Lightning Talk
FreeBSD storage performance

Benjeman Meekhof – AGLT2

USATLAS Tier2/Tier3 Workshop

University of Michigan, Ann Arbor

May 27th, 2008

Background

- Config: Dell PE2950, 2 Perc6/e RAID controllers, 4x15PD in MD1000 shelves.

- RAID setup: Two hardware RAID6 volumes, 30 PD each, 128k stripe, readahad, write-cache on. Devices mfid0 and mfid2 in OS.

- Hardware devices striped in software using "gstripe" – g for GEOM, disk framework enabling RAID and FS independent journals.

FreeBSD Storage Performance

5/27/2008

Soft Updates

- By default, FreeBSD's UFS2 filesystem uses "soft updates" to ensure filesystem consistency in the event of incomplete meta-data operations.

- Meta-data = directories, files, other constructs…data about where the data is for any filesystem.   If corrupted, filesystem is unusable.

- Soft updates can be described as "ordered meta-data updates" - avoid writing meta-data twice.

- More complicated but more efficient than a journaled system.

- Meta-data changes are kept and sorted in memory before going to disk. On-disk meta-data is possibly several minutes behind running system in heavy usage.

- journaling FS writes to journal on disk then to filesystem on disk - two hardware writes compared to only one using soft updates.

- After crash, filesystem is consistent and usable but needs fsck in background to possibly free space marked as "used" (no metadata in filesystem, but the disk blocks were written with data and marked used).

Tuning and Testing

- Testing done with 128K software stripe over two hardware based volumes using 128K stripe in hardware:
  - Command: "gstripe label -v -s 128k test /dev/mfid0 /dev/mfid2"
  - Bad results with default newfs (UFS2) settings
  - Tried many software stripe settings with no effect(-s)

FreeBSD striped UFS2 default settings (16384 blocksize)
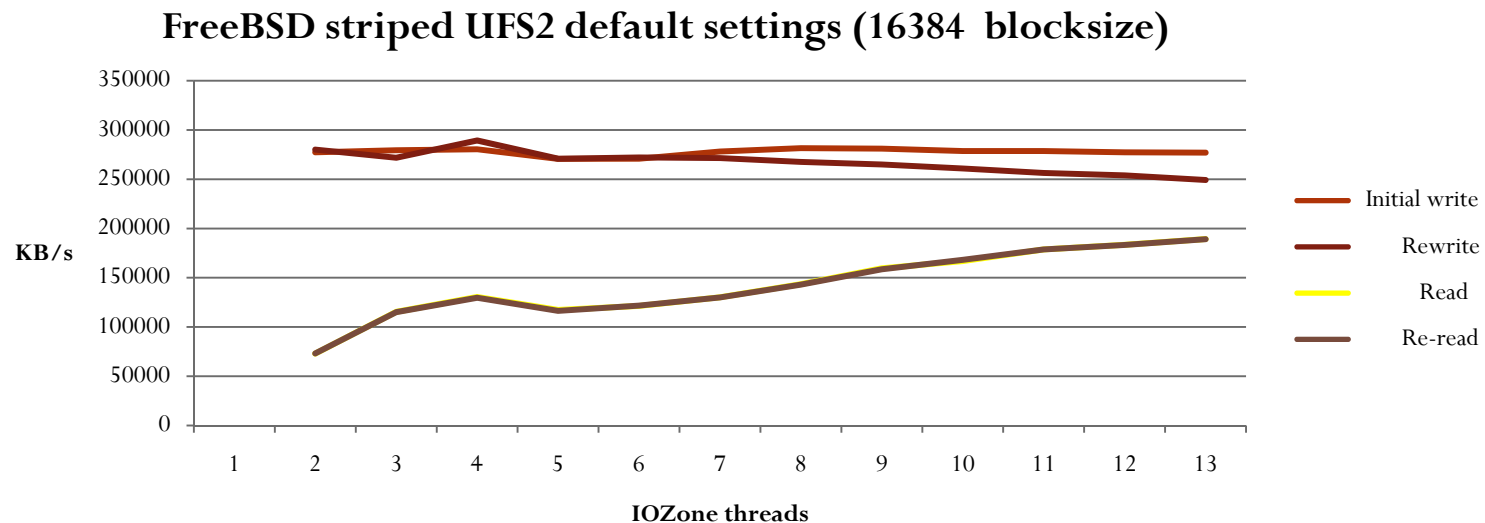
Final Results

- Finally the one key was to increase UFS2 blocksize to 65536 bytes with all other params equal:
  - gstripe label -v -s 128k test /dev/mfid0 /dev/mfid2
  - newfs -U -b 65536 /dev/stripe/test
  - Strange read anomalies around 4 threads…why???

FreeBSD striped UFS2, soft updates, 65536 FS blocksize

FreeBSD Storage Performance
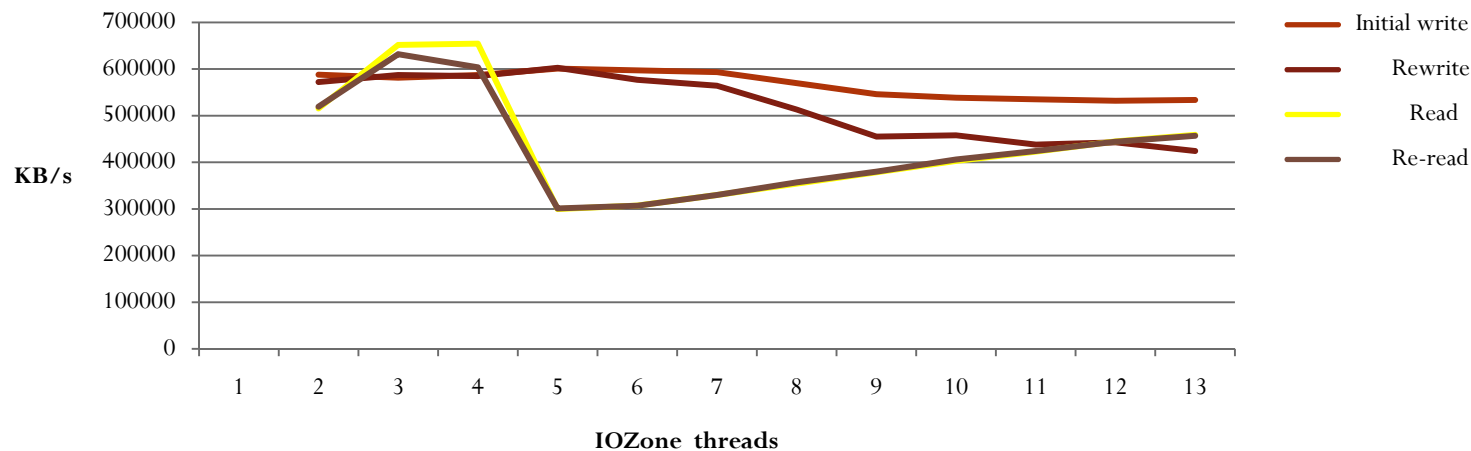
Other Notes

- GEOM journal was tested but did not perform better in general than the same filesystems using soft updates.

- ZFS is built in but experimental in 7.0-Release – should be able to outperform UFS2

- FreeBSD people are supportive, and answers to postings on freebsd-performance list confirmed my results with UFS2 even on other hardware.

- Basically, wait until ZFS is stable to see what FreeBSD can do in high-performance storage – these results are the best you can hope for with UFS2 for now.