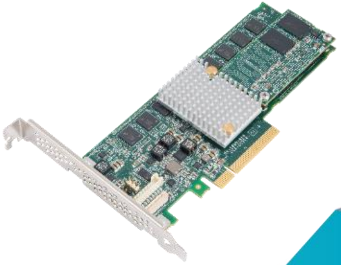
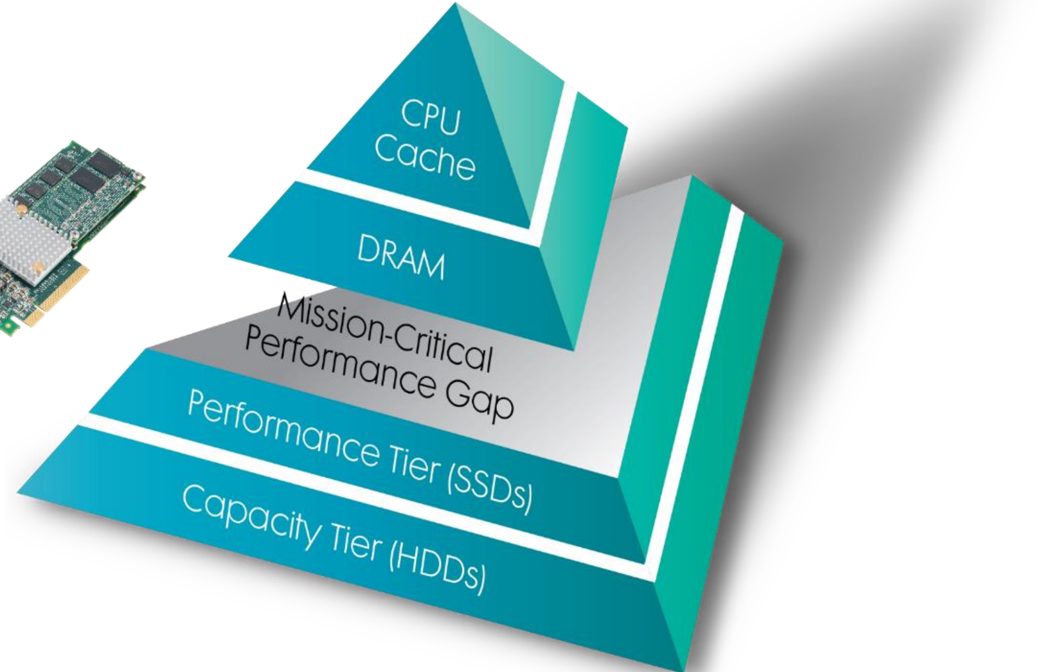




CPU Cache

DRAM

Mission-Critical Performance Gap

Performance Tier (SSDs)

Capacity Tier (HDDs)

# NVRAM & Software Defined Storage

Fabian Trumper

Senior Product Marketing Manager, Performance Solutions Group

# What is Software-Defined-Storage?

- Logical Storage Layer
- Abstracts HW differences
- Pools together resources
- Uses commercially available HW
- Can be Block, File or Object
- Can be Scale-Out or Scale-Up

ceph

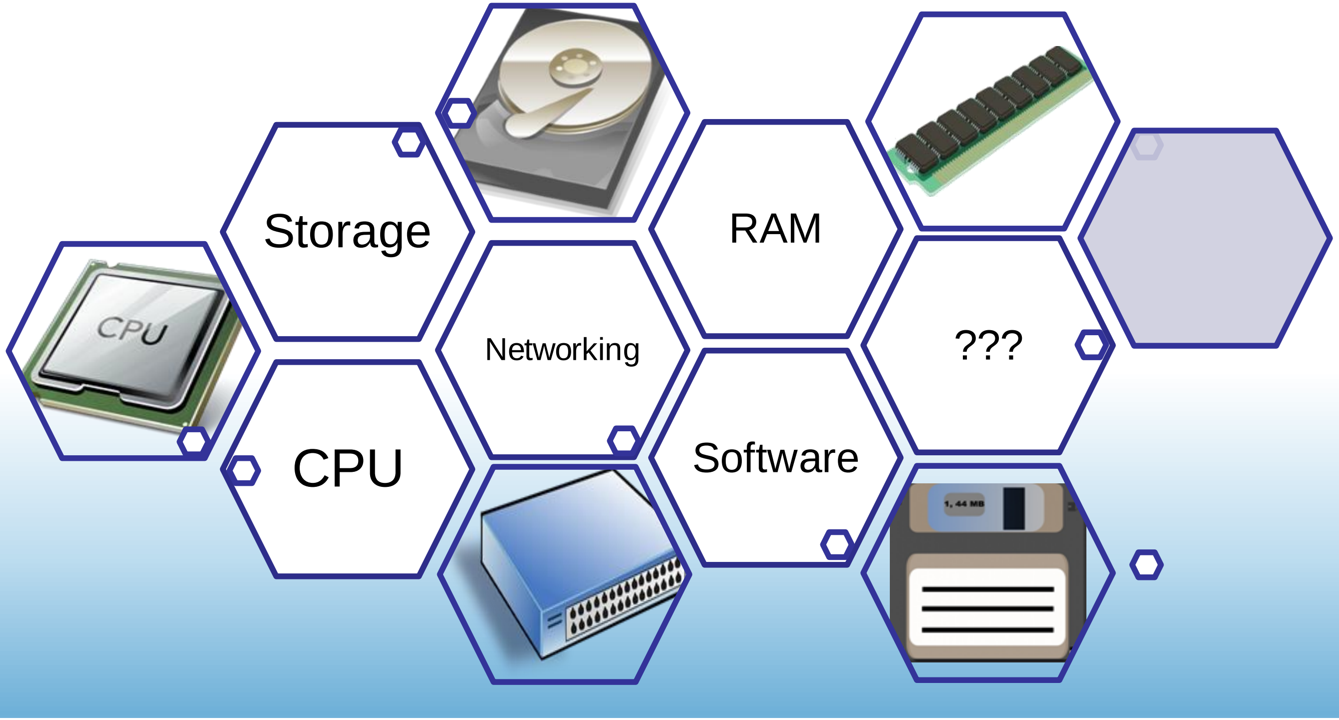
Gluster FS

SwiftStack

OpenZFS

lustre

# Building Blocks for SDS



Storage

RAM

Networking

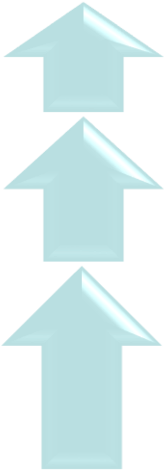
???

CPU

Software

Are you making the most of it?

# Two Approaches to Scaling



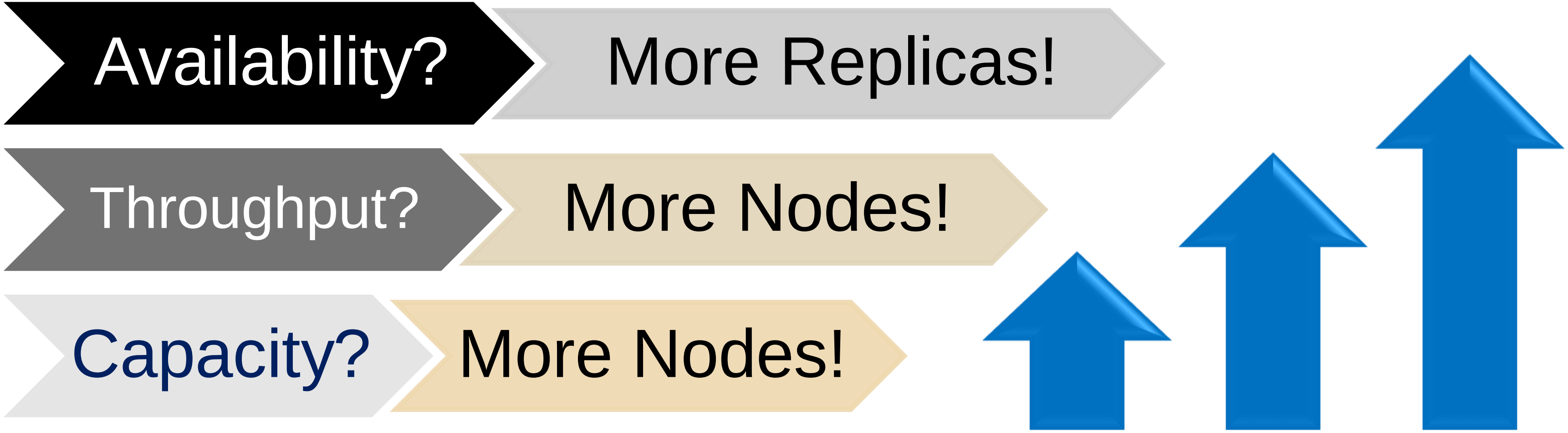
Scale Up

Scale Out

Which one is better?

Depends on what you want to achieve

# Scale Out Approach



Availability? More Replicas!

Throughput? More Nodes!

Capacity? More Nodes!

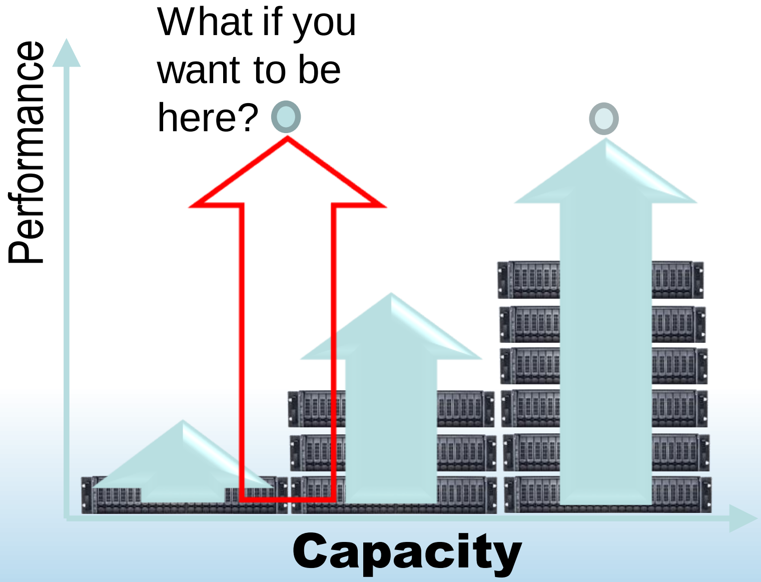
Simple, but it works…

(if you have very smart software to make it simple)

# Can Scale-Out solve everything?

## Over Provisioning Costs

- Racks
- Network
- Power
- Cooling
- Rent, etc

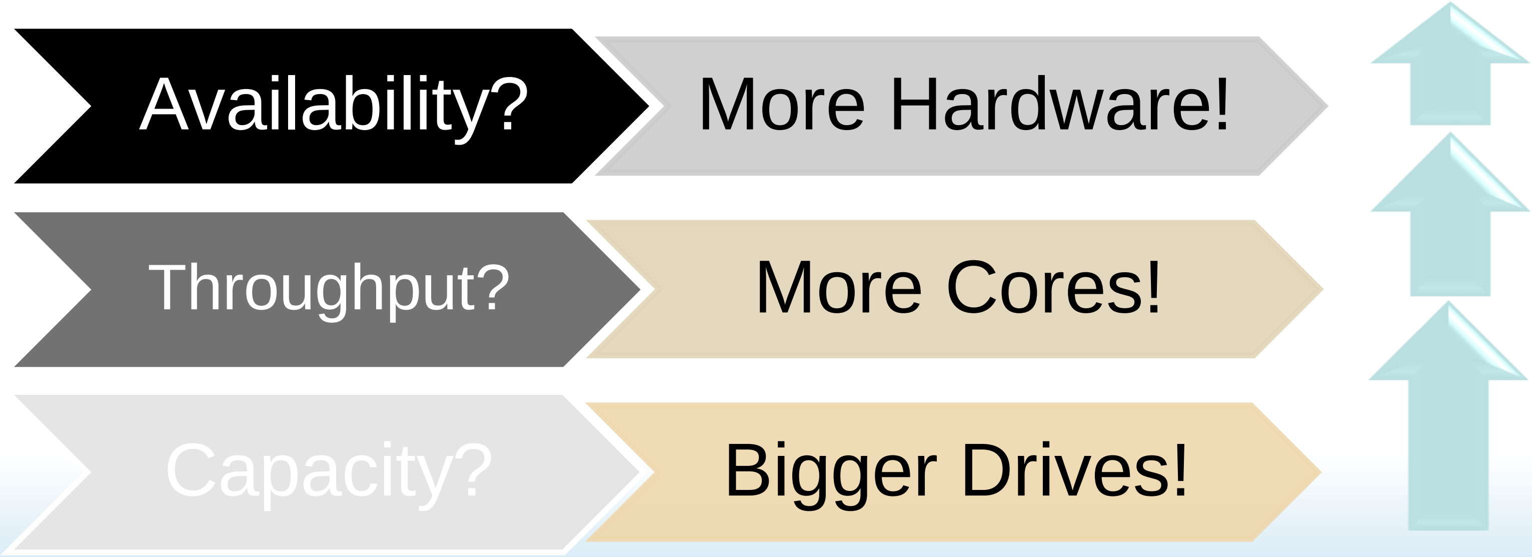


## Diminishing returns on investment

- Difficult to get it just right; efficiency decreases
- Some tasks aren't easily parallelized!

Amdahl's Law: $$\frac{1}{\alpha + \frac{1-\alpha}{P}}$$

# Scale up Approach



Availability? More Hardware!

Throughput? More Cores!

Capacity? Bigger Drives!

Mostly Hardware Options

Not very popular among the Software Define Storage folks

But can we take a page of that book to the SDS world?

# What Defines Performance?

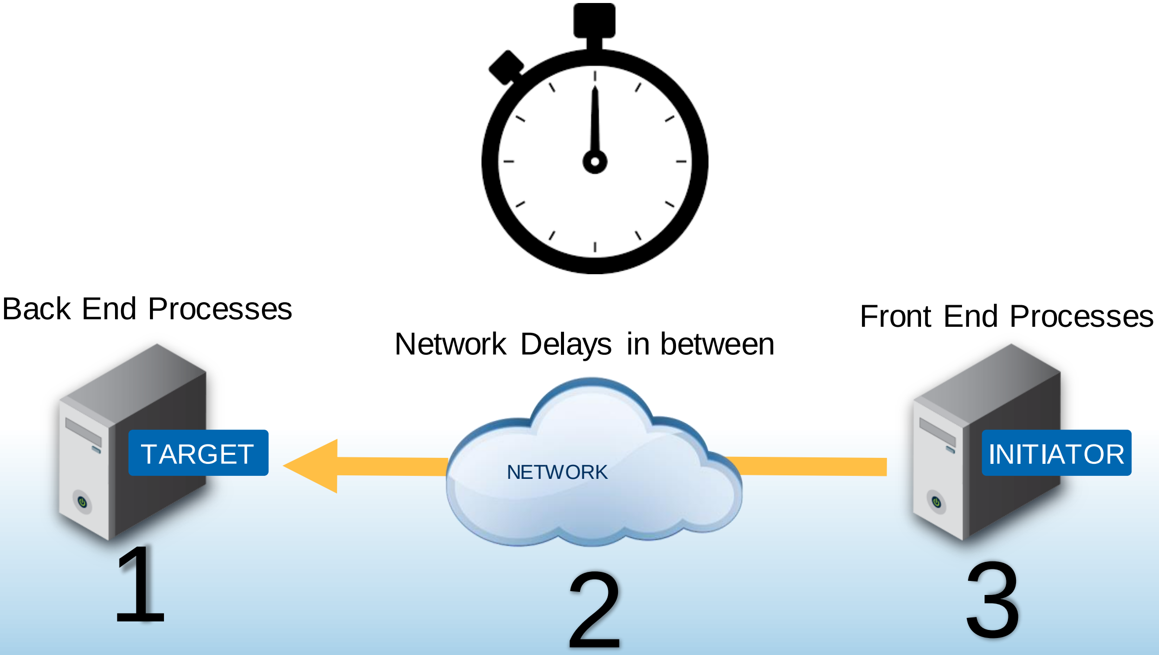
*Throughput = Bandwidth / Latency*

*What if…*

Double your bandwidth → 2X Throughput!

Cut your latency by half → 2X Throughput!

# Let’s look at latency first…



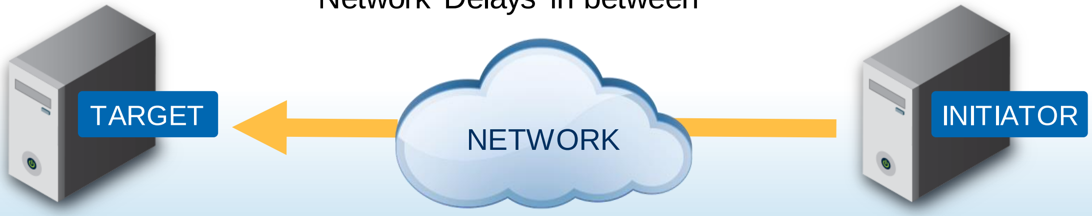
# Can We reduce Latency?

## Caching!

Back End Processes

Network Delays in between

Front End Processes



Read Cache?

Write Cache?

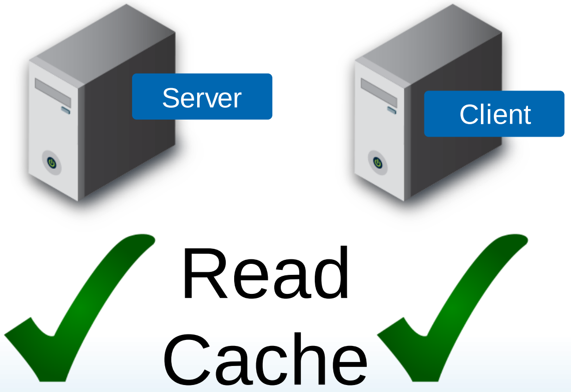
Read Cache?

Write Cache?

# Read Caching works quite well

## SSD’s are perfect for this!

- Large Capacity → Better Hit Ratio
- Low Latency → Better Throughput
- Works for both Client and Server
- DRAM is used extensively as well



Server

Client

✓ Read Cache ✓

# Write Caching is More challenging

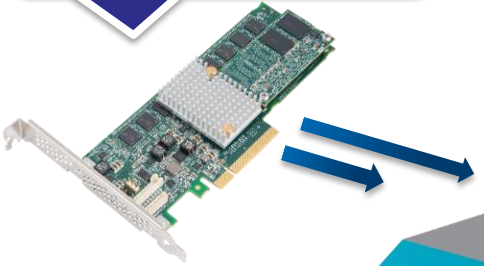
## What about a Write Cache?

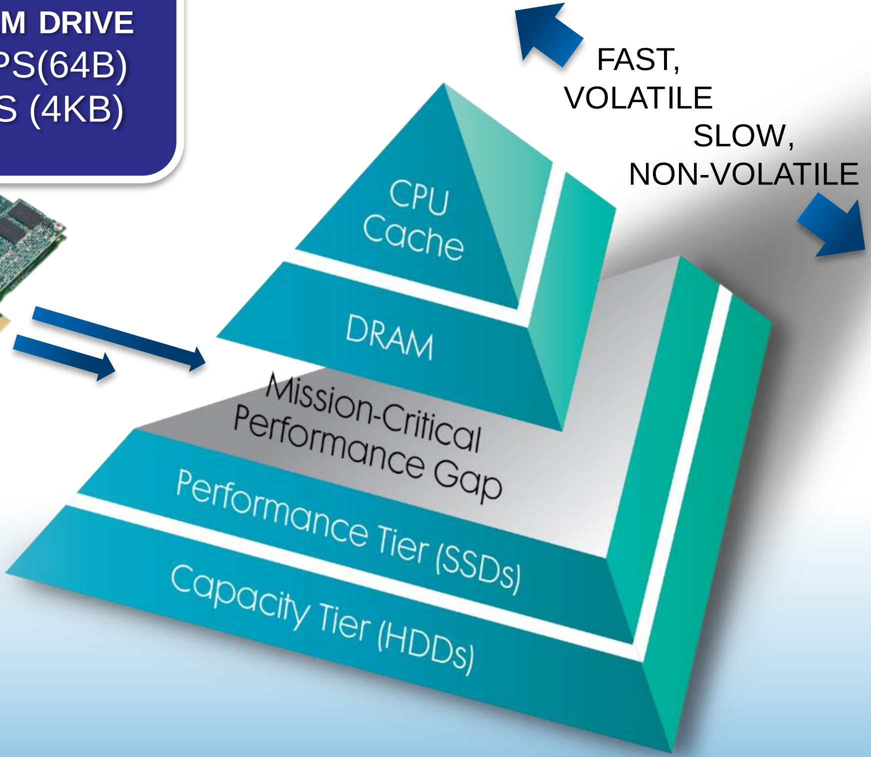
- Adoption of SSD’s as primary storage calls for a new layer to handle Write Caching!

? Write Cache ✗

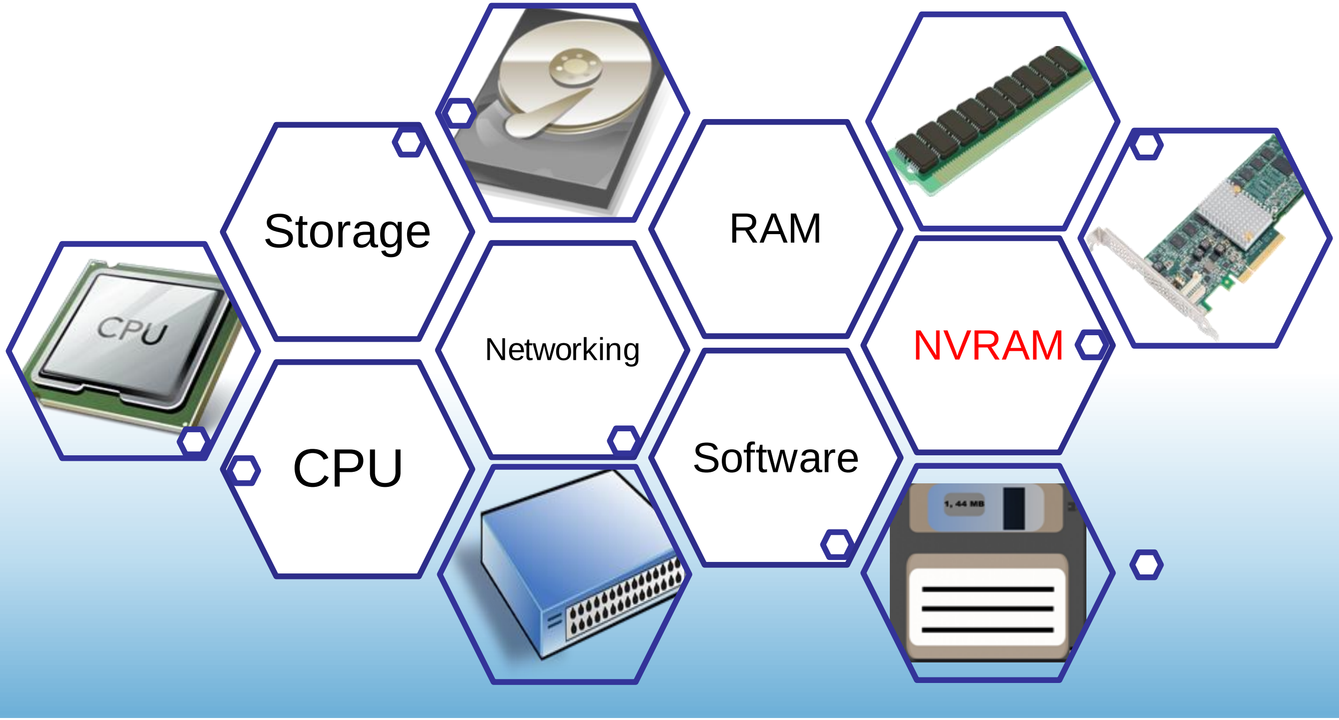
# FLASHTEC™ NVRAM DRIVES
# ESTABLISHING A NEW STORAGE TIER

**Flashtec NVRAM DRIVE**
10 Million IOPS(64B)
1 Million IOPS (4KB)





# Adding NVRAM to the mix



Storage

CPU

Networking

RAM

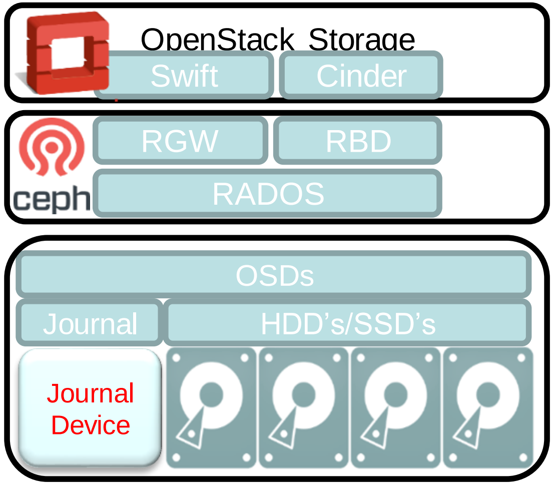
Software

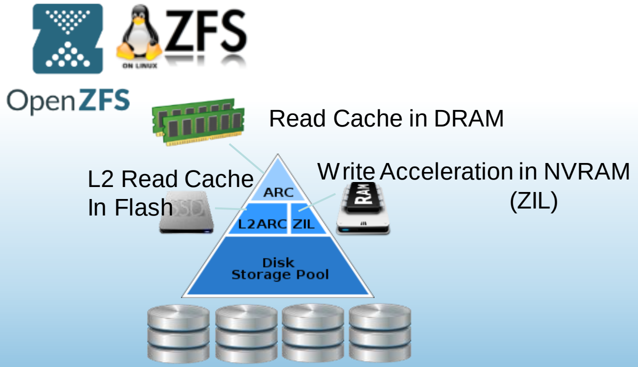
NVRAM

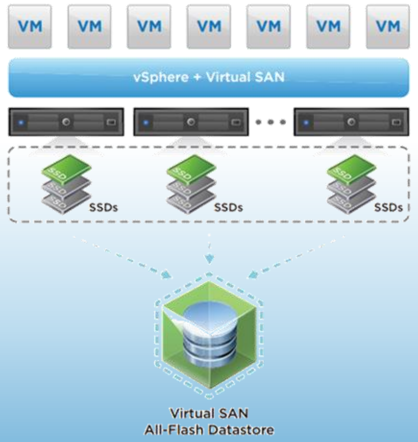
## What are the benefits?

# Write Caching Implementations

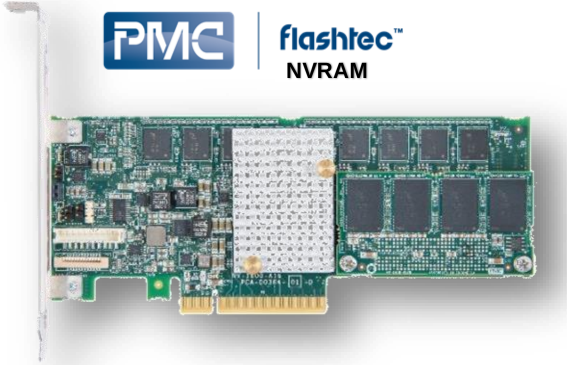
- Examples of Write Cache Usage
  - ZFS uses ZIL (ZFS Intent Log)
  - Ceph OSD’s Journaling Device
  - VMWare VSAN Write Cache



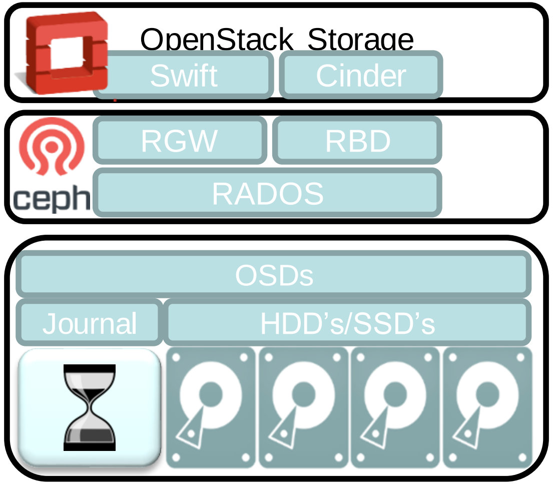




# Ceph Case Study



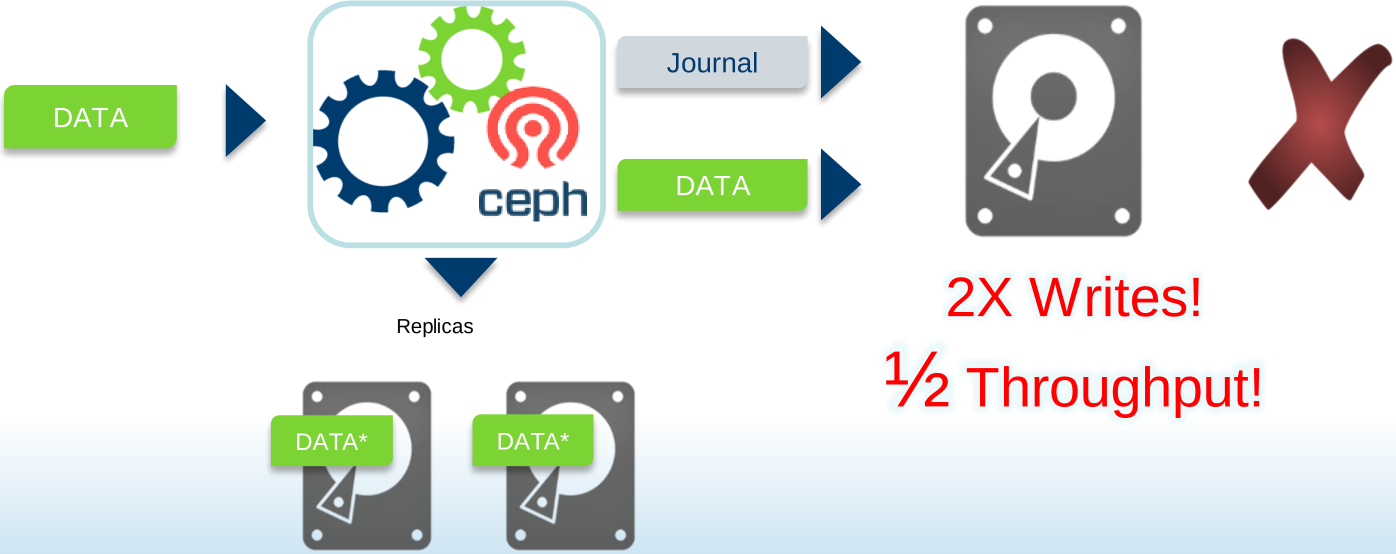
+



How much does NVRAM benefit Ceph?

# Journaling @ Ceph

Every CEPH OSD Writes Twice!

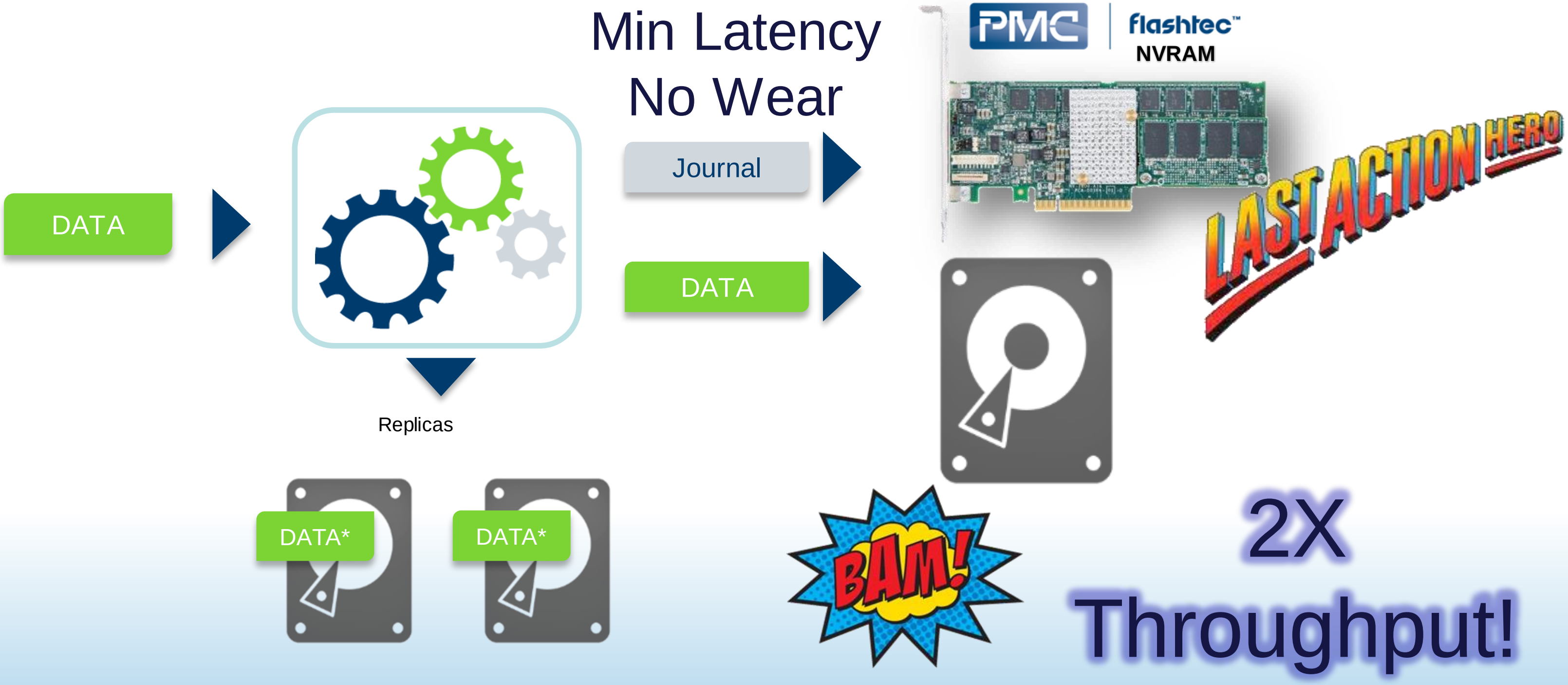


2X Writes!

½ Throughput!

+ Replica Copies

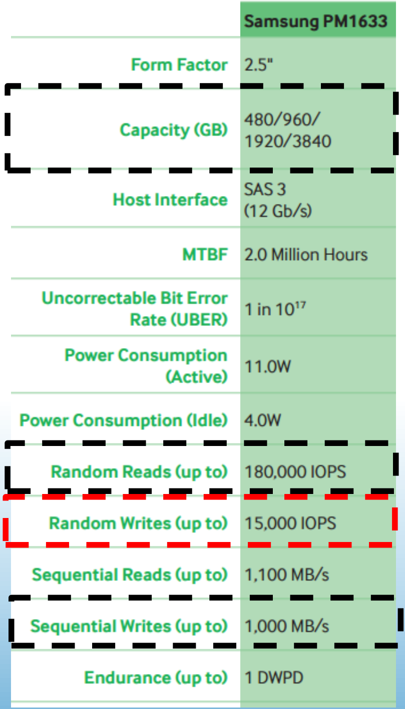
# NVRAM to the rescue!



Min Latency
No Wear

2X
Throughput!

# Can we do even better?

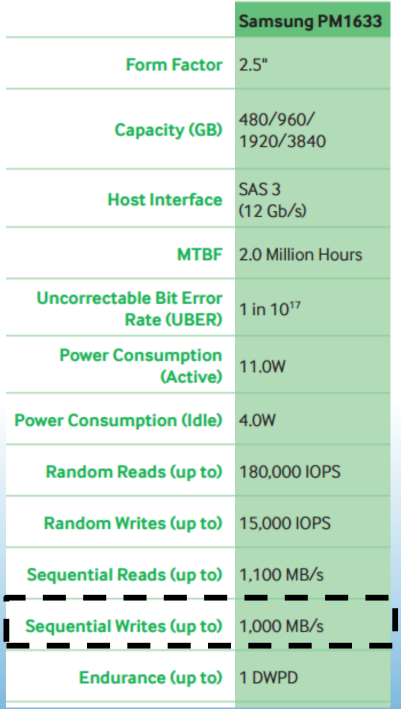


|  | Samsung PM1633 |
|---|---|
| Form Factor | 2.5" |
| Capacity (GB) | 480/960/ 1920/3840 |
| Host Interface | SAS 3 (12 Gb/s) |
| MTBF | 2.0 Million Hours |
| Uncorrectable Bit Error Rate (UBER) | 1 in $10^{17}$ |
| Power Consumption (Active) | 11.0W |
| Power Consumption (Idle) | 4.0W |
| Random Reads (up to) | 180,000 IOPS |
| Random Writes (up to) | 15,000 IOPS |
| Sequential Reads (up to) | 1,100 MB/s |
| Sequential Writes (up to) | 1,000 MB/s |
| Endurance (up to) | 1 DWPD |

- 3.84TB Capacity – GREAT! ✓
- 180K IOPS Random Read – GREAT! ✓
- 15K Random Write (60 MB/s) – FAIL! ✗

!?

# Write Caching + Re-Ordering



| | Samsung PM1633 |
|---|---|
| Form Factor | 2.5" |
| Capacity (GB) | 480/960/<br>1920/3840 |
| Host Interface | SAS 3<br>(12 Gb/s) |
| MTBF | 2.0 Million Hours |
| Uncorrectable Bit Error Rate (UBER) | 1 in $10^{17}$ |
| Power Consumption (Active) | 11.0W |
| Power Consumption (Idle) | 4.0W |
| Random Reads (up to) | 180,000 IOPS |
| Random Writes (up to) | 15,000 IOPS |
| Sequential Reads (up to) | 1,100 MB/s |
| Sequential Writes (up to) | 1,000 MB/s |
| Endurance (up to) | 1 DWPD |

1 GB/s Sequential Write! – GREAT! ✓

Can we turn Random Workloads to Sequential?

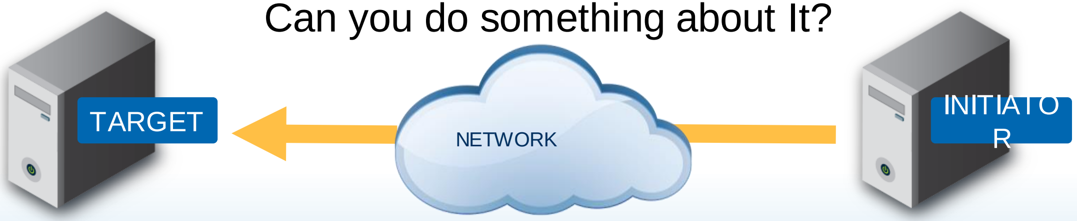
HDD’s have been doing it for years!

With an NVRAM memory layer, SDS Storage Stacks could use these age old techniques to improve throughput and reduce costs at the same time

# What About Network Latency?

Round-Trip Time + Networking Protocol Delay

Can you do something about It?



RDMA PeerDirect + NVRAM

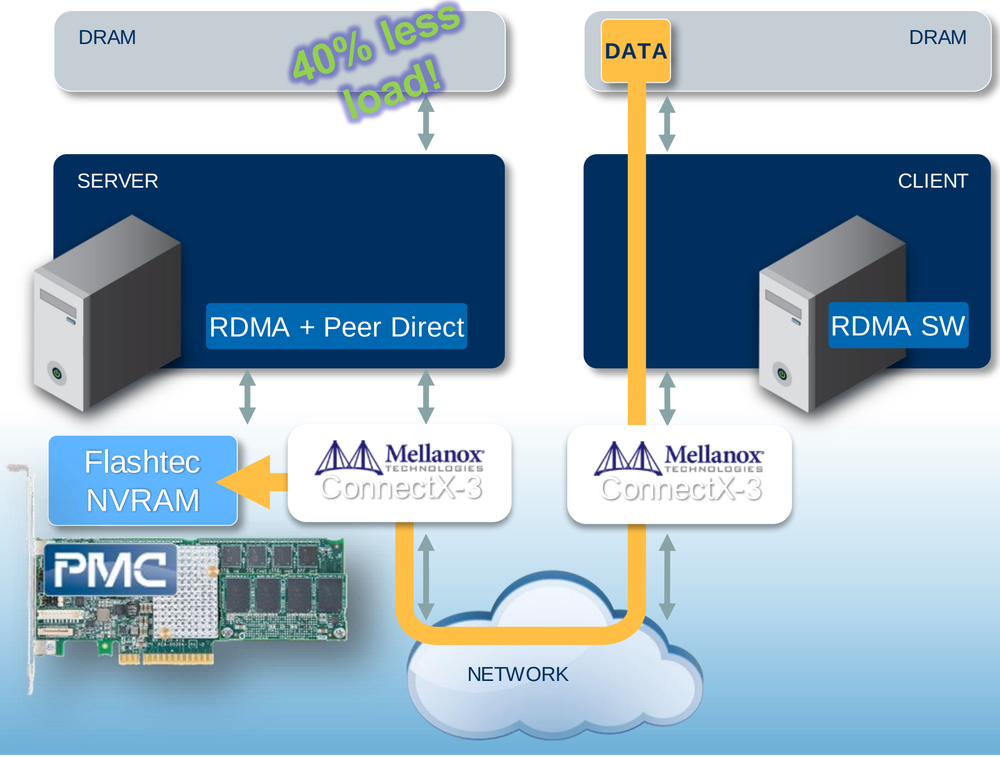
NVMe Over Fabrics in the future

# RDMA + PeerDirect + NVRAM

LIVE DEMO

Booth 213

**4KB Write Latency: 7us**



40% less load!

PMC

DELIVERING THE WORLD'S FASTEST ENTERPRISE STORAGE

See PMC @ Booth 213

PMC

flashtec™

Thank You!