

Re: strange performance dip shown by iozone

Source: <http://unix.derkeiler.com/Mailing-Lists/FreeBSD/performance/2004-02/0045.html>

mi+mx_at_aldan.algebra.com

Date: 02/20/04

To: David Schultz <das@FreeBSD.ORG>
Date: Fri, 20 Feb 2004 11:47:03 -0500

On Wed, Feb 18, 2004, mi+mx@aldan.algebra.com wrote:

=> I'm trying to tune the amrd-based RAID5 and have made several iozone
=> runs on the array and --- for comparison --- on the single disk
=> connected to the Serial ATA controller directly.

[...]

=> The filesystems displayed different performance (reads are better
=> with RAID, writes --- with the single disk), but both have shown a
=> notable dip in writing (and re-writing) speed when iozone used the
=> record lengths of 128 and 256. Can someone explain that? Is that a
=> known fact? How can that be avoided?

=This is known as the small write problem for RAID 5. Basically,
=any write smaller than the RAID 5 stripe size is performed using
=an expensive read-modify-write operation so that the parity can be
=recomputed.

I don't think, this is a valid explanation. First, there is no
"performance climb" as the record length goes up, there is a "dip". In
case of RAID5 it starts at higher level at reflen 4, decreases slowly to
128 and then drops dramatically at record lengths of 256 and 512, to climb
back up at 1024 and stay up. Here is the iozone's output to illustrate:

```
Size: RAID5: Single disk:
      KB reflen write write (Kb/second)
2097152 4 18625 17922
2097152 8 16794 17004
2097152 16 15744 23967
2097152 32 15514 20476
2097152 64 14693 18245
2097152 128 12518 17598
2097152 256 6370 29418
2097152 512 8596 35997
2097152 1024 16015 36098
2097152 2048 15588 35207
2097152 4096 16016 36832
2097152 8192 15907 37927
2097152 16384 15810 32620
```

freebsd-performance: Re: strange performance dip shown by iozone

I'd dismiss it as the controller's heuristics' artifact, but the single disk results show a similar (if not as profound) pattern of write performance changes. Could there be something about the FS?

Also, is the RAID5 writing speed supposed to be _so much_ worse, than that of a single disk?

=The solution is to not do that. If you expect lots of small random
=writes and you can't do anything about it, you need to either use
=RAID 1 instead of RAID 5, or use a log-structured filesystem, such as
=NetBSD's LFS.

This partition is intended to store huge backup files (database dumps mostly). Reading and writing will, likely, be limited by the (de)compression speed anyway, so the I/O performance is satisfactory as it is. I just wanted to have some benchmarks to help us decide, what to get for other uses in the future.

Thanks!

-mi

freebsd-performance@freebsd.org mailing list

<http://lists.freebsd.org/mailman/listinfo/freebsd-performance>

To unsubscribe, send any mail to "freebsd-performance-unsubscribe@freebsd.org"