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.
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.

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

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

MEDICAL INDUSTRY
Medical imaging needs are vast, and regulatory requirements can be demanding.
The Data Storage Shortfall

- Data stores are growing exponentially, while IT budgets are not.
- HDDs are becoming more dense, but $/GB decline is slowing.
- Software and hardware advances are needed to close the gap.
## The Datacenter Is Changing

<table>
<thead>
<tr>
<th>Development Model</th>
<th>Application Architecture</th>
<th>Deployment and Packaging</th>
<th>Application Infrastructure</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterfall</td>
<td>Monolithic</td>
<td>Bare Metal</td>
<td>Data Center</td>
<td>Scale Up</td>
</tr>
<tr>
<td>Agile</td>
<td>N-Tier</td>
<td>Virtual Services</td>
<td>Hosted</td>
<td>Scale Out</td>
</tr>
<tr>
<td>DevOps</td>
<td>Microservices</td>
<td>Containers</td>
<td>Hybrid Cloud</td>
<td>Software-defined Storage</td>
</tr>
</tbody>
</table>

#redhat #rhsummit
PERFORMANCE THAT SCALES

Performance should scale up as capacity does

Software-defined storage intelligently uses hardware to provide performance at very large scale.

• Traditional appliances perform better when they are empty than they do when they are full of disks

• Performance in software-defined storage clusters improves as clusters get larger, not the other way around

• Intel, SanDisk, Fujitsu, and Mellanox regularly contribute performance optimizations
VIRTUALIZED STORAGE SCALES BETTER

**Compute Node** + **Storage Appliance**
THE ROBUSTNESS OF SOFTWARE

Software can do things hardware can’t

Storage services based on software are more flexible than hardware-based implementations

• Can be deployed on bare metal, inside containers, inside VMs, or in the public cloud

• Can deploy on a single server, or thousands, and can be upgraded and reconfigured on the fly

• Grows and shrinks programmatically to meet changing demands
DIFFERENT KINDS OF STORAGE

**BLOCK STORAGE**
Data as sequential uniform blocks

**FILE STORAGE**
Data as buckets of hierarchical folders and files

**OBJECT STORAGE**
Data as a predictably mapped, loosely structured cluster of objects
HOW STORAGE FITS

RED HAT STORAGE

PHYSICAL
- RED HAT CEPH STORAGE
- RED HAT GLUSTER STORAGE
- RED HAT ENTERPRISE LINUX

VIRTUAL
- RED HAT CEPH STORAGE
- RED HAT GLUSTER STORAGE
- RED HAT ENTERPRISE LINUX
- RED HAT ENTERPRISE VIRTUALIZATION

PRIVATE CLOUD
- RED HAT CEPH STORAGE
- RED HAT GLUSTER STORAGE
- RED HAT ENTERPRISE LINUX
- RED HAT OPENSTACK PLATFORM

CONTAINERS
- RED HAT CEPH STORAGE
- RED HAT GLUSTER STORAGE
- RED HAT ENTERPRISE LINUX
- OPENSHIFT ENTERPRISE by Red Hat

PUBLIC CLOUD
- RED HAT CEPH STORAGE
- RED HAT GLUSTER STORAGE
- RED HAT ENTERPRISE LINUX

#redhat #rhsummit
RED HAT GLUSTER STORAGE

Open source, software-defined storage for unstructured file data at petabyte scale

- Media, video
- Machine, Log Data
- GeoSpatial
- Persistent Storage
- Documents
COMPARING THROUGHPUT AND COSTS AT SCALE

 STORAGE PERFORMANCE SCALABILITY

 NUMBER OF STORAGE NODES

 READS/Writes THROUGHPUT (MBps)

 Traditional Enterprise NAS Storage

 Software Defined Scale-out Storage (GlusterFS)

 STORAGE COSTS SCALABILITY

 NUMBER OF STORAGE NODES

 TOTAL STORAGE COSTS ($)

 Traditional Enterprise NAS Storage

 Software Defined Scale-out Storage (GlusterFS)
WHAT IS A SYSTEM?
Can be physical, virtual or cloud

PHYSICAL

VIRTUAL

CLOUD
Bricks taken from multiple hosts become one addressable unit

- High availability as needed
- Load balanced data
- Managed by Gluster
MULTI-PROTOCOL ACCESS

Primarily accessed as scale-out file storage with optional Swift obj APIs
WHY PERSISTENT STORAGE FOR CONTAINERS?

“For which workloads or application use cases have you used/do you anticipate to use containers?”

Data Apps: 77%
Cloud Apps: 71%
Systems of Engagement: 62%
Systems of Record: 62%
Web and Commerce Software: 57%
Mobile Apps: 52%
Social Apps: 46%

Scalable, Cost Effective, Distributed Storage for Containers

Base: 194 IT operations and development decision-makers at enterprise in APAC, EMEA, and North America
Source: A commissioned study conducted by Forrester Consulting on behalf of Red Hat, January 2015
GOT IT?
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
- Optimized 72x 7.2K HDD: 2 DVDs per second
- 72x SSD: 4 DVDs per second
OR…

High-Def CCTV Camera Recording Streams

200 CCTV streams within latency threshold

Optimized

500 CCTV streams within latency threshold

or

72x 7.2K HDD

72x SSD

or

72x 7.2K HDD

or

72x SSD
KEEP IT SIMPLE, STUPID
START WITH THE WORKLOAD, DUMMY
WHY DO YOU ASK THE WRONG QUESTIONS?
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....
plans to **add physical nodes to increase "performance"**

(currently [text 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. ??
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 ----------

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

Standard servers - 32KB file throughput by architecture

Design choice dramatically affects system capabilities

Same Hardware

Native / 12-HDD RAID 6 dist·rep 3x2
Native / 12-HDD RAID 6 disp 1x(4+2)
Native / 12-HDD JBOD dist·disp 12x(4+2)
NFS / 12-HDD RAID 6 dist·rep 3x2
NFS / 12-HDD RAID 6 disp 1x(4+2)
NFS / 12-HDD JBOD dist·disp 12x(4+2)

0 50 100 150 200

Files/s/drive

Read
Write

#redhat #rhsummit
Design choice as well has a large impact on the efficiency of your $$
Client concurrency is important for maximizing system throughput.

**SMALL FILE JPEG WORKLOAD**
SMALL FILE JPEG WORKLOAD

**Server Aggregate Network Utilization**
- 1.3 Gbps (10% of Theoretical Maximum)

**Server Aggregate CPU Utilization**
- 25% Consumption
SMALL FILE JPEG WORKLOAD

Server Aggregate Memory Utilization

768 GiB Maximum

Server HDD Busy

We are reaching a disk bottleneck on reads

#redhat #rhsummit
IF A FILE IS VERY VERY SMALL

IS IT STILL A FILE?
IT'S LIKE A SERIES OF TUBES
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,
      ...}
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

Create

<table>
<thead>
<tr>
<th></th>
<th>32KB files/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untuned</td>
<td>2000</td>
</tr>
<tr>
<td>Tuned w/ cache cold</td>
<td>3200</td>
</tr>
<tr>
<td>Tuned w/ cache hot</td>
<td>3500</td>
</tr>
</tbody>
</table>

Read

<table>
<thead>
<tr>
<th></th>
<th>32KB files/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untuned</td>
<td>5000</td>
</tr>
<tr>
<td>Tuned w/ cache cold</td>
<td>15000</td>
</tr>
<tr>
<td>Tuned w/ cache hot</td>
<td>20000</td>
</tr>
</tbody>
</table>
SMALLFILE METADATA WORKLOAD

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

![Graph 1: 320k files `ls -lR` (single client thread)]

- Untuned
- Tuned w/ cache cold
- Tuned w/ cache hot

- Smaller = better

![Graph 2: 320k files `ls -lR` (4 clients, 8 threads/client)]

- Untuned
- Tuned w/ cache cold
- Tuned w/ cache hot

- Smaller = better
LARGE FILE DVD WORKLOAD

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

Server Aggregate Network Utilization

- Theoretical Maximum: 15 Gbps
- Writes appear to hit a network bottleneck

Server Aggregate CPU Utilization

- 25% Consumption
- Server CPU is more taxed on writes and less on reads versus the distributed-replicated volume
LARGE FILE DVD WORKLOAD

Server Aggregate Memory Utilization

Server HDD Aggregate Bandwidth

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 guessing large file performance:

- **Writes** = (Slowest of NIC / DISK) / # replicas * .7(overhead)
  
  \[ \frac{1200 \text{ MB}}{2} \times .7 = 420 \text{ MB/sec} \]

- **Reads** = (Slowest of NIC / Disk) * .6(overhead)
  
  \[ 1200 \times .6 = 720 \text{ MB/sec} \]

*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

Massive difference in capabilities based on the system configuration

STREAMING VIDEO CAPTURE LIMIT PER CLUSTER CONFIGURATION

- RAID 6 write-through, dist-rep 3x2
- JBOD, dist-disp 12x(4+2)
- RAID 0 write-back, dist-disp 12x(4+2)
- RAID 0 write-back + lvocache write-through, dist-disp 12x(4+2)
- RAID 0 write-back + lvocache write-back, dist-disp 12x(4+2)

Average Latency

Concurrent Camera Streams

Same Hardware
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

Stores sharded VM metadata and disk files
VM PERFORMANCE

Scaling Hyperconverged VMs

<table>
<thead>
<tr>
<th></th>
<th>Write</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>No VM</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sharded 1VM</td>
<td>100</td>
<td>400</td>
</tr>
<tr>
<td>Sharded 2 VMs</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>Sharded 3 VMs</td>
<td>300</td>
<td>1200</td>
</tr>
</tbody>
</table>
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
THANK YOU

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

Dustin Black
@dustinlblack

Ben Turner
@bennyturns

#redhat #rhsummit
Install The Summit Mobile App

https://install.events/reraa17

PLEASE
Help us out by reviewing our session

#redhat #rhsummit