RED HAT SUMMIT

GLUSTER CAN DO THAT!

Architecting and Performance Tuning Efficient Gluster Storage Pools

Dustin Black
Senior Architect, Software-Defined Storage
@dustinlblack

Ben Turner
Principal Quality Engineer
@bennyturns

2017-05-02

GLUSTER 101
IN 5 SECONDS

THE DATA EXPLOSION

WEB, MOBILE, SOCIAL MEDIA, CLOUD
Our digital assets have grown exponentially due to web scale services like Facebook, Flickr, Snapchat, YouTube, and Netflix.

VIDEO ON-DEMAND SERVICES
Rapid growth of video on-demand has culminated in 50% of households using this service.

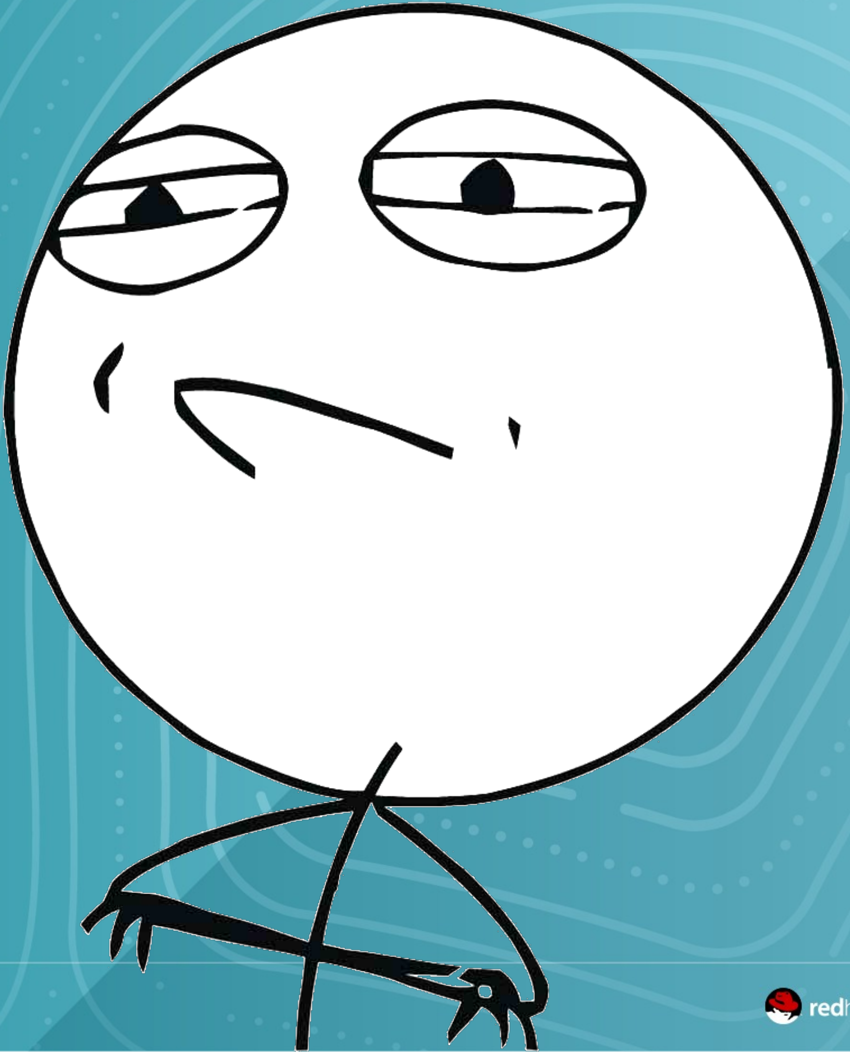
MEDIA AND ENTERTAINMENT INDUSTRIES
A staggering amount of content is created during today's optimized production processes.

MEDICAL INDUSTRY
Medical imaging needs are vast, and regulatory requirements can be demanding.

NOT SURE IF YOU GOT IT?

https://people.redhat.com/dblack/summit2017

GLUSTER
CAN DO THAT!*



*If you build it right

A SIX-NODE POOL CAN PROCESS...

JPEG Web
Image Files
(32KB)

72x 7.2K HDD

1,700 JPEGs
per second

or

Optimized
72x 7.2K HDD

12,000 JPEGs
per second

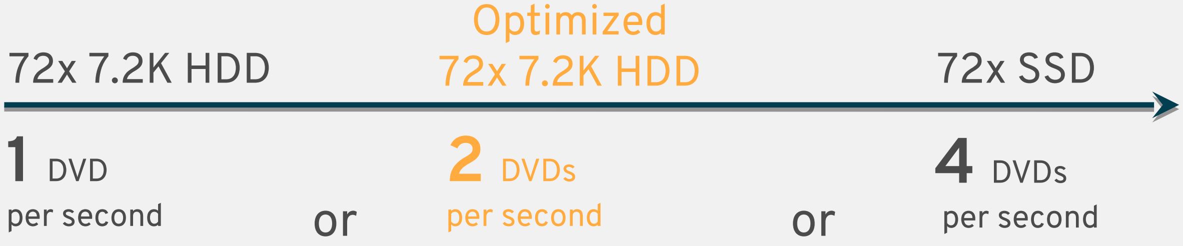
or

72x SSD

23,000 JPEGs
per second

OR...

DVD
Movie Files
(4GB)



72x 7.2K HDD

1 DVD
per second

or

Optimized
72x 7.2K HDD

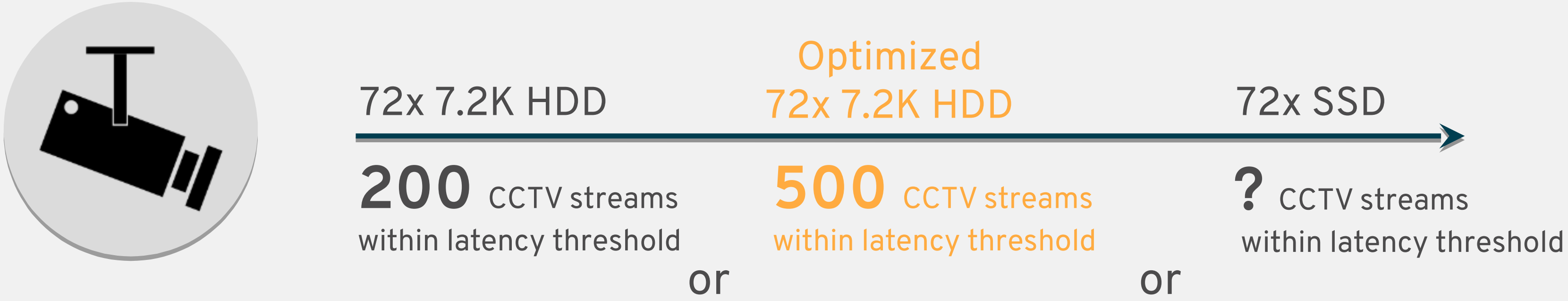
2 DVDs
per second

or

72x SSD

4 DVDs
per second

OR...



High-Def
CCTV Camera
Recording Streams

72x 7.2K HDD

200 CCTV streams
within latency threshold

or

Optimized
72x 7.2K HDD

500 CCTV streams
within latency threshold

or

72x SSD

? CCTV streams
within latency threshold

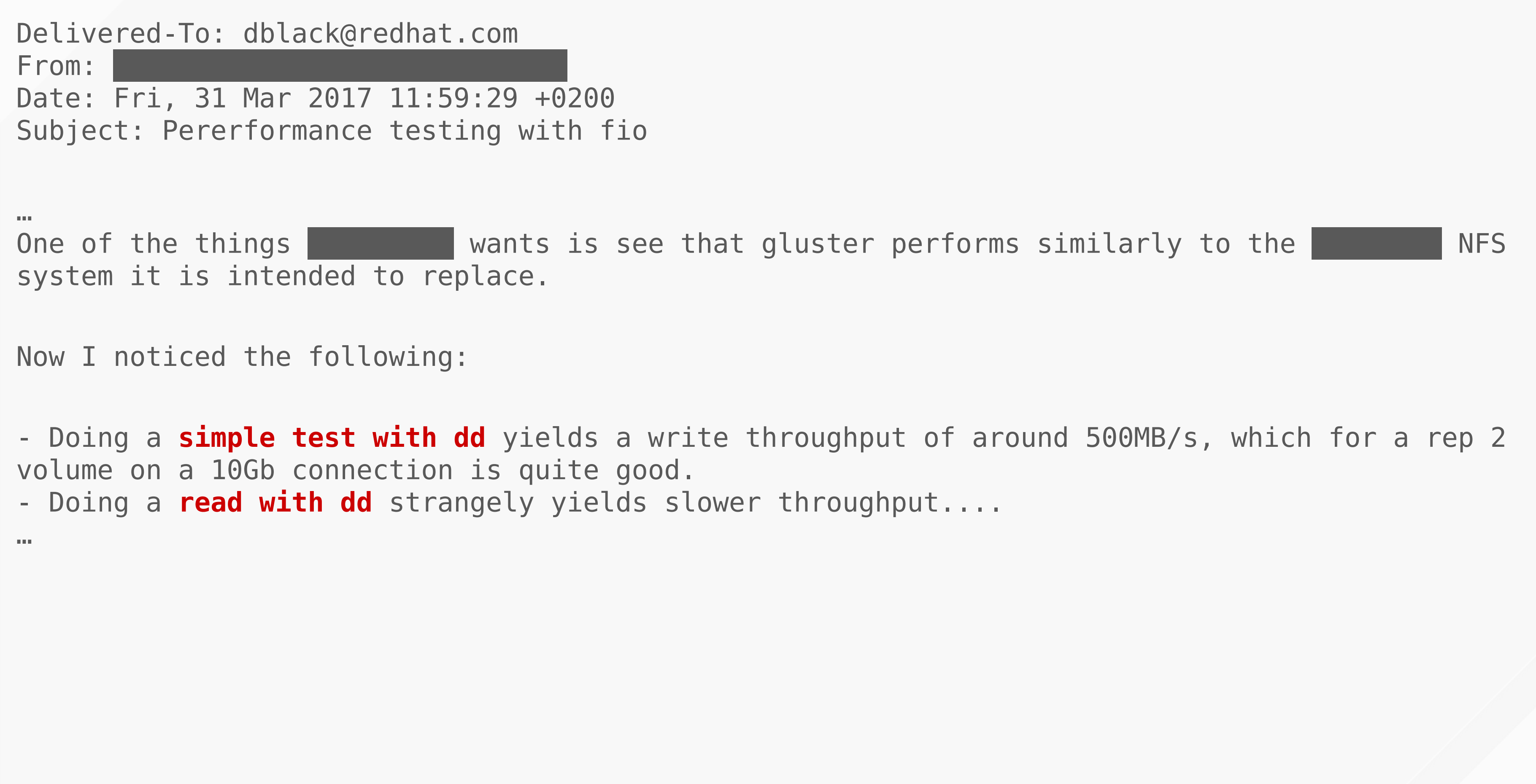
KISS

KEEP
IT
SIMPLE,
STUPID

SWTWD

Start
With
The
Workload,
Dummy

WHY DO YOU ASK THE WRONG QUESTIONS?



Delivered-To: dblack@redhat.com
From: [redacted]
Date: Fri, 31 Mar 2017 11:59:29 +0200
Subject: Pererformance testing with fio

…
One of the things [redacted] wants is see that gluster performs similarly to the [redacted] NFS system it is intended to replace.

Now I noticed the following:

- Doing a **simple test with dd** yields a write throughput of around 500MB/s, which for a rep 2 volume on a 10Gb connection is quite good.
- Doing a **read with dd** strangely yields slower throughput....

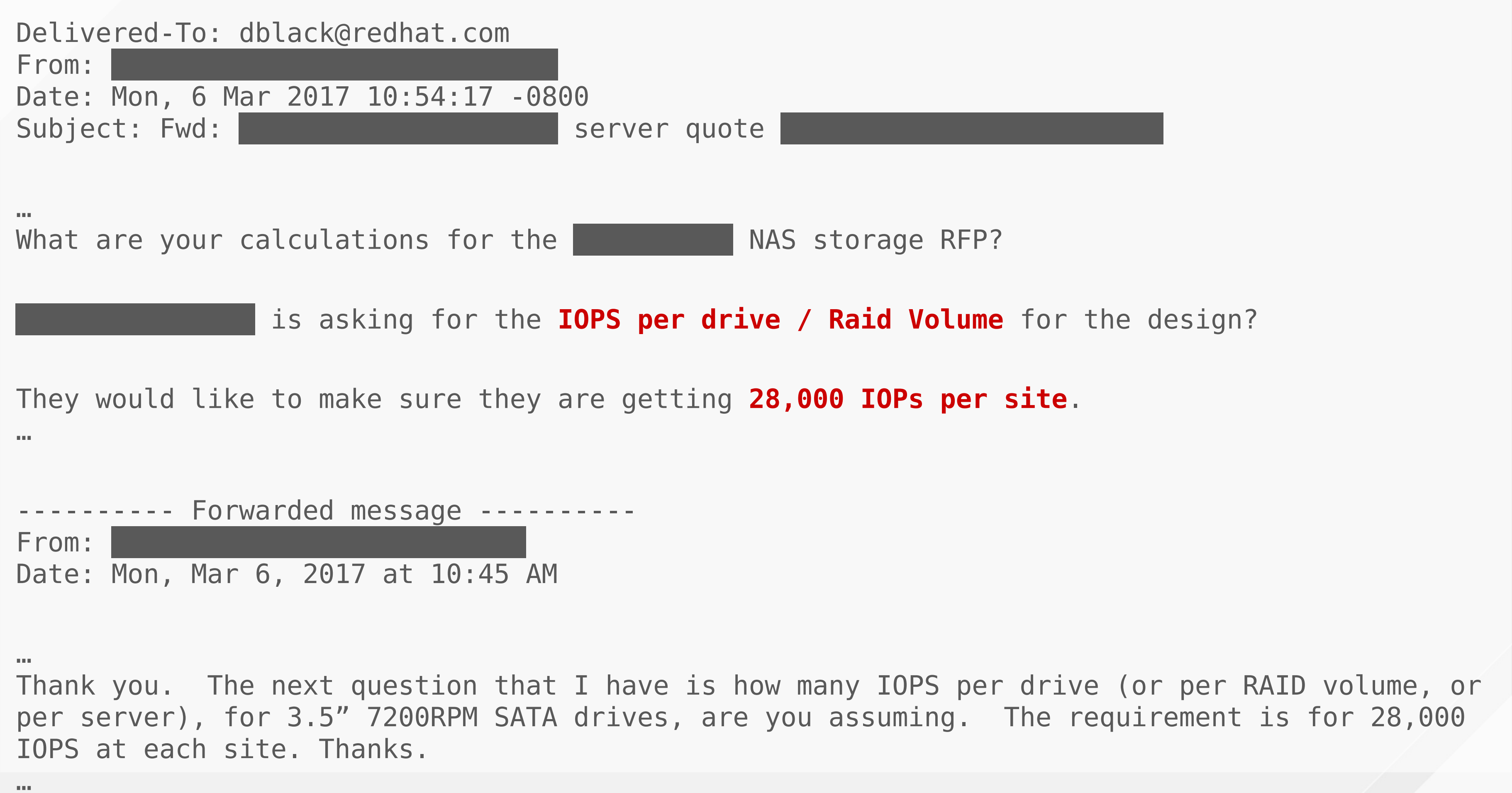
…

Delivered-To: dblack@redhat.com
From: [redacted]
Date: Sun, 5 Feb 2017 20:16:40 +0900
Subject: RHGS scale-out options

…
[redacted] plans to **add physical nodes to increase "performance"**
(currently [redacted] is experiencing performance problem)
…
Current Env : 80 X 2-way distributed replicated vols on 6 nodes
To-Be : add 6 more nodes... becomes 80 X 2-way distributed replicated vols on 12 nodes

I'm not sure which one is the best way to increase performance.

1. extend current cluster from 6 to 12 nodes and add bricks from new 6 nodes into existing 80 vols
2. extend current cluster from 6 to 12 nodes and migrate some vols to new new 6 nodes.
3. create another RHGS gluster cluster with new 6 nodes and migrate some vols to new RHGS cluster
4. ??
…



Delivered-To: dblack@redhat.com
From: [redacted]
Date: Mon, 6 Mar 2017 10:54:17 -0800
Subject: Fwd: [redacted] server quote [redacted]

…
What are your calculations for the [redacted] NAS storage RFP?

[redacted] is asking for the **IOPS per drive / Raid Volume** for the design?

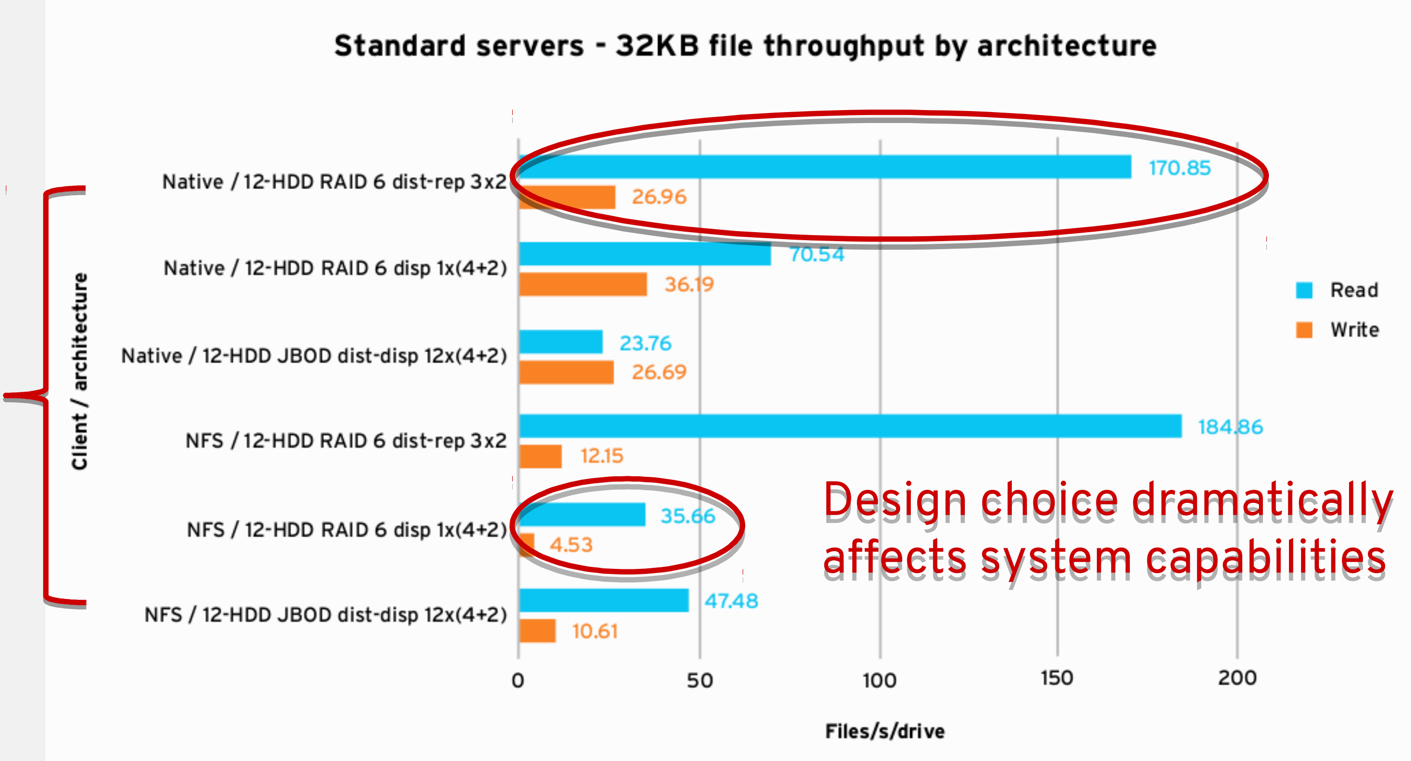
They would like to make sure they are getting **28,000 IOPs per site**.
…

---------- Forwarded message ----------
From: [redacted]
Date: Mon, Mar 6, 2017 at 10:45 AM

…
Thank you. The next question that I have is how many IOPS per drive (or per RAID volume, or per server), for 3.5” 7200RPM SATA drives, are you assuming. The requirement is for 28,000 IOPS at each site. Thanks.
…

THE WORKLOAD IS COMING

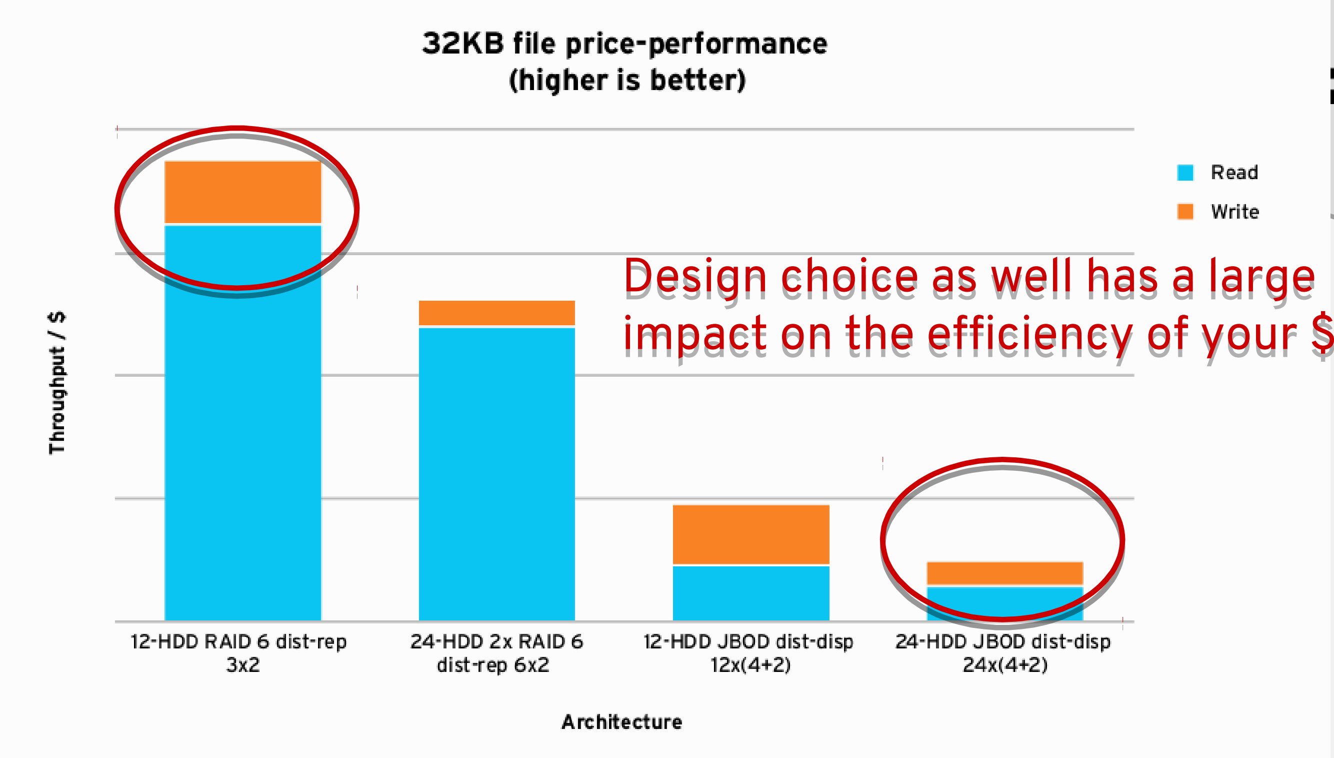
SMALL FILE JPEG WORKLOAD



Same Hardware

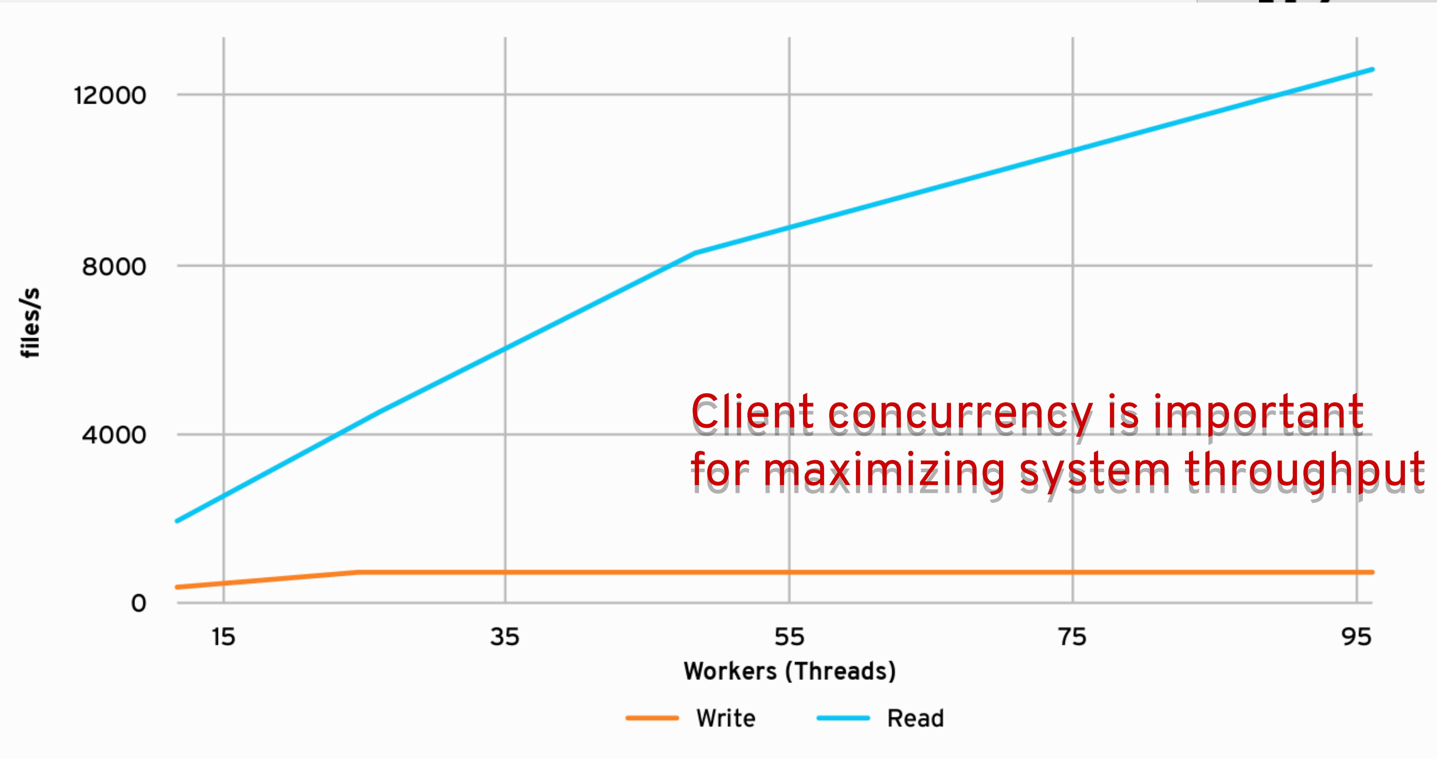
Design choice dramatically affects system capabilities

SMALL FILE JPEG WORKLOAD



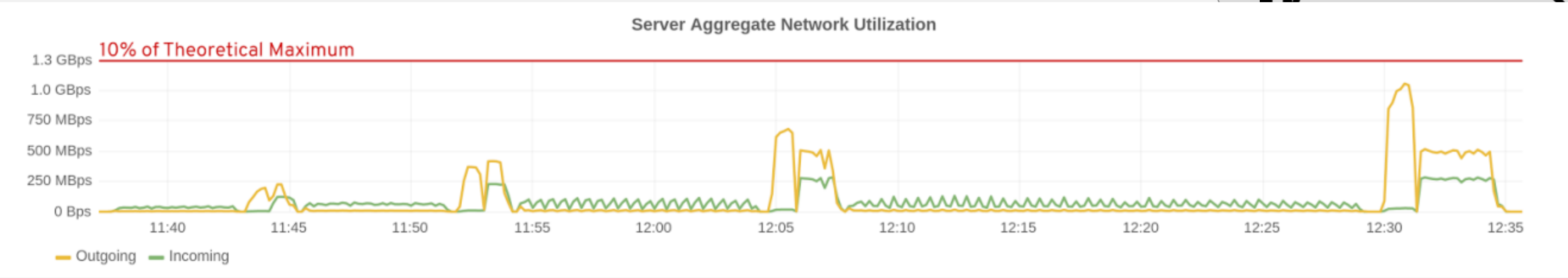
Design choice as well has a large impact on the efficiency of your $$

SMALL FILE JPEG WORKLOAD



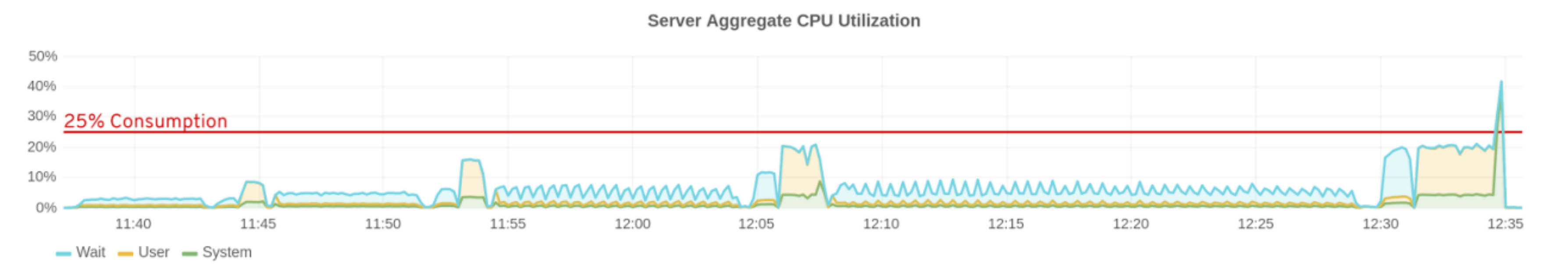
Client concurrency is important for maximizing system throughput

SMALL FILE JPEG WORKLOAD



Server Aggregate Network Utilization

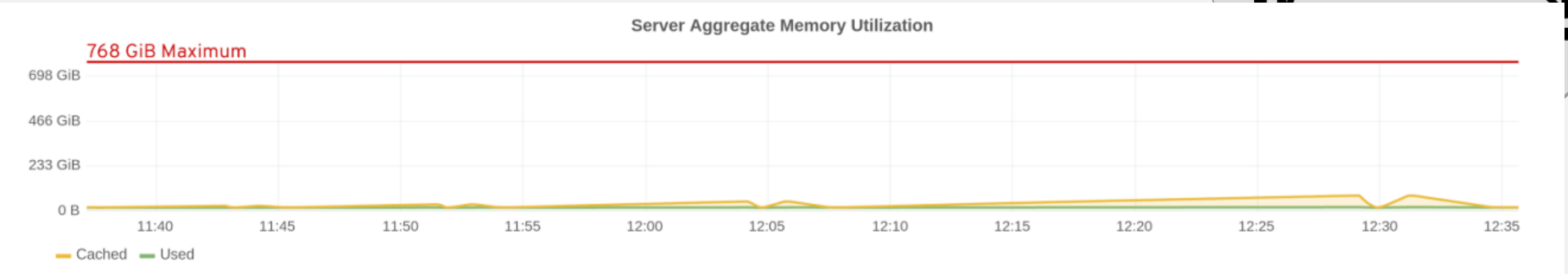
10% of Theoretical Maximum

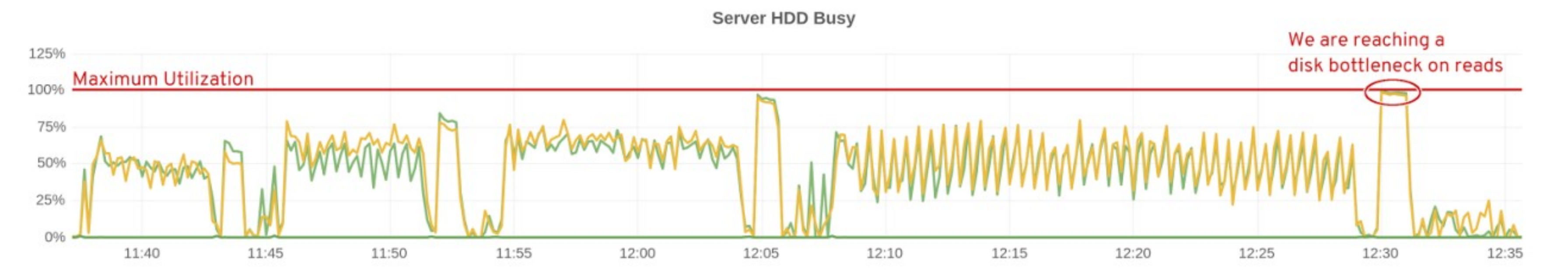


Server Aggregate CPU Utilization

25% Consumption

SMALL FILE JPEG WORKLOAD



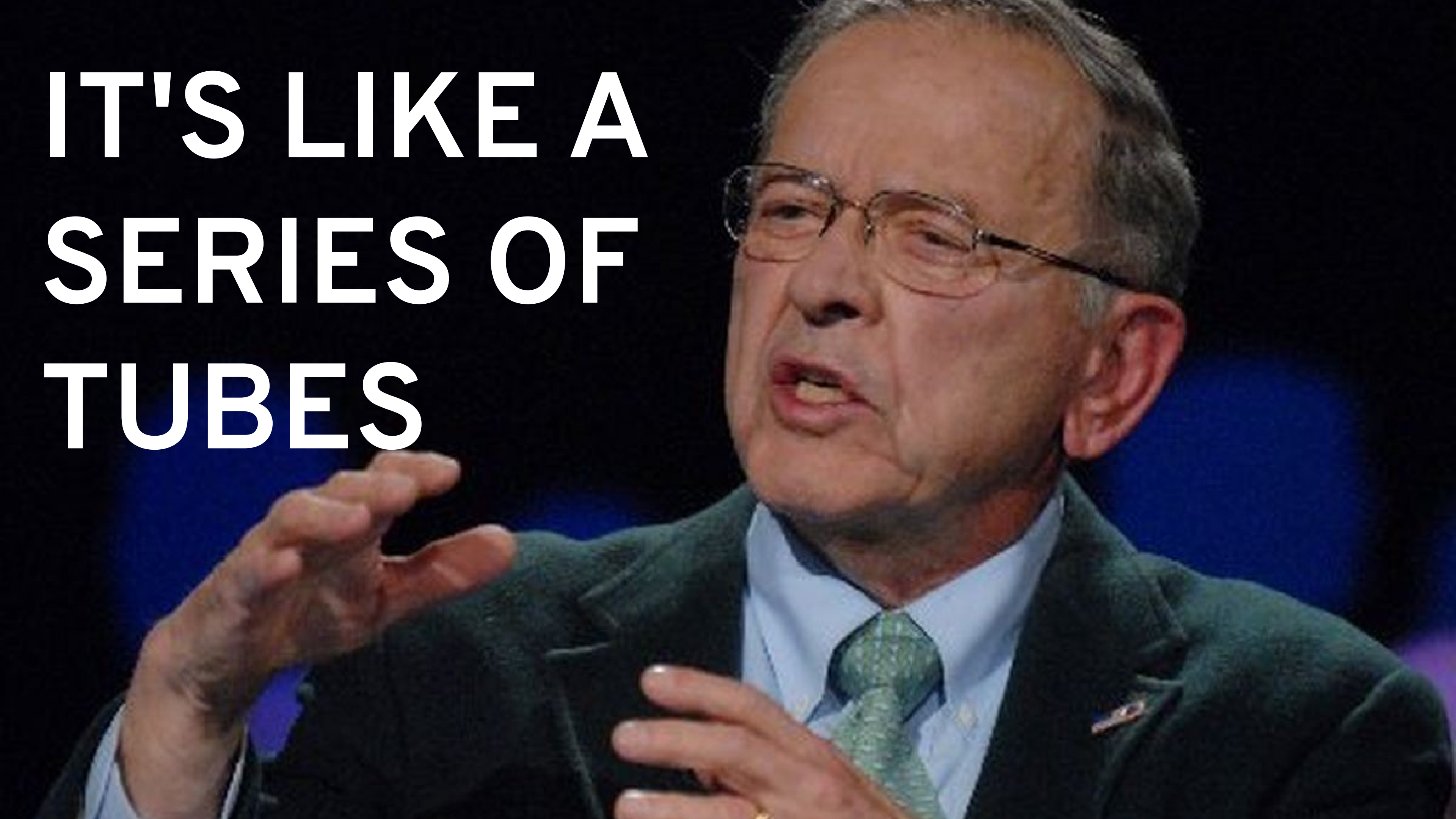


We are reaching a disk bottleneck on reads

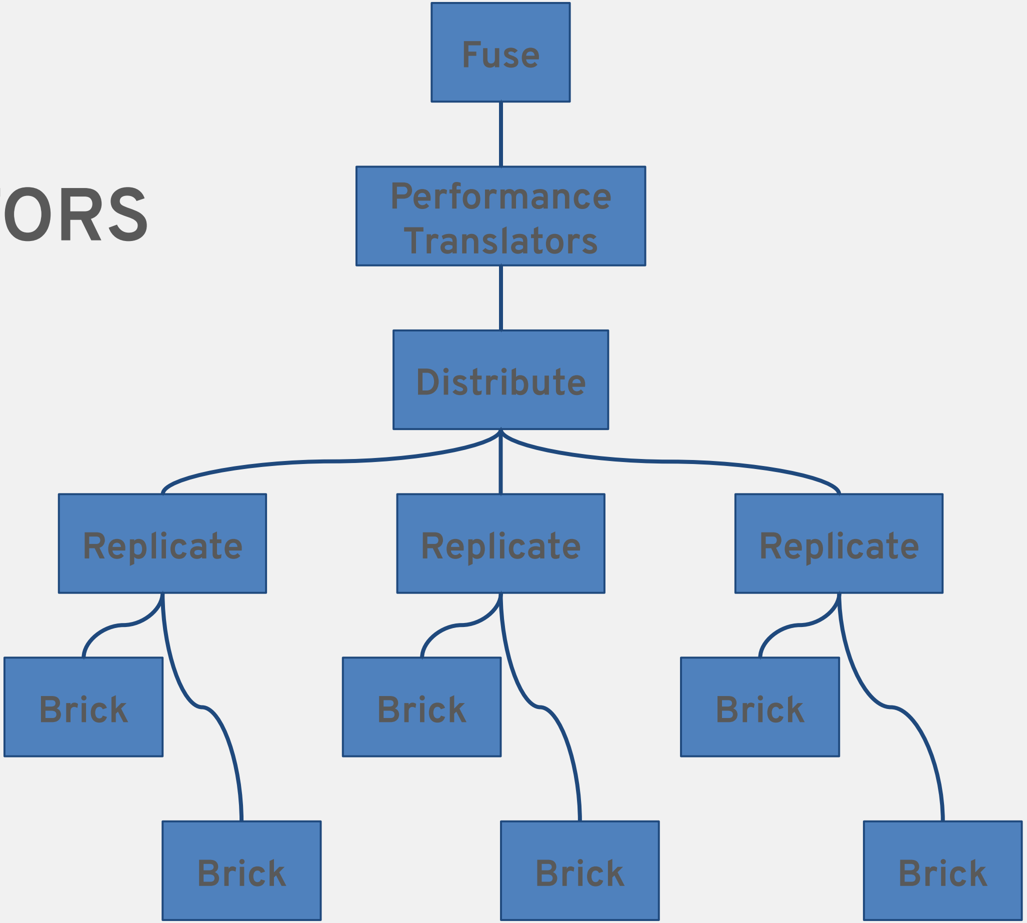
IF A FILE IS VERY VERY SMALL

IS IT STILL A FILE?

IT'S LIKE A SERIES OF TUBES



GLUSTER TRANSLATORS



```
struct xlator_fops fops = {
        .open        = ra_open,
        .create      = ra_create,
        .readv       = ra_readv,
        .writev      = ra_writev,
        .flush       = ra_flush,
        .fsync       = ra_fsync,
        .truncate    = ra_truncate,
        .ftruncate   = ra_ftruncate,
        .fstat       = ra_fstat,
        .discard     = ra_discard,
        .zerofill    = ra_zerofill,
};


struct volume_options options[] = {
        { .key  = {"force-atime-update"},
          .type = GF_OPTION_TYPE_BOOL,
          .default_value = "false"
        },
        { .key  = {"page-count"},
          .type = GF_OPTION_TYPE_INT,
          .min  = 1,
          .max  = 16,
```

SMALL FILE AND METADATA WORKLOADS

What the Gluster community is doing:

Improve efficiency of individual calls

Store metadata in client cache

Prefetch metadata

Compound file operations

Coming Soon! Negative lookups and parallel readdirp

TUNING FOR SMALL FILE & METADATA

Since small file workloads are metadata intensive, I use the same tuning for both.

RAID 10 or RAID 6 are recommended for bricks

Tuned profile: rhgs-throughput-performance

Event Threads = 4

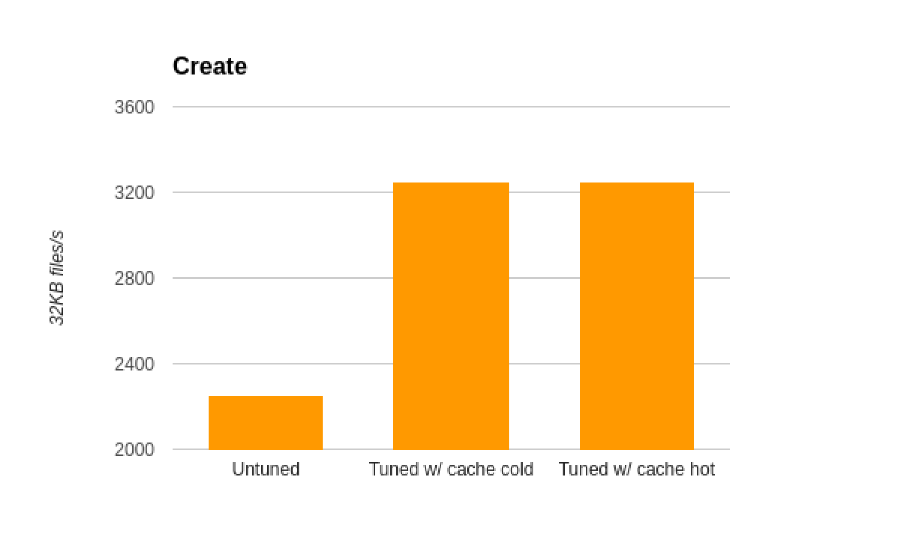
lookup-optimize = on

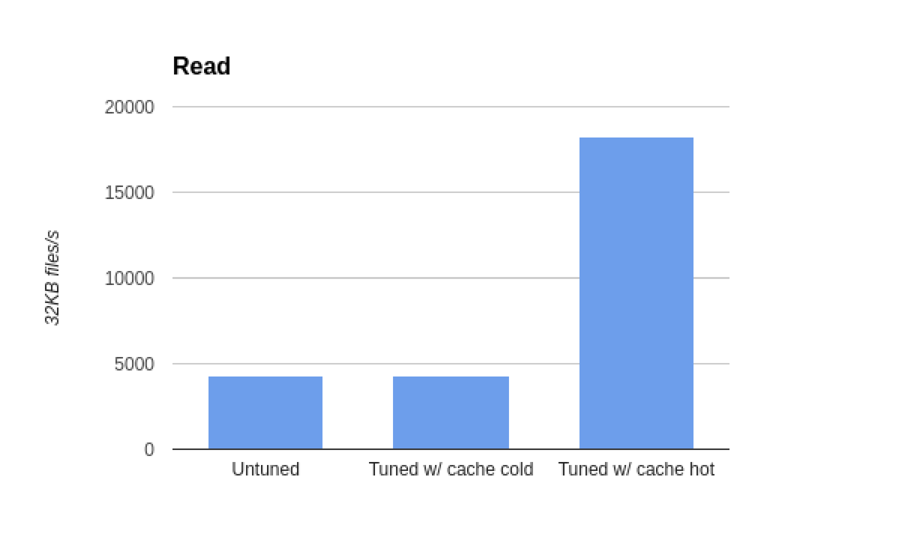
Features.cache-invalidation = on

Performance.stat-prefetch = on

SMALLFILE CREATES & READS

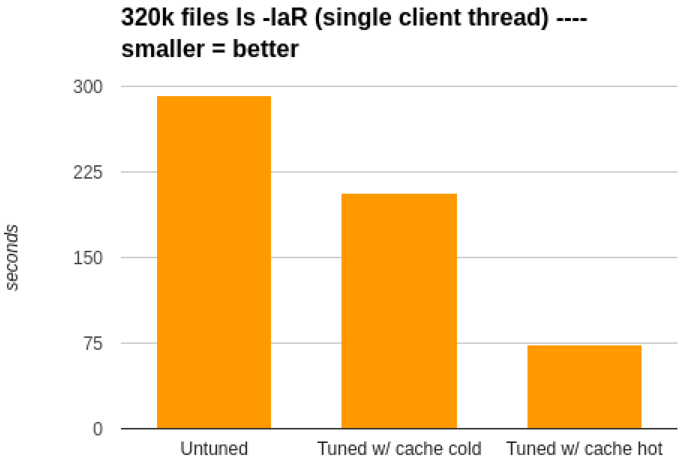
Create & read of 32 KB files
untuned vs tuned w/ cold cache vs tuned w/ hot cache



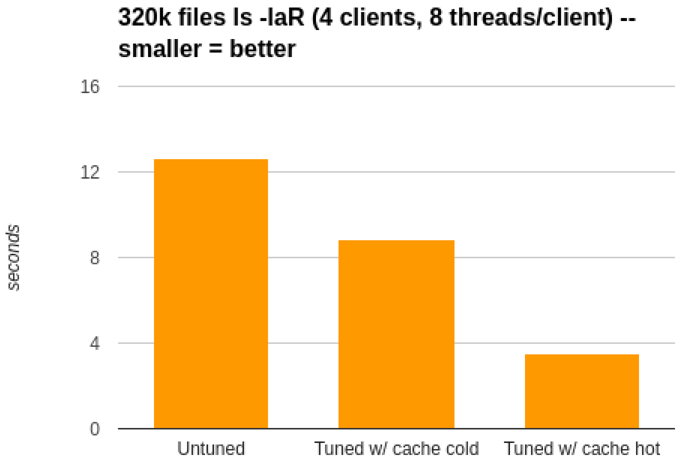


SMALLFILE METADATA WORKLOAD

Single and multi-threaded ls -l workloads
untuned vs tuned w/ cold cache vs tuned w/ hot cache

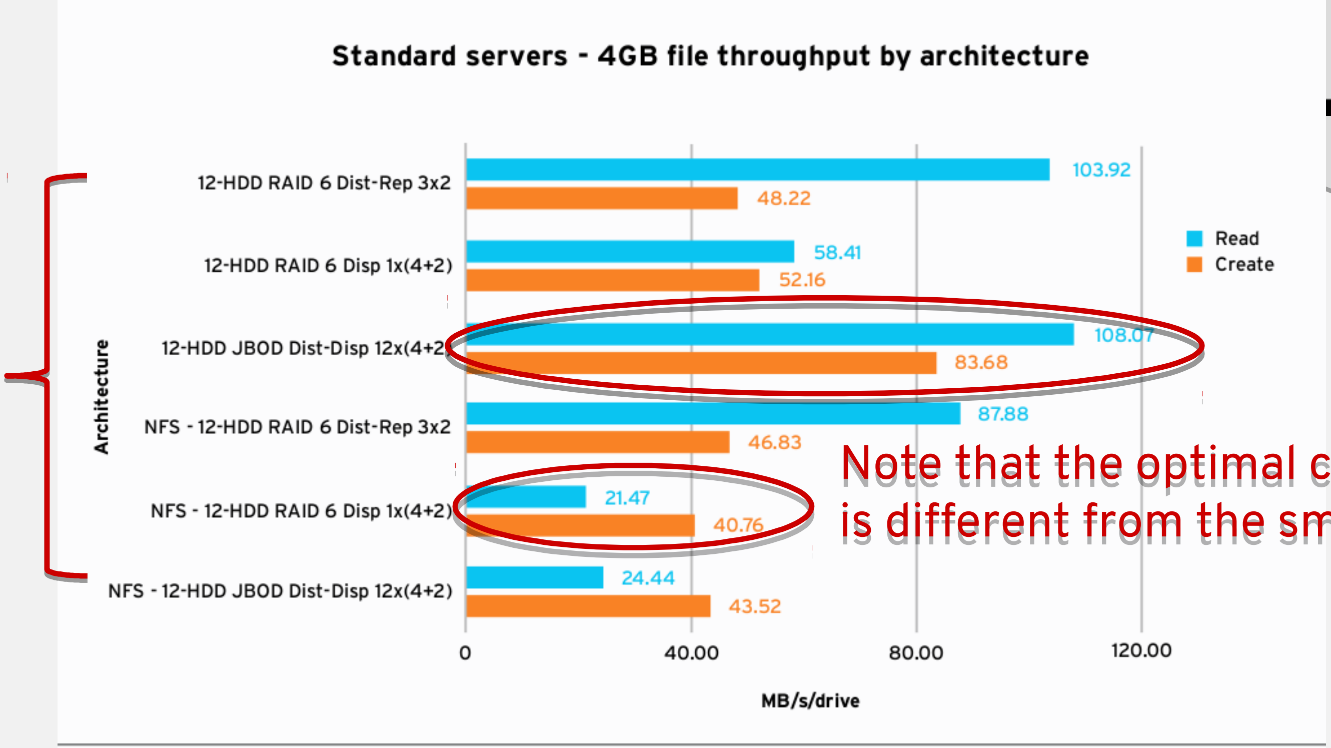


320k files ls -laR (single client thread) ---- smaller = better



320k files ls -laR (4 clients, 8 threads/client) -- smaller = better

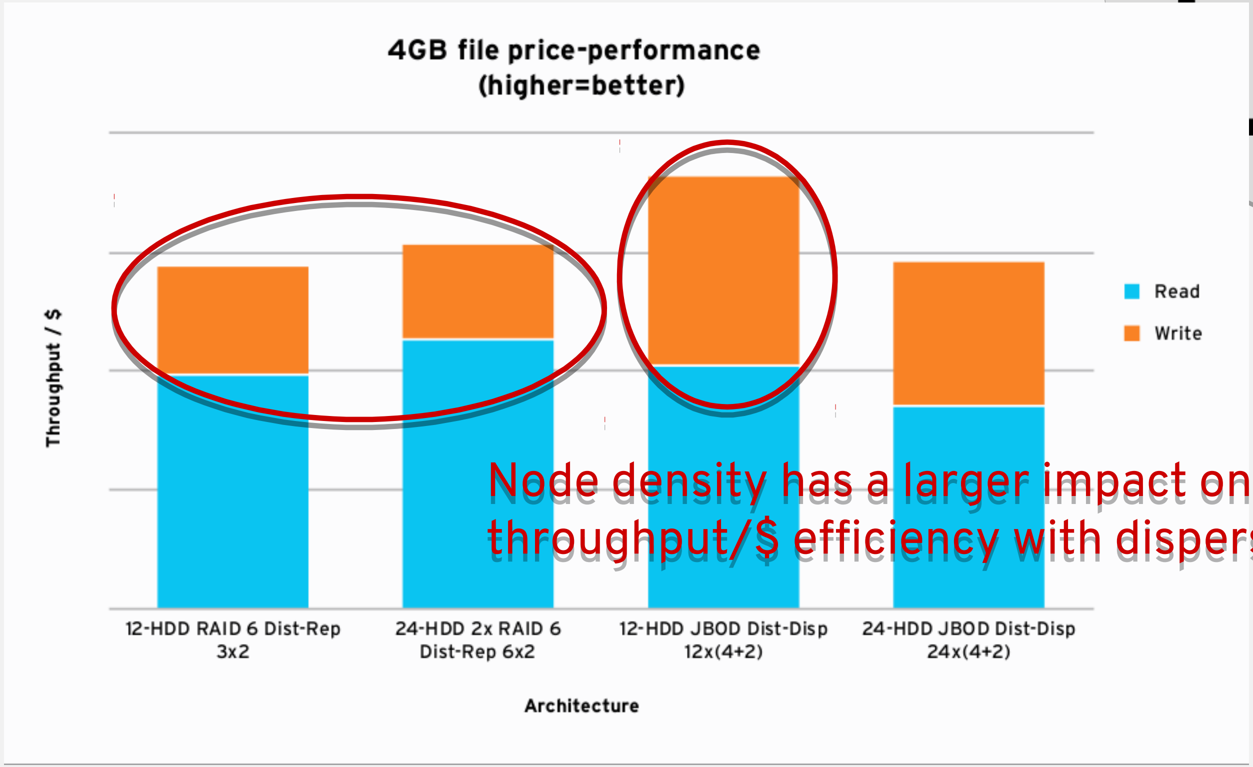
LARGE FILE DVD WORKLOAD



Same Hardware

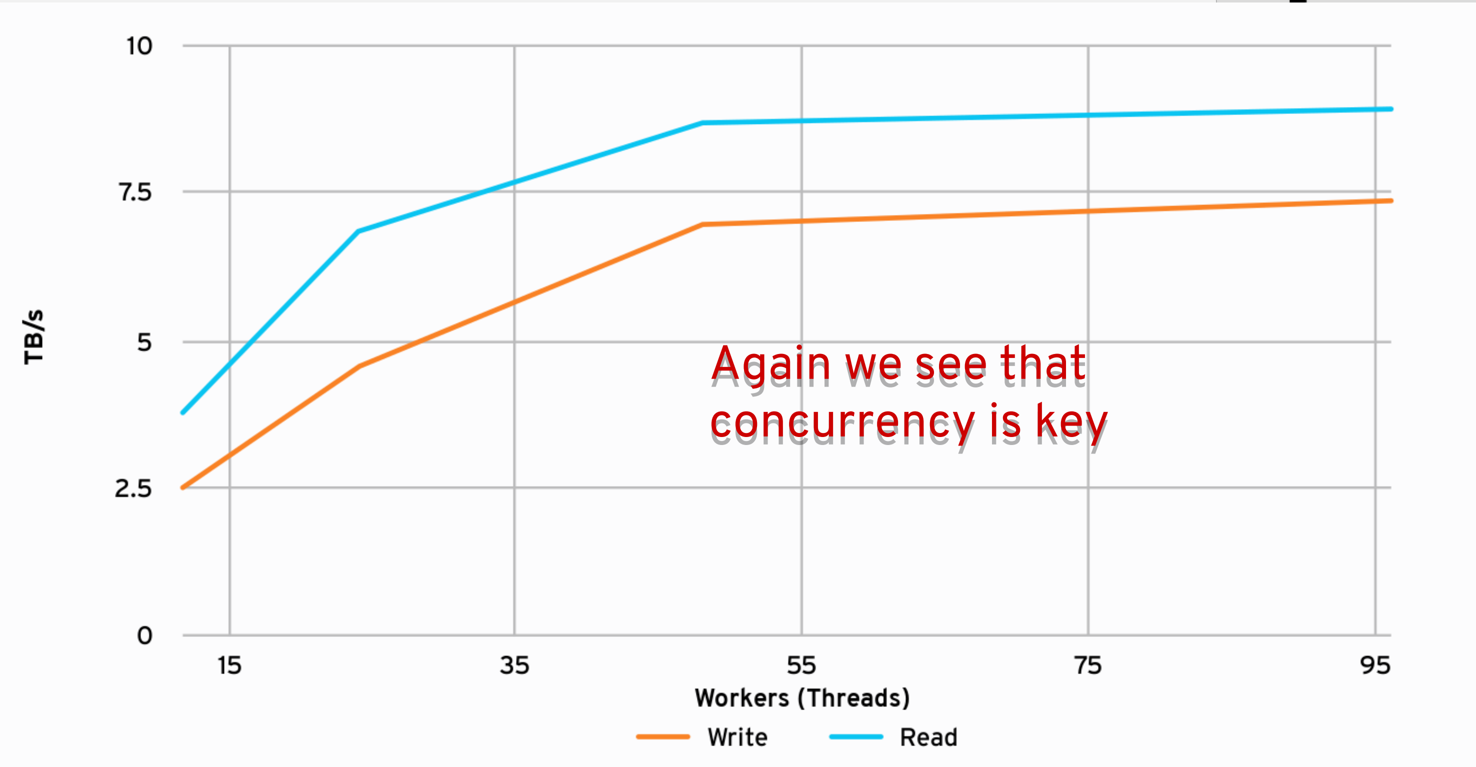
Note that the optimal configuration is different from the small file results

LARGE FILE DVD WORKLOAD



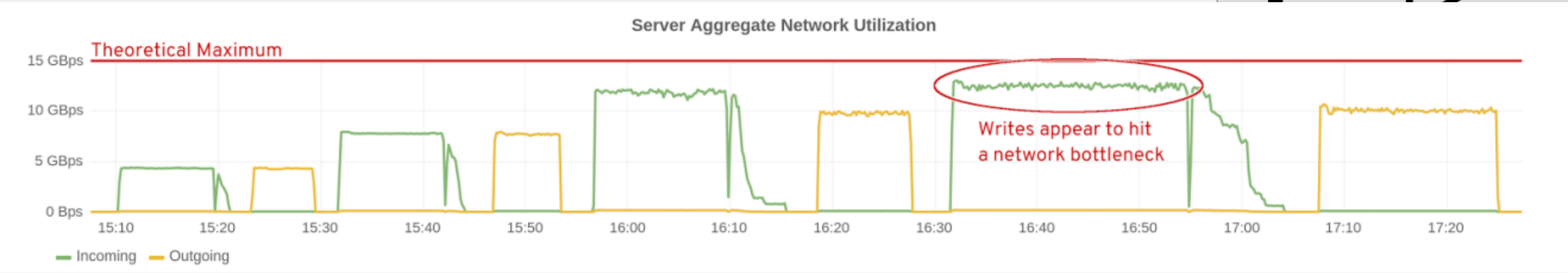
Node density has a larger impact on your throughput/$ efficiency with disperse volumes.

LARGE FILE DVD WORKLOAD

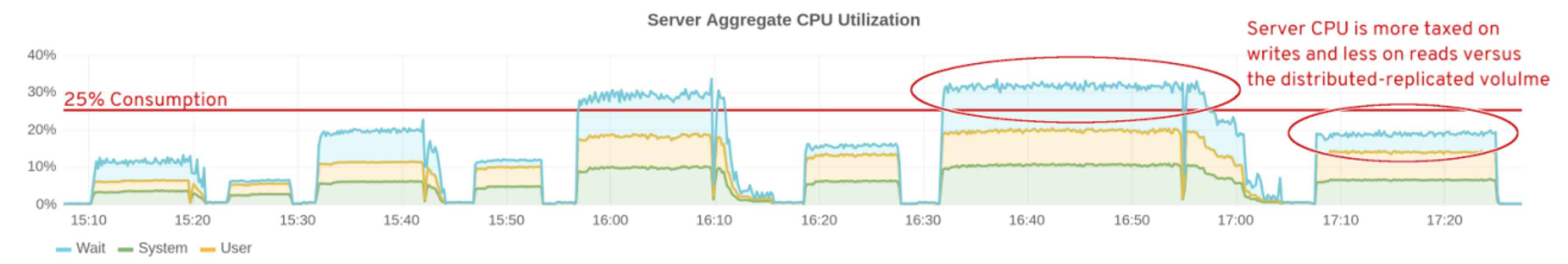


Again we see that concurrency is key

LARGE FILE DVD WORKLOAD

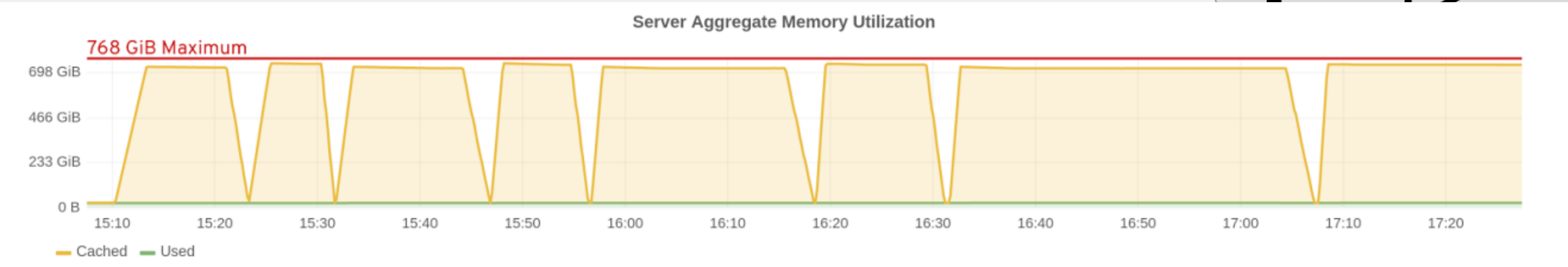


Writes appear to hit a network bottleneck

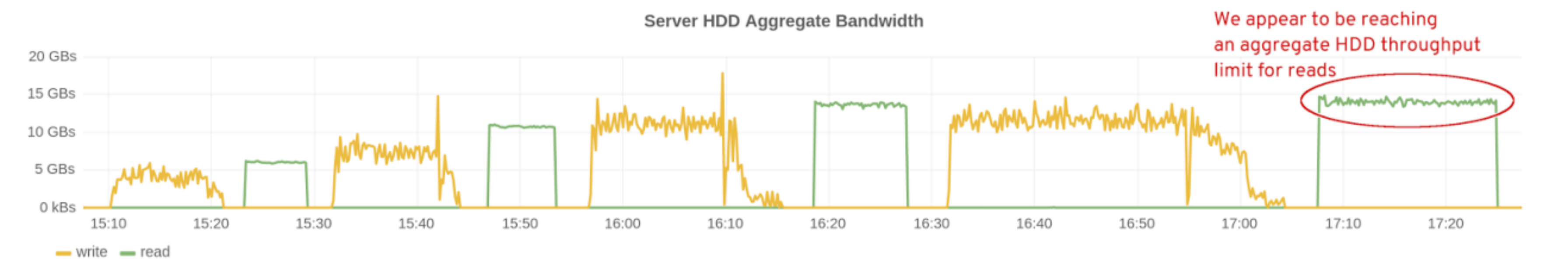


Server CPU is more taxed on writes and less on reads versus the distributed-replicated volulme

LARGE FILE DVD WORKLOAD



768 GiB Maximum



We appear to be reaching an aggregate HDD throughput limit for reads

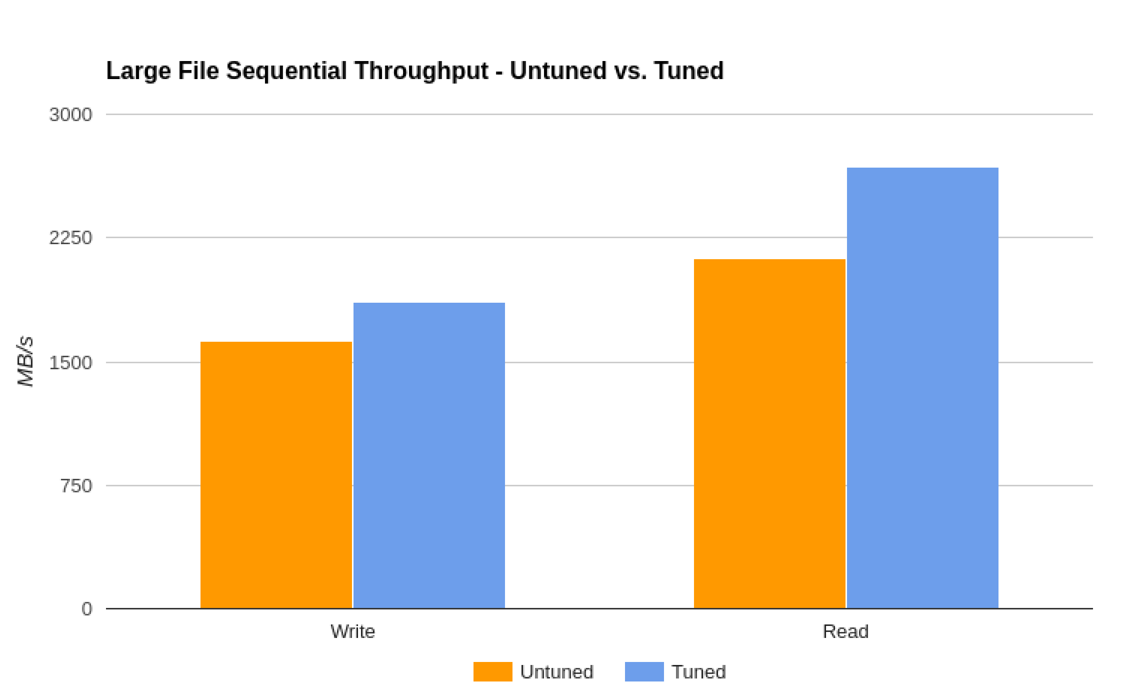
TUNING FOR LARGE FILE SEQUENTIAL

How Dustin got his performance gains from tuning!

- RAID 6 or EC are recommended for bricks
- Tuned profile: rhs-high-throughput
  - Read-ahead on bricks
  - Deadline scheduler
  - vm.dirty-ratio
- Jumbo Frames
- Event Threads = 4
- Smallfile tuning may have some effect, especially with metadata operations.

LARGE FILE SEQUENTIAL

4 Servers, 4 Clients, 4 Workers/Client, 16GB File/Worker



SCOPING FOR LARGE FILE WORKLOADS

Now that you understand the workload, how can you size your cluster?

Formula for *guesstimating* large file performance:

Writes = (Slowest of NIC / DISK) / # replicas * .7(overhead)

1200 MB / 2 * .7 = 420 MB / sec

Reads = (Slowest of NIC / Disk) * .6(overhead)

1200 * .6 = 720 MB / sec

This is just a rule of thumb, actual results are highly dependant on hardware.

TAKEAWAYS FOR LARGE FILE WORKLOADS

EC on JBOD outperforms replica 2 on RAID 6 high worker concurrency workloads

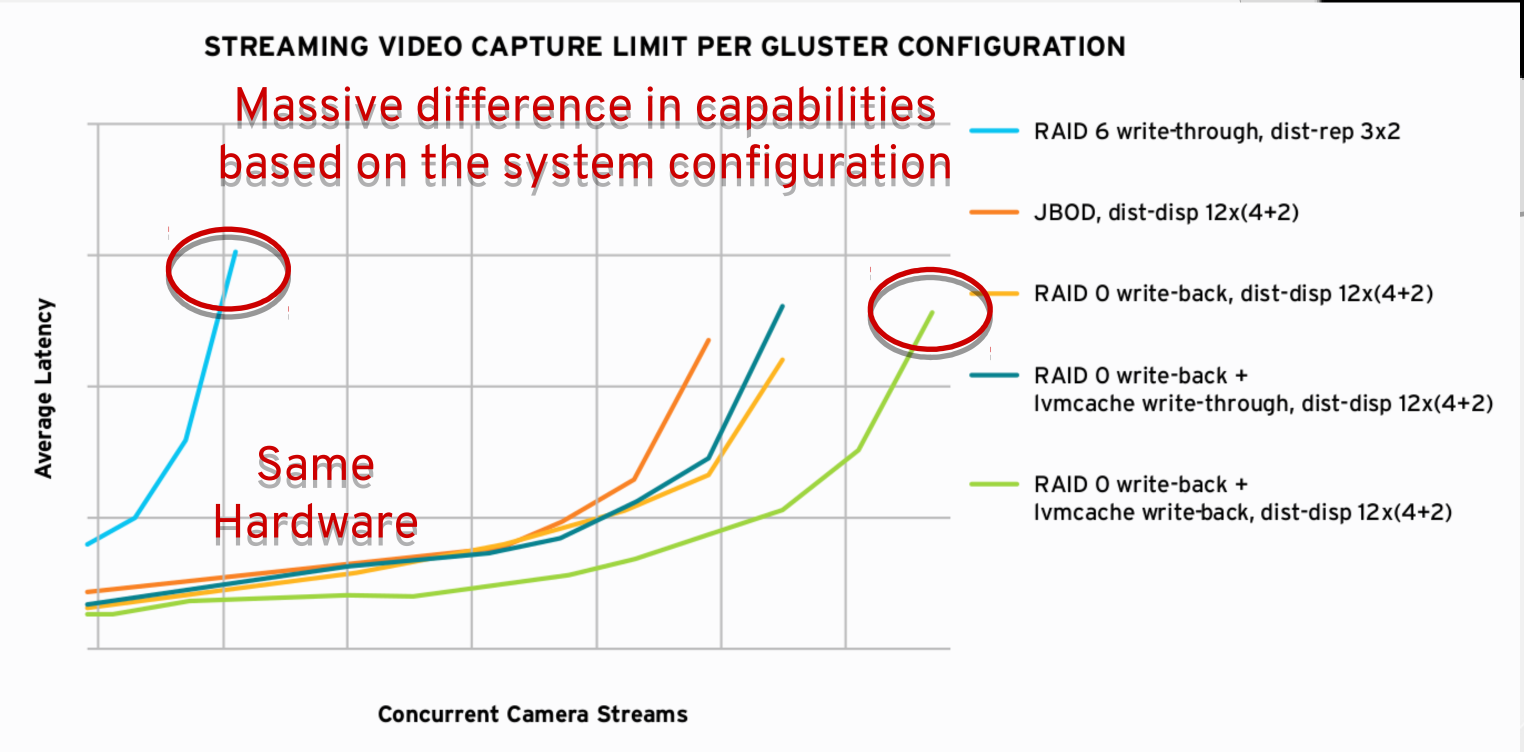
Replica 2 on RAID 6 outperforms EC on JBOD when there are less files / clients / threads and on single threaded workloads

Read ahead on block devices as well as jumbo frames provide the most performance benefit of the tunables

Again, start with the workload when designing your storage cluster. The proper brick architecture from the start will yield far better performance than any of the tunables mentioned. Design in a way that avoids problems, don’t try tune your way out of them.

YOUR WORKLOAD CAN'T BE SLOW IF YOU NEVER RUN IT

CCTV STREAMING WORKLOAD



HYPERCONVERGED RHV / RHGS

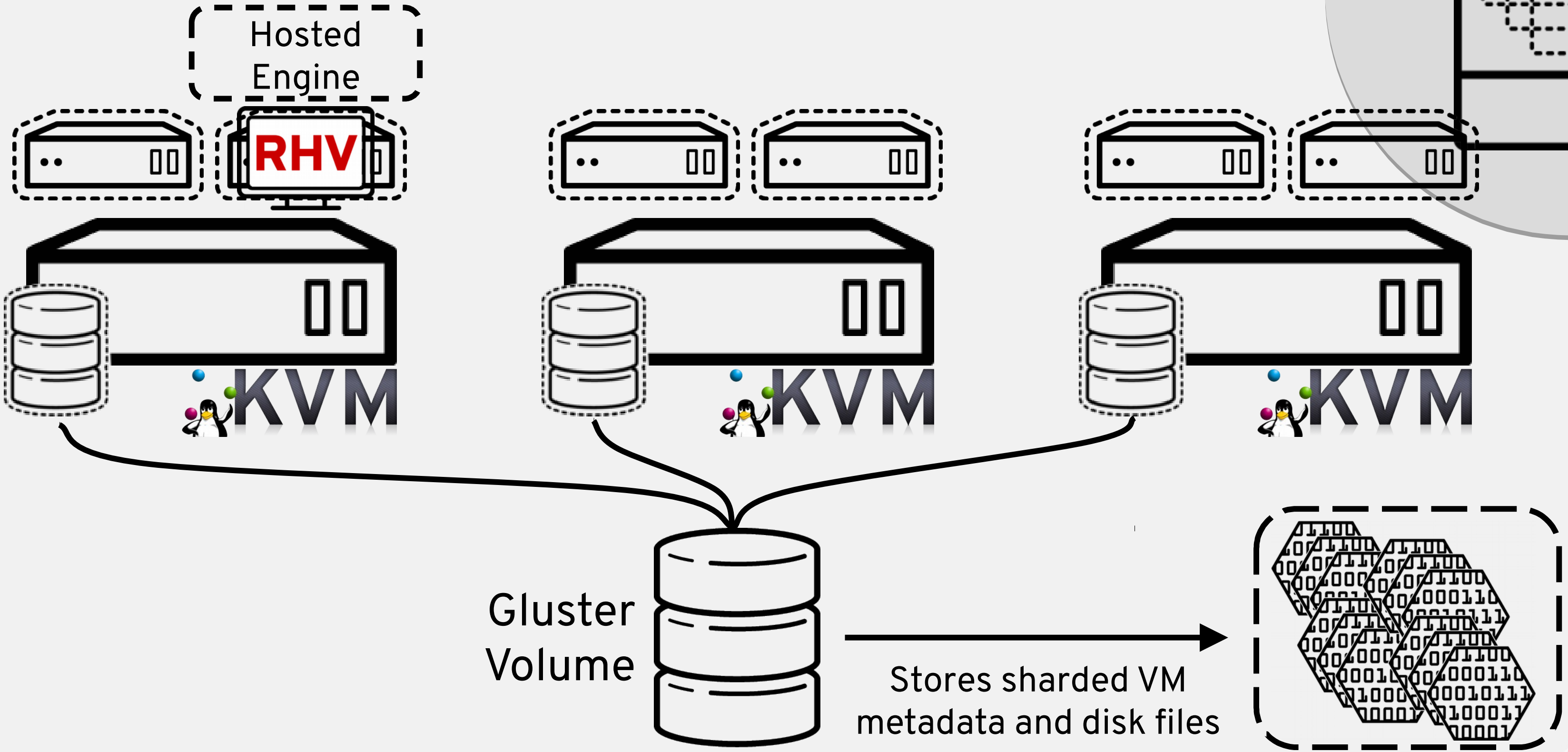
Setup Details

- Storage and compute on the same systems
- Cost advantage
- Management using the same linux based tools

```
# gdeploy -c robo.conf

# hosted-engine --deploy --config-append=<path to hosted engine answer
file>
```

Hyperconverged Infrastructure Example Arch

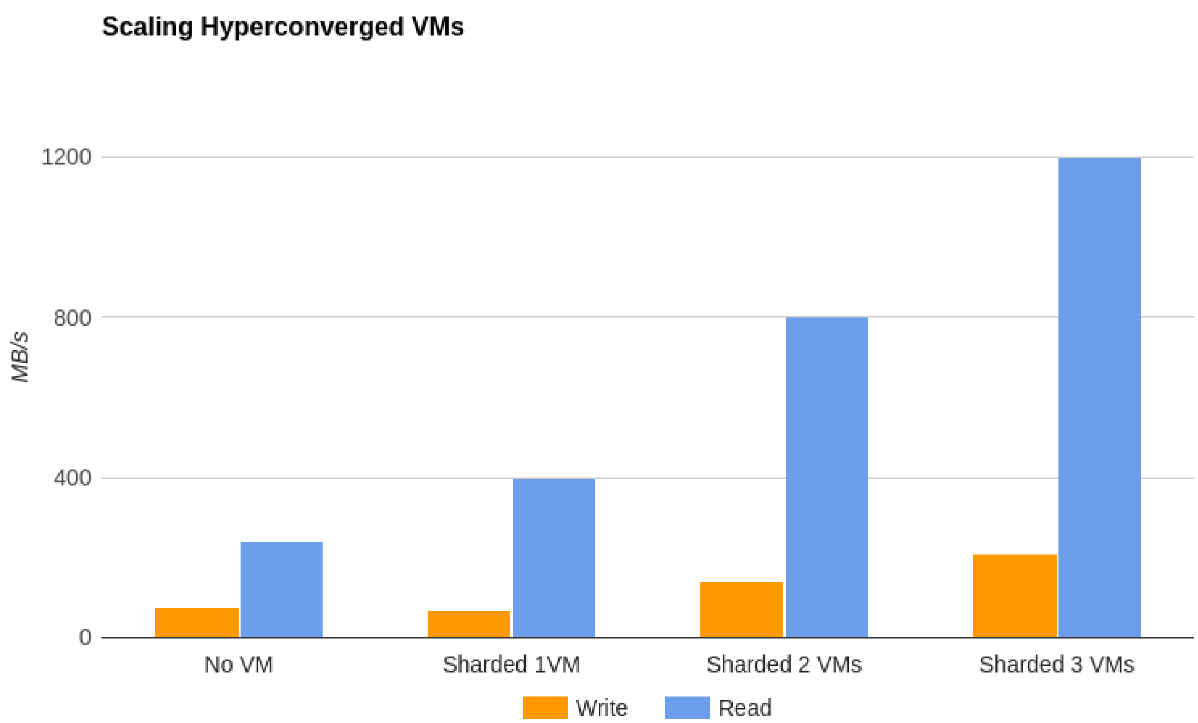


Hosted Engine

Gluster Volume

Stores sharded VM metadata and disk files

VM PERFORMANCE



PERFORMANCE TEST TOOL - GBENCH

Gbench was used to gather the performance data

- https://github.com/gluster/gbench
- Wraps IOZone, smallfile, FIO
- Run multiple iterations and averages it
- Multi host capable

RED HAT SUMMIT

THANK YOU

- plus.google.com/+RedHat
- facebook.com/redhatinc
- linkedin.com/company/red-hat
- twitter.com/RedHatNews
- youtube.com/user/RedHatVideos

Dustin Black
@dustinlblack

Ben Turner
@bennyturns

Install The Summit Mobile App

https://install.events/reraa17

PLEASE

Help us out by reviewing our session