 Hrs (228.311 Years)
- Endurance = >80 TB written
- Uncorrectable Bit Error Rate = Can't find this anywhere?
- 10 Year Warranty
- It's not Samsung

# Resiliency

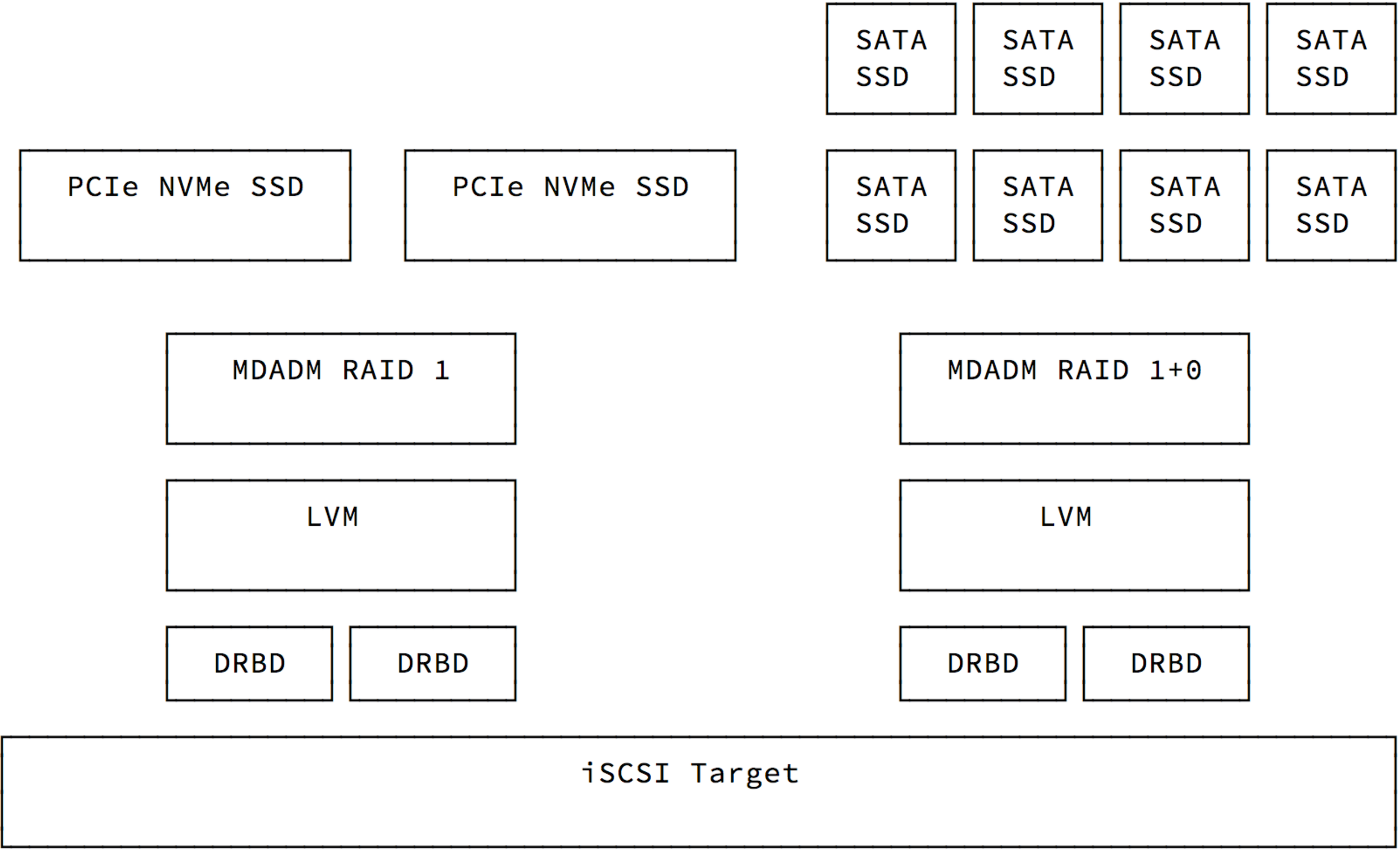
Node-Node replication & failover

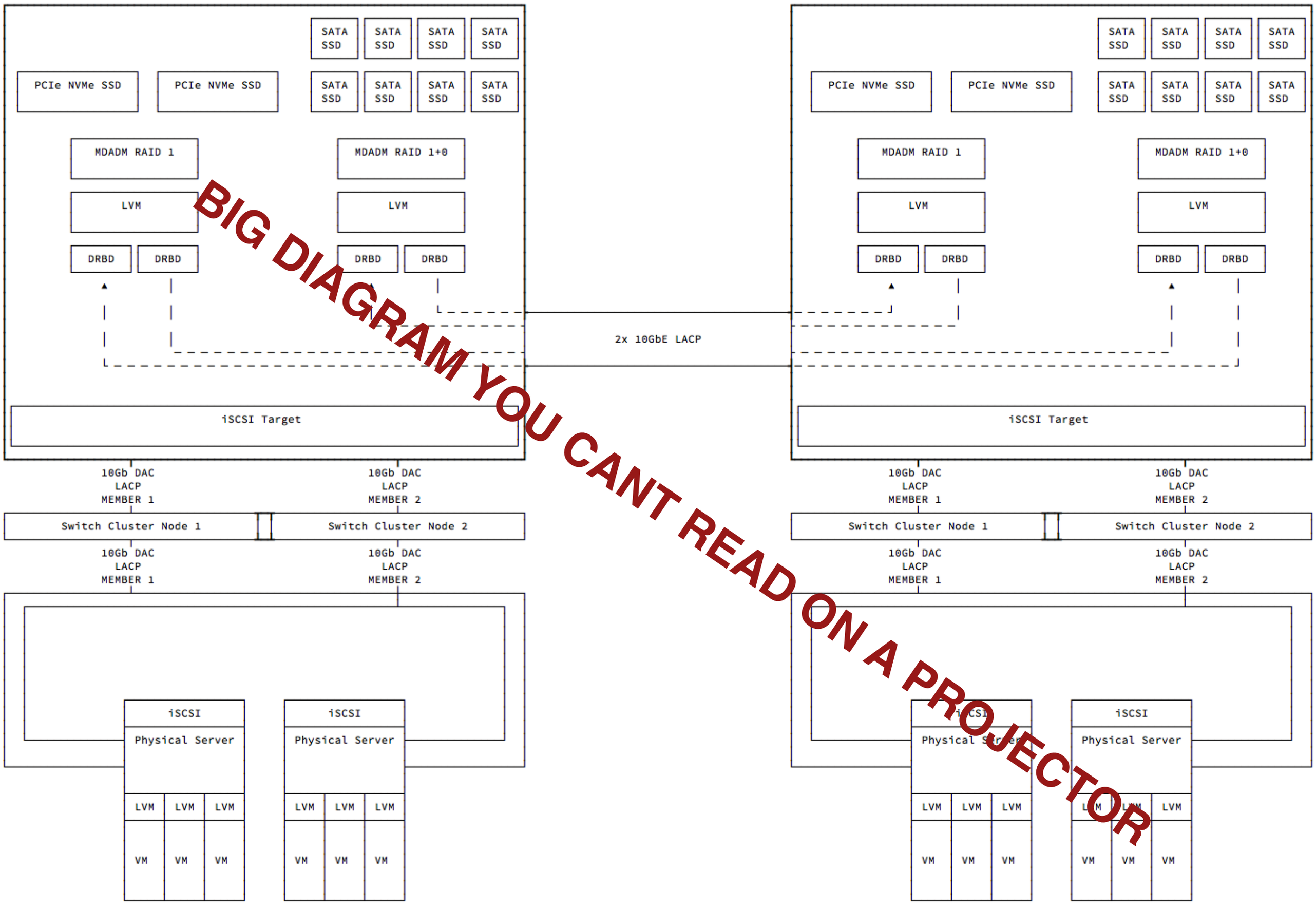
No proprietary RAID controllers or drivers

Low latency means greatly reduced 'in flight' data risk

**Network, Storage, Node and Site Level Redundancy**

# Node Topology




SATA SSD
SATA SSD
SATA SSD
SATA SSD
PCIe NVMe SSD
PCIe NVMe SSD
SATA SSD
SATA SSD
SATA SSD
SATA SSD
MDADM RAID 1
MDADM RAID 1+0
LVM
LVM
DRBD
DRBD
DRBD
DRBD
2x 10GbE LACP
iSCSI Target
10Gb DAC LACP MEMBER 1
10Gb DAC LACP MEMBER 2
Switch Cluster Node 1
Switch Cluster Node 2
10Gb DAC LACP MEMBER 1
10Gb DAC LACP MEMBER 2
iSCSI
Physical Server
LVM
LVM
LVM
VM
VM
VM
BIG DIAGRAM YOU CANT READ ON A PROJECTOR

# Storage as code

```yaml
---
storage_alpha:
  manage_ssd: true
  manage_drbd: true
  manage_iscsi: true
  manage_mdadm: true
  arrays:
    '/dev/md480':
      devices:
        - /dev/sda
        - /dev/sdb
        - /dev/sdc
        - /dev/sdd
        - /dev/sde
        - /dev/sdf
        - /dev/sdg
        - /dev/sdh
      level: 10
      force: false
      options: '--layout=f2 --chunk=512K'
      metadata: '1.2'
    '/dev/md200':
      devices:
        - /dev/nvme0n1
        - /dev/nvme1n1
      level: 1
      options: '--chunk=512K'
      metadata: '1.2'
      force: true
```

Node Group &
Volume Configuration

```yaml
  volume_groups:
    nvme:
      physical_volumes:
        - /dev/md200
      logical_volumes:
        nvmetest:
          size: 100G
    sandisk:
      physical_volumes:
        - /dev/md480
      logical_volumes:
        sdtest:
          size: 100G

iscsi_targets:
  'iqn.2015-01.com.example:storage.backups':
    backing: '/dev/mapper/nvme-nvmetest'
    ipaddress: '10.51.10.145'
    user: 'sam'
    password: 'sam'
```

# Replication & Network Configuration (Defaults)

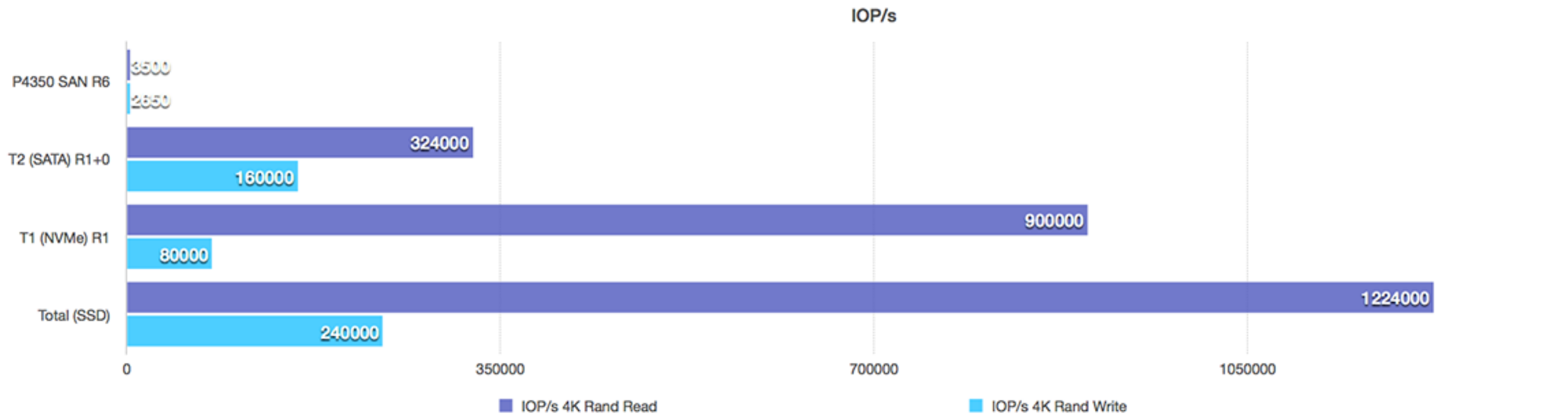
```
52    drbd:
53      r0:
54        host1: 's1-san5'
55        host2: 's1-san6'
56        ip1: '172.30.1.5'
57        ip2: '172.30.1.6'
58        device: '/dev/drbd0'
59        port: '7789'
60        verify_alg: 'sha1'
61        manage: true
62        initial_setup: true
63        rate: '4194304K'
64        logical_volume: '/dev/mapper/nvme-nvmetest'
65      net_parameters:
66        'max-epoch-size':'8000'
67        'no-tcp-cork' :''
68        'unplug-watermark':'16'
69        'sndbuf-size':'0'
70      syncer_parameters:
71        'al-extents' :'3389'
72      disk_parameters:
73        'disk-flushes': 'no'
74        'md-flushes': 'no'
75        'disk-barrier': 'no'
76        'no-disk-flushes': ''
77        'no-md-flushes': ''
78        'no-disk-barrier': ''
79        'c-plan-ahead': '10'
80        'c-max-rate': '1024M'
81        'c-max-rate': '512M'
82        'c-fill-target': '500k'
```

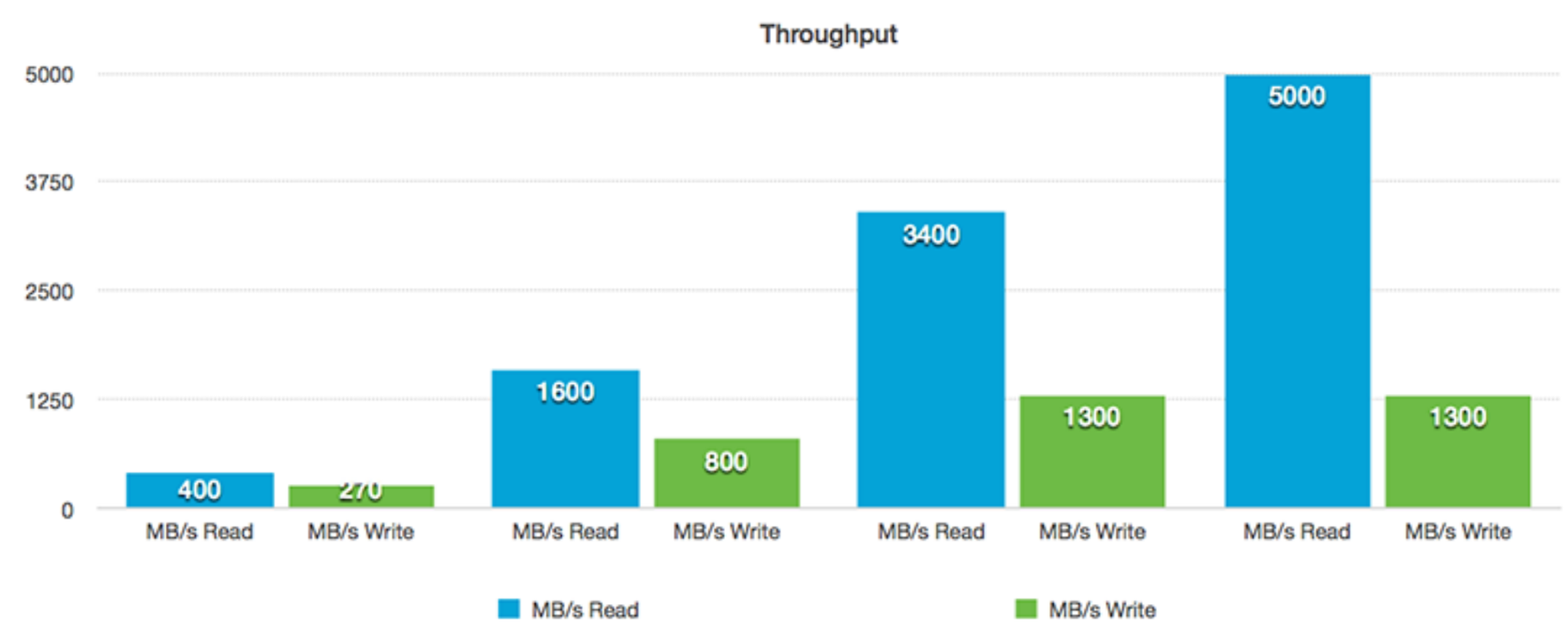
```
84  network_storage:
85    eth0:
86      bond_master: 'drbd'
87      mtu: '9000'
88      method: 'manual'
89    eth1:
90      bond_master: 'drbd'
91      mtu: '9000'
92      method: 'manual'
93    eth2:
94      bond_master: 'storage'
95      mtu: '9000'
96      method: 'manual'
97    eth3:
98      bond_master: 'storage'
99      mtu: '9000'
100     method: 'manual'
101   storage:
102     ipaddress: '%{storage_ip}'
103     netmask: '255.255.255.0'
104     bond_updelay: '200'
105     bond_downdelay: '200'
106     bond_miimon: '100'
107     bond_mode: '802.3ad'
108     mtu: '9000'
109     gateway: '10.51.40.1'
110     dns_search: 'serv.abb.ixa.net.au.'
111     dns_nameservers: '10.51.10.235'
112     manage_order: '20'
113   drbd:
114     ipaddress: '%{drbd_ip}'
```

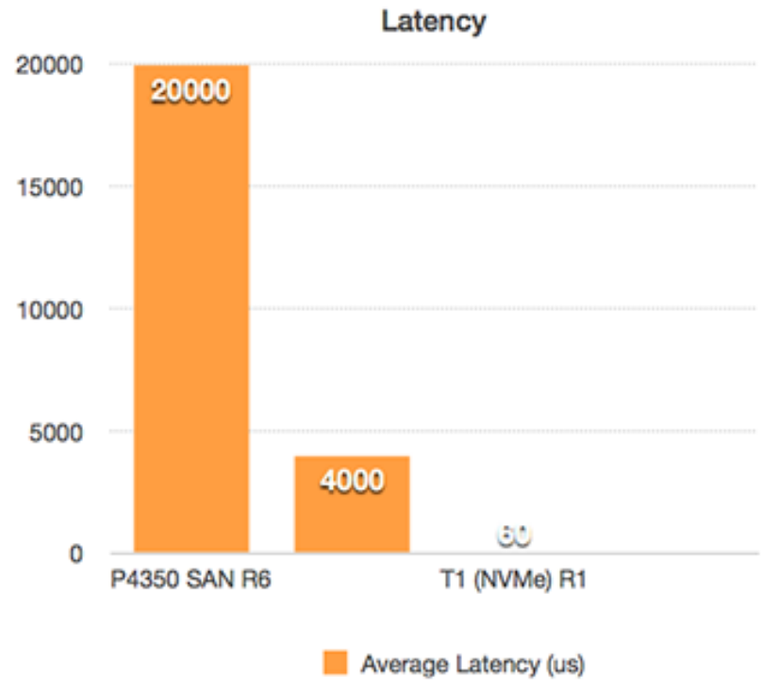
# Performance

You just want to see the numbers right?

*Please keep in mind these are early results and may be subject to change*



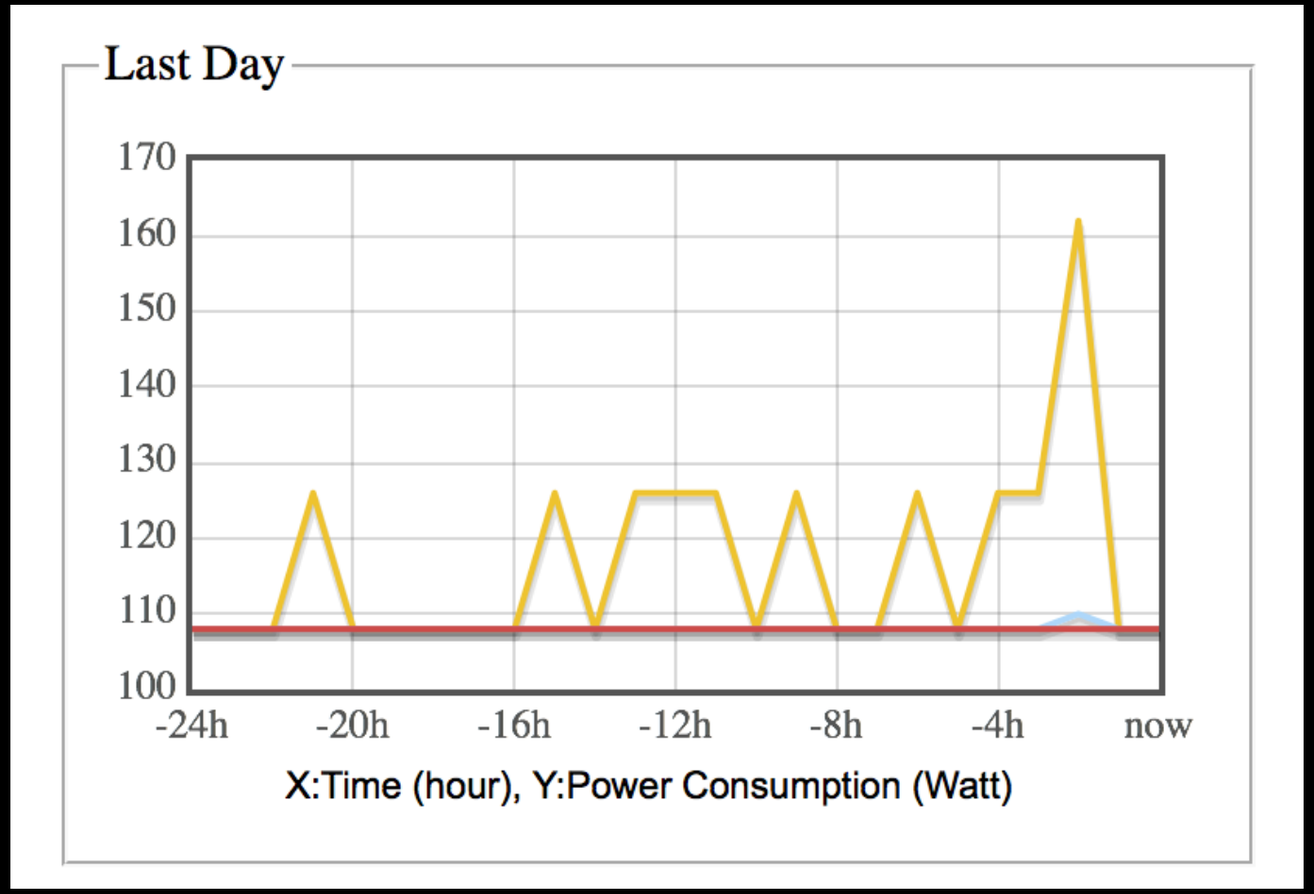




Sustained Performance

| TEST | P4350 SAN R6 | T2 (SATA) R1+0 | T1 (NVME) R1 | TOTAL (SSD) |
|---|---|---|---|---|
| **IOP/s 4K Rand Read** | 3500 | 324,000 | 900,000 | **1,224,000** |
| **IOP/s 4K Rand Write** | 2650 | 160,000 | 80,000 | **240,000** |
| **MB/s Read** | 400 | 1600 | 3,400 | **5,000** |
| **MB/s Write** | 270 | 800 | 1,300 | **1,300** |
| **Average Latency (us)** | 20000 | 4,000 | 60 | |

# Power Consumption

100W idle, 180W maximum



Cost

# $10,000/node

## $100/10K IOP/s

1M+ IOP/s 4K rand reads

5GB/s throughput

Highly redundant

Scalable design

Managed in code

1RU form factor

https://github.com/sammcj/storage

Sam McLeod | smcleod.net | @s_mcleod