Test for read/write speeds, array rebuild times, scrub times on 2-6 hard drives for various array types

How do the ZFS pool configurations compare to each other based on a number of disks

Papers read:

28/11/2023

**ZFS STORAGE POOL LAYOUT**

<https://static.ixsystems.co/uploads/2020/09/ZFS_Storage_Pool_Layout_White_Paper_2020_WEB.pdf>

Positives:

* Clearly explains the different types of pool configurations
* Shows formulas and examples for working out read/write speeds, iops, redundancy levels and capacity

Negatives:

* Only gives theoretical numbers, no real world figures, arguably unimportant and performance figures will depend on a variety of factors, although things like CPU are unlikely to affect things these days
* Does not include resilvering times and probability of loosing the pool

How it relates

* Basically what I want to do plus writing a script to give some real world performance metrics

21/11/2023

**A PERFORMANCE COMPARISON**

**OF ZFS AND BTRFS ON LINUX**

<https://www.diva-portal.org/smash/get/diva2:822493/FULLTEXT01.pdf>

Positves:

* Points out using warm cache
* Experiments done 20 times, discarding first time to reach steady-state
* Important to think about cache and memory

Negatives:

* Results consistently paint ZFS as the worst which is not comparable to many other results
* From 2015, more up to date than other papers but still out of date, would be interesting to see if the performance numbers match more recent results

**The Zettabyte Filesystem**

<https://www.cs.hmc.edu/~rhodes/cs134/readings/The%20Zettabyte%20File%20System.pdf>

Positives:

* A paper about ZFS from the creators highlighting what makes ZFS different and how it works

Negatives:

* Biased

**Benchmarking file system benchmarking: it \*IS\* rocket science**

24/11/2023

<https://www.researchgate.net/publication/262350224_Benchmarking_file_system_benchmarking_it_IS_rocket_science>

How it relates:

* Seems to show IOmeter would be best used to measure IOPS
* Can talk about how it points out that some benchmarks like Filebench are unable to separate between different levels of the file system, however it is what most others use so the results are more valuable when able to compare them with others
* Is it necessary to separate the parts out
* All parts of the filesystem are so intertwined that it is difficult to separate and test the individual parts separately
* Makes interesting points about caching, need to limit the amount of cache otherwise the files can fit entirely in memory, at which point the speeds comes from the memory system not the file system
* Interesting to test how the file system uses cache but hard to control

Existing benchmarks

**On the implementation of ZFS (Zettabyte File System) storage system**

25/11/2023

<https://www.researchgate.net/publication/315866344_On_the_implementation_of_ZFS_Zettabyte_File_System_storage_system>

Positives:

* Provides very granular results between different ZFS compression algorithms

Negatives:

* According to the benchmarking rocket science paper, bonnie++ is not able to isolate the on-disk and caching layers, i.e. the performance will improve as time goes on as files begin to be cached
* According to the open source github repository, LZ4 has had performance improvements since this papers publication <https://github.com/lz4/lz4/releases>

How it relates:

* Shows that, with the exception of very specific circumstances such as very low available CPU resources or entirely writing compressible files, LZ4 is almost always the best compression algorithm to use and so would not be worth repeating the tests with different compression algorithms

## \*\*\*[ZFS: Resilver Performance of Various RAID Schemas](https://louwrentius.com/zfs-resilver-performance-of-various-raid-schemas.html)

26/11/2023

<https://louwrentius.com/zfs-resilver-performance-of-various-raid-schemas.html>

Positive:

* Tests a wide variety of pool configurations and number of disks to find resilvering times
* Tests resilvering times at both 25% and 50%

Negatives:

* Does not test anything more than 50%
* Only tests data immediately after it has been written to disk and where it is not fragmented. In reality, by the time a disk failed and the array would need rebuilding, the data would be heavily fragmented which could lead to increased resilvering times
* Is not stated whether the test has been run multiple times to validate results
* RAID-Z tests 3, 4, 5 and 9 disks, RAID-Z2 tests 4, 6 and 10 and RAID-Z3 tests 5, 7 and 11. Would have been nice to see some consistency with the amount of disks being used

How it relates to my question:

* Provides some values to compare my results against
* My results will be more granular as I will test every possible configuration on 2-6 drives
* Will likely just fill the array to 10% as this research shows that as the array fills up, resilvering times mostly increase linearly with some extra as the disk tracks become closer to the center

**\*\*\*Workload Dependent Performance Evaluation of the Btrfs and ZFS Filesystems**

<https://picture.iczhiku.com/resource/paper/shIdkpUhfRUjHbcx.pdf>

Positives:

* Provides some relevant benchmarks
* Explains some of the architecture for ZFS
* Repeated benchmarks 10 times

Negatives:

* Published in 2009 so rather out of date

How it relates:

* Provides some relevant benchmarks to compare to
* Can talk about how it gave me some background on ZFS
* Mentions that zfs shouldn’t be used for situations where there aren’t a lot of cpu resources

**Performance Evaluation of ZFS and LVM (with ext4)**

**for Scalable Storage System**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5930130>

positives:

* Provides some benchmarks
* Possible the ZFS over sata results are relevant to me
* Demonstrates that ZFS gives better performance the more drives it has due to its dynamic striping, relevant to my research as I can talk about this and compare it to whether ZFS does give better “throughput”

Negatives:

* Only benchmarks 1-3 drives
* Emphasis is on comparing LVM to ZFS over the network so may not be completely relevant

**Ext4, XFS, BtrFS and ZFS Linux File Systems on**

**RADOS Block Devices (RBD):**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7881982>

Positives:

* Gives benchmarks
* Uses tools that others have used such as filebench so results can be comparable

Negaives:

* Paper focuses on using RADOS so may not strictly be relevant
* They apparently use RAID 0 for ZFS which is an odd choiuce

A Review on Performance Analysis of ZFS & BTRFS

[https://ieeexplore.ieee.org/---------------------------------------stamp/stamp.jsp?tp=&arnumber=8698103](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8698103)

more of a literature review, does not do any kind of analysis or experiment, just sums up other papers

Could test the performance of ZFS on one drive, is a little redundant though

2 Disks, Mirror, Read/Write, IOPS, resilver speed

2 Disks RAID-Z, Read/Write, IOPS, resilver speed

3 Disks, Mirror, Read/Write, IOPS, resilver speed

3 Disks, RAID-Z, Read/Write, IOPS, resilver speed

3 Disks, RAID-Z2, Read/Write, IOPS, resilver speed

4 Disks, Mirror, Read/Write, IOPS, resilver speed

4 Disks, RAID-Z, Read/Write, IOPS, resilver speed

4 Disks, RAID-Z2, Read/Write, IOPS, resilver speed

4 Disks, RAID-Z3, Read/Write, IOPS, resilver speed

5 Disks, Mirror, Read/Write, IOPS, resilver speed

5 Disks, RAID-Z, Read/Write, IOPS, resilver speed

5 Disks, RAID-Z2, Read/Write, IOPS, resilver speed

5 Disks, RAID-Z3, Read/Write, IOPS, resilver speed

6 Disks, Mirror, Read/Write, IOPS, resilver speed

6 Disks, RAID-Z, Read/Write, IOPS, resilver speed

6 Disks, RAID-Z2, Read/Write, IOPS, resilver speed

6 Disks, RAID-Z3, Read/Write, IOPS, resilver speed

Is it worth testing stripe? Would never be used in the real world as if one disk dies, all data is lost

What about a striped mirror?

Could do some benchmarks limiting memory and without

A black screen with a white line

Description automatically generated

videoserver Filebench Version 1.5-alpha3 0.000: Allocated 177MB of shared memory 0.004: Video Server Version 3.0 personality successfully loaded 0.564: Populating and pre-allocating filesets 0.564: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 655360.000MB total size 0.564: Removing passivevids tree (if exists) 0.572: Pre-allocating directories in passivevids tree 0.573: Pre-allocating files in passivevids tree 2250.276: activevids populated: 16 files, avg. dir. width = 4, avg. dir. depth = 2.0, 0 leafdirs, 163840.000MB total size 2250.276: Removing activevids tree (if exists) 2250.283: Pre-allocating directories in activevids tree 2250.784: Pre-allocating files in activevids tree 3573.183: Waiting for pre-allocation to finish (in case of a parallel pre-allocation) 3616.100: Population and pre-allocation of filesets completed 3616.118: Starting 1 vidreaders instances 3616.120: Starting 1 vidwriter instances 3617.158: Running... 4217.241: Run took 600 seconds... 4217.242: Per-Operation Breakdown serverlimit 228411ops 381ops/s 0.0mb/s 99.829ms/op [0.000ms - 5003.326ms] vidreader 228507ops 381ops/s 95.2mb/s 106.343ms/op [0.014ms - 2975.416ms] replaceinterval 1ops 0ops/s 0.0mb/s 10000.109ms/op [10000.109ms - 10000.109ms] wrtclose 2ops 0ops/s 0.0mb/s 0.015ms/op [0.015ms - 0.016ms] newvid 2ops 0ops/s 34.1mb/s 284946.183ms/op [280801.369ms - 289090.998ms] wrtopen 2ops 0ops/s 0.0mb/s 360.100ms/op [0.608ms - 719.592ms] vidremover 2ops 0ops/s 0.0mb/s 7726.783ms/op [464.074ms - 14989.493ms] 4217.242: IO Summary: 228515 ops 380.806 ops/s 381/0 rd/wr 129.3mb/s 108.904ms/op 4217.242: Shutting down processes pool: zfstest state: ONLINE config: NAME STATE READ WRITE CKSUM zfstest ONLINE 0 0 0 sdb ONLINE 0 0 0 sdc ONLINE 0 0 0 sdd ONLINE 0 0 0 sde ONLINE 0 0 0 errors: No known data errors videoserver Filebench Version 1.5-alpha3 0.001: Allocated 177MB of shared memory 0.005: Video Server Version 3.0 personality successfully loaded 0.233: Populating and pre-allocating filesets 0.233: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 655360.000MB total size 0.234: Removing passivevids tree (if exists) 0.242: Pre-allocating directories in passivevids tree 0.242: Pre-allocating files in passivevids tree 1939.260: activevids populated: 16 files, avg. dir. width = 4, avg. dir. depth = 2.0, 0 leafdirs, 163840.000MB total size 1939.260: Removing activevids tree (if exists) 1939.267: Pre-allocating directories in activevids tree 1939.269: Pre-allocating files in activevids tree 3129.085: Waiting for pre-allocation to finish (in case of a parallel pre-allocation) 3183.487: Population and pre-allocation of filesets completed 3183.507: Starting 1 vidreaders instances 3183.509: Starting 1 vidwriter instances 3184.542: Running... 3784.657: Run took 600 seconds... 3784.658: Per-Operation Breakdown serverlimit 229915ops 383ops/s 0.0mb/s 109.586ms/op [0.001ms - 5022.489ms] vidreader 230041ops 383ops/s 95.8mb/s 63.611ms/op [0.016ms - 1756.934ms] replaceinterval 3ops 0ops/s 0.0mb/s 10000.128ms/op [10000.094ms - 10000.151ms] wrtclose 3ops 0ops/s 0.0mb/s 0.022ms/op [0.013ms - 0.031ms] newvid 3ops 0ops/s 51.2mb/s 184168.004ms/op [137046.461ms - 220726.169ms] wrtopen 4ops 0ops/s 0.0mb/s 147.262ms/op [1.869ms - 392.540ms] vidremover 4ops 0ops/s 0.0mb/s 478.530ms/op [179.607ms - 589.618ms] 3784.658: IO Summary: 230055 ops 383.352 ops/s 383/0 rd/wr 147.0mb/s 66.020ms/op 3784.658: Shutting down processes pool: zfstest state: ONLINE config: NAME STATE READ WRITE CKSUM zfstest ONLINE 0 0 0 sdb ONLINE 0 0 0 sdc ONLINE 0 0 0 sdd ONLINE 0 0 0 sde ONLINE 0 0 0 sdf ONLINE 0 0 0 errors: No known data errors videoserver Filebench Version 1.5-alpha3 0.000: Allocated 177MB of shared memory 0.004: Video Server Version 3.0 personality successfully loaded 0.992: Populating and pre-allocating filesets 0.992: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 655360.000MB total size 0.992: Removing passivevids tree (if exists) 0.999: Pre-allocating directories in passivevids tree 1.000: Pre-allocating files in passivevids tree 1729.626: activevids populated: 16 files, avg. dir. width = 4, avg. dir. depth = 2.0, 0 leafdirs, 163840.000MB total size 1729.626: Removing activevids tree (if exists) 1729.647: Pre-allocating directories in activevids tree 1729.650: Pre-allocating files in activevids tree 2813.261: Waiting for pre-allocation to finish (in case of a parallel pre-allocation) 2877.042: Population and pre-allocation of filesets completed 2877.042: Starting 1 vidreaders instances 2877.046: Starting 1 vidwriter instances 2878.083: Running... 3478.226: Run took 600 seconds... 3478.226: Per-Operation Breakdown serverlimit 229829ops 383ops/s 0.0mb/s 116.479ms/op [0.001ms - 5003.542ms] vidreader 229959ops 383ops/s 95.8mb/s 35.886ms/op [0.015ms - 1221.401ms] replaceinterval 5ops 0ops/s 0.0mb/s 10000.117ms/op [10000.093ms - 10000.133ms] wrtclose 5ops 0ops/s 0.0mb/s 0.031ms/op [0.011ms - 0.096ms] newvid 5ops 0ops/s 85.3mb/s 94986.683ms/op [83254.878ms - 99920.156ms] wrtopen 6ops 0ops/s 0.0mb/s 24.418ms/op [0.892ms - 86.415ms] vidremover 6ops 0ops/s 0.0mb/s 204.072ms/op [153.622ms - 259.849ms] 3478.226: IO Summary: 229981 ops 383.211 ops/s 383/0 rd/wr 181.1mb/s 37.953ms/op 3478.226: Shutting down processes pool: zfstest state: ONLINE config: NAME STATE READ WRITE CKSUM zfstest ONLINE 0 0 0 sdb ONLINE 0 0 0 sdc ONLINE 0 0 0 sdd ONLINE 0 0 0 sde ONLINE 0 0 0 sdf ONLINE 0 0 0 sdg ONLINE 0 0 0 errors: No known data errors videoserver Filebench Version 1.5-alpha3 0.000: Allocated 177MB of shared memory 0.005: Video Server Version 3.0 personality successfully loaded 0.181: Populating and pre-allocating filesets 0.182: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 655360.000MB total size 0.182: Removing passivevids tree (if exists) 0.190: Pre-allocating directories in passivevids tree 0.191: Pre-allocating files in passivevids tree 1628.455: activevids populated: 16 files, avg. dir. width = 4, avg. dir. depth = 2.0, 0 leafdirs, 163840.000MB total size 1628.455: Removing activevids tree (if exists) 1628.481: Pre-allocating directories in activevids tree 1628.482: Pre-allocating files in activevids tree 2600.198: Waiting for pre-allocation to finish (in case of a parallel pre-allocation) 2679.050: Population and pre-allocation of filesets completed 2679.051: Starting 1 vidreaders instances 2679.054: Starting 1 vidwriter instances 2680.089: Running... 3280.259: Run took 600 seconds... 3280.260: Per-Operation Breakdown serverlimit 229746ops 383ops/s 0.0mb/s 119.297ms/op [0.001ms - 5019.507ms] vidreader 229876ops 383ops/s 95.7mb/s 23.844ms/op [0.018ms - 798.206ms] replaceinterval 6ops 0ops/s 0.0mb/s 10000.162ms/op [10000.095ms - 10000.393ms] wrtclose 6ops 0ops/s 0.0mb/s 0.018ms/op [0.012ms - 0.025ms] newvid 6ops 0ops/s 102.4mb/s 80760.431ms/op [68243.649ms - 85797.353ms] wrtopen 7ops 0ops/s 0.0mb/s 14.242ms/op [3.508ms - 29.752ms] vidremover 7ops 0ops/s 0.0mb/s 163.821ms/op [113.321ms - 253.538ms] 3280.260: IO Summary: 229902 ops 383.061 ops/s 383/0 rd/wr 198.1mb/s 25.954ms/op 3280.260: Shutting down processes pool: zfstest state: ONLINE config: NAME STATE READ WRITE CKSUM zfstest ONLINE 0 0 0 mirror-0 ONLINE 0 0 0 sdb ONLINE 0 0 0 sdc ONLINE 0 0 0 errors: No known data errors mirror videoserver Filebench Version 1.5-alpha3 0.000: Allocated 177MB of shared memory 0.004: Video Server Version 3.0 personality successfully loaded 0.200: Populating and pre-allocating filesets 0.201: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 655360.000MB total size 0.201: Removing passivevids tree (if exists) 0.209: Pre-allocating directories in passivevids tree 0.209: Pre-allocating files in passivevids tree 3989.161: activevids populated: 16 files, avg. dir. width = 4, avg. dir. depth = 2.0, 0 leafdirs, 163840.000MB total size 3989.161: Removing activevids tree (if exists) 3989.169: Pre-allocating directories in activevids tree 3991.420: Pre-allocating files in activevids tree 6635.083: Waiting for pre-allocation to finish (in case of a parallel pre-allocation) 6732.823: Population and pre-allocation of filesets completed 6732.835: Starting 1 vidreaders instances 6732.839: Starting 1 vidwriter instances 6733.873: Running... 7333.961: Run took 600 seconds... 7333.963: Per-Operation Breakdown serverlimit 152912ops 255ops/s 0.0mb/s 141.927ms/op [0.000ms - 5393.883ms] vidreader 153008ops 255ops/s 63.7mb/s 183.630ms/op [0.015ms - 5393.865ms] replaceinterval 1ops 0ops/s 0.0mb/s 10000.355ms/op [10000.355ms - 10000.355ms] wrtclose 1ops 0ops/s 0.0mb/s 0.029ms/op [0.029ms - 0.029ms] newvid 1ops 0ops/s 17.1mb/s 322159.216ms/op [322159.216ms - 322159.216ms] wrtopen 2ops 0ops/s 0.0mb/s 66.705ms/op [0.735ms - 132.675ms] vidremover 2ops 0ops/s 0.0mb/s 1238.683ms/op [953.302ms - 1524.064ms] 7333.967: IO Summary: 153014 ops 254.986 ops/s 255/0 rd/wr 80.8mb/s 185.745ms/op 7333.967: Shutting down processes pool: zfstest state: ONLINE config: NAME STATE READ WRITE CKSUM zfstest ONLINE 0 0 0 mirror-0 ONLINE 0 0 0 sdb ONLINE 0 0 0 sdc ONLINE 0 0 0 sdd ONLINE 0 0 0 errors: No known data errors mirror videoserver Filebench Version 1.5-alpha3 0.000: Allocated 177MB of shared memory 0.005: Video Server Version 3.0 personality successfully loaded 0.189: Populating and pre-allocating filesets 0.190: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 655360.000MB total size 0.190: Removing passivevids tree (if exists) 0.198: Pre-allocating directories in passivevids tree 0.199: Pre-allocating files in passivevids tree

Filebench Version 1.5-alpha3 0.000: Allocated 177MB of shared memory 0.007: Video Server Version 3.0 personality successfully loaded 0.217: Populating and pre-allocating filesets 0.218: passivevids populated: 64 files, avg. dir. width = 20, avg. dir. depth = 1.4, 0 leafdirs, 131072.000MB total size 0.218: Removing passivevids tree (if exists) 0.236: Pre-allocating directories in passivevids tree 0.255: Pre-allocating files in passivevids tree 834.077: activevids populated: 16 files, avg. dir. width = 4, avg. dir. depth = 2.0, 0 leafdirs, 32768.000MB total size 834.077: Removing activevids tree (if exists) 834.085: Pre-allocating directories in activevids tree 835.252: Pre-allocating files in activevids tree 1316.629: Waiting for pre-allocation to finish (in case of a parallel pre-allocation) 1359.341: Population and pre-allocation of filesets completed 1359.341: Starting 1 vidreaders instances 1359.343: Starting 1 vidwriter instances 1360.359: Running... 1960.444: Run took 600 seconds... 1960.444: Per-Operation Breakdown serverlimit 229809ops 383ops/s 0.0mb/s 105.441ms/op [0.001ms - 4999.823ms] vidreader 229915ops 383ops/s 95.8mb/s 81.328ms/op [0.015ms - 3940.711ms] replaceinterval 9ops 0ops/s 0.0mb/s 10000.138ms/op [10000.122ms - 10000.160ms] wrtclose 9ops 0ops/s 0.0mb/s 0.021ms/op [0.013ms - 0.033ms] newvid 9ops 0ops/s 30.7mb/s 55064.279ms/op [48648.814ms - 62854.277ms] wrtopen 10ops 0ops/s 0.0mb/s 217.470ms/op [1.007ms - 758.920ms] vidremover 10ops 0ops/s 0.0mb/s 316.273ms/op [1.410ms - 583.169ms] 1960.444: IO Summary: 229953 ops 383.201 ops/s 383/0 rd/wr 126.5mb/s 83.493ms/op 1960.444: Shutting down processes