Storage Performance Benchmarking: Part 3 – Block Components

Ken Cantrell, NetApp
Mark Rogov, EMC
David Fair, SNIA ESF Chair, Intel

March 8, 2016
SNIA Legal Notice

• The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
• Member companies and individual members may use this material in presentations and literature under the following conditions:
  • Any slide or slides used must be reproduced in their entirety without modification
  • The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
• This presentation is a project of the SNIA Education Committee.
• Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
• The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
About The Speakers

Ken Cantrell
NetApp
Manager Perf Engineering
@kencantrelljr

Mark Rogov
EMC
Advisory Systems Engineer
@rogovmark

Dr. David Fair
SNIA ESF Chair & Intel
Ethernet Networking Marketing Manager
Storage Performance Benchmarking

METRICS AND TERMINOLOGY

SOLUTION UNDER TEST

BLOCK COMPONENTS

FILE COMPONENTS

WORKLOAD DEFINITIONS

JULY 30, 2015

OCT 20, 2015

TODAY

FUTURE WEBCASTS

© 2015 Storage Networking Industry Association. All Rights Reserved.
Session 1 – Terminology and Context

**OPS**
Count every protocol operation per second

**IOPS**
Count every IO operation per second

**RESPONSE TIME**
Time target takes to reply to an IO

**MB/S**
Payload sum of every operation per second

**OPS COUNT EVERY PROTOCOL OPERATION PER SECOND**

**IOPS COUNT EVERY IO OPERATION PER SECOND**

**RESPONSE TIME TIME TARGET TAKES TO REPLY TO AN IO**

**MB/S PAYLOAD SUM OF EVERY OPERATION PER SECOND**

**TERMINOLOGY**

**CONTEXT MAKES METRICS MATTER**

**GRAPH FUN**

**CONCEPT**

- System 1
- System 2

**RESPONSE TIME**
Time target takes to reply to an IO

**MB/S**
Payload sum of every operation per second

**OPS**
Count every protocol operation per second

**IOPS**
Count every IO operation per second

**CONTEXT MAKES METRICS MATTER**

**GRAPH FUN**
Session 2 – The Slowest Component Matters Most

SLOW COMPONENT MATTERS MOST

BOTTLENECKS ALWAYS EXIST

3 PERFORMANCE PRINCIPLES

DISK BOUND

CLIENT BOUND

PHYSICAL STORAGE
BACK-END CONNECT
STORAGE CONTROLLER
FRONT-END CONNECT
CLIENTS / HOSTS

Do Less Work
Do Work Faster
Increase Parallelism

© 2015 Storage Networking Industry Association. All Rights Reserved.
Enterprise Storage Capacity Shipped In 3Q’15

33.1 EXABYTES

WORLD POPULATION

7.4 BILLION

SPLIT EQUALLY

4.5 GIGABYTES PER INDIVIDUAL

OR

METRICS AND TERMINOLOGY

1541 COPIES OF OUR FIRST WEBCAST PowerPoint

© 2015 Storage Networking Industry Association. All Rights Reserved.
Eventually, All Data Goes To Block Storage

BLOCK STORAGE

SOLUTIONS UNDER TEST

WORKLOADS
<table>
<thead>
<tr>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTRO</strong></td>
</tr>
<tr>
<td><strong>R/W</strong></td>
</tr>
<tr>
<td><strong>TECH</strong></td>
</tr>
<tr>
<td><strong>RAID</strong></td>
</tr>
<tr>
<td><strong>FUN</strong></td>
</tr>
<tr>
<td><strong>END</strong></td>
</tr>
</tbody>
</table>
Let's Take A Drive… And Test It!

**RANDOM 4KiB I/O PERFORMANCE COMPARISON**

<table>
<thead>
<tr>
<th>IOPS</th>
<th>READS</th>
<th>WRITES</th>
</tr>
</thead>
<tbody>
<tr>
<td>514</td>
<td></td>
<td>133</td>
</tr>
</tbody>
</table>

- **READS**: 514
- **WRITES**: 133

**SAS HDD (146GB, 15K RPM)**
Detour! What Does “Random” Mean?

Imagine that the keyboard is a disk drive.

A QUICK BROWN FOX JUMPED OVER A LAZY DOG

Keys are all over
What Does “Sequential” Mean?

Imagine that the keyboard is a disk drive.

Every key is next to previous.

1 2 3 4 5 6
“Sequential Read” Example
Let’s Take A Drive… And Test It!

**Random 4KiB I/O Performance Comparison**

<table>
<thead>
<tr>
<th></th>
<th>Reads</th>
<th>Writes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOPS</td>
<td>514</td>
<td>133</td>
</tr>
</tbody>
</table>

**SAS HDD (146GB, 15K RPM)**
Let’s Take Two Drives… And Test Them!

**RANDOM 4KiB I/O PERFORMANCE COMPARISON**

<table>
<thead>
<tr>
<th></th>
<th>Reads</th>
<th>Writes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS HDD (146GB, 15K RPM)</td>
<td>514</td>
<td>133</td>
</tr>
<tr>
<td>CONSUMER SSD (400GB, SATA 6Gb/s, MLC)</td>
<td>61,178</td>
<td>34,406</td>
</tr>
</tbody>
</table>

© 2015 Storage Networking Industry Association. All Rights Reserved.
And Add More SSDs

**RANDOM 4KiB I/O PERFORMANCE COMPARISON**

- **SAS HDD** (146GB, 15K RPM)
- **Consumer SSD** (400GB, SATA 6Gb/s, MLC)
- **Consumer SSD** (128GB, M.2x4 AHCI, MLC)
- **Enterprise SSD** (1400GB, PCIe x8 AHCI, MLC)
- **Enterprise SSD** (1600GB, U.2 NVMe, MLC)

<table>
<thead>
<tr>
<th>Type</th>
<th>Reads IOPS</th>
<th>Writes IOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDD</td>
<td>514</td>
<td>133</td>
</tr>
<tr>
<td>SAS HDD</td>
<td>172,881</td>
<td>34,406</td>
</tr>
<tr>
<td>Consumer SSD</td>
<td>61,178</td>
<td>6,307</td>
</tr>
<tr>
<td>Consumer SSD</td>
<td>113,260</td>
<td>83,021</td>
</tr>
<tr>
<td>Enterprise SSD</td>
<td>750,846</td>
<td>113,260</td>
</tr>
<tr>
<td>Enterprise SSD</td>
<td>585,884</td>
<td>83,021</td>
</tr>
</tbody>
</table>

© 2015 Storage Networking Industry Association. All Rights Reserved.
Agenda

INTRO

INTRODUCTION

R/W

READING, WRITING; WHAT IS THE DIFFERENCE?

TECH

HOW DOES THIS TECH WORK ANYWAY?

RAID

WHAT IF YOU NEED MORE THAN ONE?

FUN

PERFORMANCE?

END

SUMMARY
How Does This Tech Work?

**FLASH**

**HDD OR DISK DRIVE**
Spinning Drives And Sectors

SEEK THE TRACK
SPIN TO THE SECTOR

READ
WRITE
Flash And NAND Gates

**Every NAND can be set to 0 individually**

WRITE 0

WRITE INPUT

1 1 0 1

COMMON INPUT

ERASE INPUT

**To set back to 1, an entire group needs to be reset**

WRITE 1

WRITE INPUT

1 1 1 1 1

COMMON INPUT

ERASE INPUT

© 2015 Storage Networking Industry Association. All Rights Reserved.
Flash Construction

**Flash Block**

- Pages: 2KiB, 4KiB, 8KiB, ...

**Pages per Block (dep on model)**
- 128, 256, ..., etc.

**Most Flash**

Write—**1 page** at a time

**Flash Device**

- **Logical to Physical Redirection Map**
  - Clean, Dirty, Data

- **Redirect on Over-Write**
  - An IO is redirected to a **clean block/page**
  - Leaving old block/page **dirty**

---

© 2015 Storage Networking Industry Association. All Rights Reserved.
Garbage Collection

**Flash Device**

- Logical to Physical Redirection Map:
  - 0x00: Clean Block
  - 0x01: Dirty Block
  - 0x02: Clean Block
  - 0xAA: Dirty Block

- Number of Blocks Defines Capacity

**Garbage Collection**

- Erase—1 Dirty Block at a Time
  - (When number of Clean Blocks is low)
Sequential Vs. Random

SSD or Flash

EVERYTHING IS RANDOM IO FOR FLASH

WRITE | READ
---|---
ERASE + WRITE | READ

HDD or Disk Drive

WRITE | READ
---|---
SEQUENTIAL | READ
SEEK/SPIN + WRITE | SEEK/SPIN + READ

SLOWER PERFORMANCE
Agenda

INTRO | INTRODUCTION
R/W | READING, WRITING; WHAT IS THE DIFFERENCE?
TECH | HOW DOES THIS TECH WORK ANYWAY?
RAID | WHAT IF YOU NEED MORE THAN ONE?
FUN | PERFORMANCE?
END | SUMMARY
Just One?

CAPACITY

PROTECTION

PERFORMANCE
RAID—Redundant Array Of Inexpensive Disks

- PHYSICAL STORAGE
- BACK-END CONNECT
- STORAGE CONTROLLER
- FRONT-END CONNECT
- CLIENTS / HOSTS

© 2015 Storage Networking Industry Association. All Rights Reserved.
RAID-0 (Striping Without Parity)

**PHYSICAL STORAGE**
- 01
- 11

**VIRTUAL/LOGICAL**
- 01
- 11

**CLIENTS / HOSTS**

**BACK END**

**FRONT END**

**CAPACITY**
- 100%

**PROTECTION**
- NONE

**PERFORMANCE**
- 100%
RAID-1 (Mirroring)

**PHYSICAL STORAGE**
- 01
- 11

**VIRTUAL/LOGICAL**
- 01
- 11

**CLIENTS / HOSTS**
- Front End

**BACK END**
- 01
- 11

**FRONT END**
- 01
- 11

**CAPACITY**
- 50%
- 1 Drive
- 100%
- 50%

© 2015 Storage Networking Industry Association. All Rights Reserved.
RAID-3, -4, -5 [-6, -DP]*

Striping With Parity*

* RAID-6/-DP requires more than one parity
RAID Partial Writes

All Single Parity RAID: RAID-3, -4, -5, and etc.

**Single Partial Write:**
- **Read** Old Data
- **Read** Old Parity
- **Calculate** New Parity
- **Write** New Data
- **Write** New Parity

2 Reads 2 Writes

100 IOs 70R/30W = 70 Read + 30 Write IOs

\[
\text{Backend} = (70R + 30 \times (2W + 2R)) = 190 \text{ IOs}
\]

RAID Penalty

© 2015 Storage Networking Industry Association. All Rights Reserved.
RAID Implementation

PHYSICAL STORAGE

BACK-END CONNECT

STORAGE CONTROLLER

FRONT-END CONNECT

CLIENTS / HOSTS

© 2015 Storage Networking Industry Association. All Rights Reserved.
Erasure Coding Implementation
Erasure Coding

\[ N + M = 2 + 1 \]

Physical Storage:

- N is number of data blocks
- M is number of protection blocks

Diagram:

- Front End
- Back End

Diagram elements correspond to the equations and concepts described.
Agenda

INTRO

INTRODUCTION

R/W

READING, WRITING; WHAT IS THE DIFFERENCE?

TECH

HOW DOES THIS TECH WORK ANYWAY?

RAID

WHAT IF YOU NEED MORE THAN ONE?

FUN

PERFORMANCE?

END

SUMMARY
What “Really” Happens With RAID-5?

HDD Potential Aggregate 4KiB Random Write Performance (As Seen at Client)

IOPS

106
106
106
106

Each Drive

© 2015 Storage Networking Industry Association. All Rights Reserved.
What “Really” Happens With RAID-5?

HDD Potential Aggregate 4KiB Random Write Performance (as seen at client)

Each drive:
- Disk 1: 106
- Disk 2: 106
- Disk 3: 106
- Disk 4: 106
- Put into 3+1R5: 169

© 2015 Storage Networking Industry Association. All Rights Reserved.
What “Really” Happens With RAID-5?

HDD Potential Aggregate 4KiB Random Write Performance (As Seen at Client)

- Each Drive: 106
- Put into 3+1R5: 169
- Caching RAID Controller: 149
What “Really” Happens With RAID-5?

**FLASH POTENTIAL AGGREGATE 4KiB RANDOM WRITE PERFORMANCE (AS SEEN AT CLIENT)**

<table>
<thead>
<tr>
<th>HDD REMINDER</th>
<th>INDIVIDUAL DEVICES</th>
<th>3+1 RAID-5</th>
<th>CACHING RAID CONTROLLER</th>
</tr>
</thead>
<tbody>
<tr>
<td>149</td>
<td>169</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2015 Storage Networking Industry Association. All Rights Reserved.
What “Really” Happens With RAID-5?

**Flash Potential Aggregate 4KiB Random Write Performance (As Seen at Client)**

<table>
<thead>
<tr>
<th>HDD Reminder</th>
<th>Individual Devices</th>
<th>3+1 RAID-5</th>
<th>Caching RAID Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DEVICE1</td>
<td>DEVICE2</td>
<td>DEVICE3</td>
</tr>
<tr>
<td>IOPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5045</td>
<td>5045</td>
<td>5045</td>
<td>318</td>
</tr>
</tbody>
</table>

© 2015 Storage Networking Industry Association. All Rights Reserved.
What “Really” Happens With RAID-5?

**Flash Potential Aggregate 4KiB Random Write Performance (As Seen at Client)**

<table>
<thead>
<tr>
<th>HDD Reminder</th>
<th>Individual Devices</th>
<th>3+1 RAID-5</th>
<th>Caching RAID Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td>318</td>
<td>5045</td>
<td>9736</td>
<td>9736</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5045</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5045</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5399</td>
<td></td>
</tr>
</tbody>
</table>

- Device1
- Device2
- Device3
- Device4
- 3+1R5
- Secret Sauce

© 2015 Storage Networking Industry Association. All Rights Reserved.
What “Really” Happens With RAID-5?

**Flash Potential Aggregate 4KiB Random Write**
(MiB/s, As Seen At Client)

<table>
<thead>
<tr>
<th>HDD Reminder</th>
<th>Individual Devices</th>
<th>3+1 RAID-5</th>
<th>Caching RAID Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>19.71</td>
<td>38</td>
<td>21.09</td>
</tr>
<tr>
<td></td>
<td>19.71</td>
<td>38.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.71</td>
<td>38.03</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>19.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>59</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>21.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.03</td>
<td></td>
</tr>
</tbody>
</table>

© 2015 Storage Networking Industry Association. All Rights Reserved.
What “Really” Happens With RAID-5?

**FLASH POTENTIAL AGGREGATE 128KiB SEQUENTIAL READS**
(MB/s, AS SEEN AT CLIENT)

- **INDIVIDUAL DEVICES**
  - Device 1: 1895 MB/s
  - Device 2: 1895 MB/s
  - Device 3: 1895 MB/s
  - Device 4: 1895 MB/s

- **3+1 RAID-5**
  - 5685 MB/s
Flash In The Real World

**PHYSICAL STORAGE**
5685 MB/s

**BACK-END CONNECT**
REQUIRES 10x 6Gb/s SAS

**STORAGE CONTROLLER**
REQUIRES 6x PCIe 3.0 LANES

**FRONT-END CONNECT**
REQUIRES 4x 16Gb FC

**CLIENTS / HOSTS**
How many hosts?

**MB/s**

10
6
4
??

© 2015 Storage Networking Industry Association. All Rights Reserved.
Flash In The Real World

NEW TECH

PHYSICAL STORAGE

BACK-END CONNECT

STORAGE CONTROLLER

FRONT-END CONNECT

CLIENTS / HOSTS

5685 MB/s

REQUIRES
10x 6Gb/s

REQUIRES
6x PCIe

REQUIRES
4x 16G

How many hosts?
Flash In The Real World

BLOCK IS THE FOUNDATION OF STORAGE PERFORMANCE
Agenda

1. **INTRO**
   - Introduction

2. **R/W**
   - Reading, Writing; What is the Difference?

3. **TECH**
   - How does this tech work anyway?

4. **RAID**
   - What if you need more than one?

5. **FUN**
   - Performance?

6. **END**
   - Summary
Storage Performance Benchmarking

METRICS AND TERMINOLOGY

FILE COMPONENTS

WORKLOAD DEFINITIONS

TODAY

JULY 30, 2015

OCT 20, 2015

FUTURE WEBCASTS

IOPS
MB/S
RESPONSE TIME

SOLUTION UNDER TEST

BLOCK COMPONENTS

FILE COMPONENTS

© 2015 Storage Networking Industry Association. All Rights Reserved.
After This Webcast

- A PDF and a PPT of the slides for this and all previous parts of this Webcast series will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
  - PPT and PDF: [http://www.snia.org/forums/esf/knowledge/webcasts](http://www.snia.org/forums/esf/knowledge/webcasts)
  - Presentation Recording: [https://www.brighttalk.com/webcast/663/164323](https://www.brighttalk.com/webcast/663/164323)

- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
  - [http://sniaesfblog.org/](http://sniaesfblog.org/)

- Follow us on Twitter @SNIAESF, @RogovMark, @KenCantrellJr, @DrJMetz

- Next Webcast – Second Half of 2016
  - “Storage Performance Benchmarking: Part 4”
QUESTIONS?
THANK YOU!
Appendix – Additional Reading
Appendix – More Reading

- Benchmarking methods for randomly sampling from a file, and why random seeks can (usually) hurt performance: [http://simpsonlab.github.io/2015/05/19/io-performance/](http://simpsonlab.github.io/2015/05/19/io-performance/)
- Excellent hard drive overview: [https://www.backblaze.com/hard-drive.html](https://www.backblaze.com/hard-drive.html)
- SSD M.2 Interface: [http://arstechnica.com/gadgets/2015/02/understanding-m-2-the-interface-that-will-speed-up-your-next-ssd/](http://arstechnica.com/gadgets/2015/02/understanding-m-2-the-interface-that-will-speed-up-your-next-ssd/)
- More complete SSD interface article, covering NVMe, U.2 and M.2: [http://blog.ocz.com/ssd-interfaces-sata-m2-u2-nvme/](http://blog.ocz.com/ssd-interfaces-sata-m2-u2-nvme/)
• RAID
  - [www.raid-recovery-guide.com/raid5-parity.aspx](http://www.raid-recovery-guide.com/raid5-parity.aspx)
  - [http://rickardnobel.se/how-raid5-works/](http://rickardnobel.se/how-raid5-works/)
  - [http://igoro.com/archive/how-raid-6-dual-parity-calculation-works/](http://igoro.com/archive/how-raid-6-dual-parity-calculation-works/)