CMU 18–746 **Storage Systems** **27 Feb 2008**

Spring 2008 **Exam 1**

Name: _____

Instructions

There are four (4) questions on the exam. You may find questions that could have several answers and require an explanation or a justification. As we've said, many answers in storage systems are "It depends!". In these cases, we are more interested in your justification, so make sure you're clear. Good luck!

If you have several calculations leading to a single answer, please place a box around your answer.

Problem 1 : Short answer. [48 points]

(a) Consider a 10 GB file accessed by a workload of random 2 KB reads. Is there a significant benefit to placing this file in the outermost zone of the disk? Explain why or why not.

No. Since transfer time will be a small portion of the access time for a random small access workload, the extra streaming bandwidth available at the outside of the disk will not be helpful.

(b) What value would you have been unable to determine for your disk in the "(a) Write-based SKIPPY" experiments of Lab 1, if the head switch time and the cylinder switch time values were identical? Explain why.

If the head switch time and cylinder switch times were the same, you would be unable to tell how many heads the disk has. The move from one head to the other look the same as a cylinder switch.

(c) In looking at the files on his 100 GB disk, Dr. Jones noticed 10 files for which the length value stored in the inode was 100 GB. Explain how all of those files could fit on the same 100 GB disk.

If the filesystem supports sparse files, the space consumed by the file and the length of the file are not necessarily the same value. The file length represents the distance from the first byte of the file to the last byte of the file. However, sections of the file may be empty, which the filesystem may choose to not store.

Alternately, the files could be hardlinks to the same data. Softlinks would not work, since the length in the inode would be length of the softlink data, not the file to which it points.

(d) Most disks' firmware prefetches physically sequential sectors from the media into the on-board buffer/cache memory. Explain (briefly) how file system data placement decisions can arrange to maximize the value of such prefetching despite having no direct knowledge of physical disk parameters (e.g., sectors-per-track or number of heads).

While filesystems may not know the exact physical parameters of any disk, they do know that disks map consequtive LBNs to sequential sectors. Thus, the filesystem can attempt to place data sequentially in LBN space in order to cause the disk to prefetch blocks that are likely to be accessed next.

(e) Imagine a file named /home/ganger/foo and a "symbolic link" file named /home/garth/foo that refers to it. If /home/ganger/foo is renamed to /home/ganger/bar, what will happen when someone tries to access /home/garth/foo?

The filesystem will return an error stating that the underlying file does not exist. This is because symbolic links point to another name in the filesystem. If the link were a hard link, this movement would not break the link.

(f) Some modern disks perform write-back caching, wherein write requests are reported complete once the corresponding data is transferred from the host into the on-board RAM. What problem for file systems can arise if the disk firmware's disk scheduler reorders the media writes in order to improve efficiency? Explain (briefly).

Recovery mechanisms for filesystems require that certain updates be made persistent, such as being written to disk media, in a specific order to work correctly. If the disk reports writes as complete without actually writing them stably onto the media, the filesystem may be unable to recover to a consistent state after a crash.

Problem 2 : Disk details. [28 points]

Consider the following disks.

	Seagate Cheetah4LP		IBM Ultrastar18ES	
Year	1996		1998	
Form factor	3.5” half-height		3.5” half-height	
Capacity	4.5 GB		9 GB	
Cylinders	6582		11474	
Surfaces	8		5	
Spindle speed	10033		7200	
Zone Information				
	firstcyl–lastcyl	*sectors per track*		
Zone 1	0–1343	195	0–377	390
Zone 2	1345–2448	187	378–1263	374
Zone 3	2450–3541	176	1264–2247	364
Zone 4	3543–4406	166	2248–3466	351
Zone 5	4408–5223	155	3467–4504	338
Zone 6	5225–5956	145	4505–5526	325
Zone 7	5958–6580	131	5527–7044	312
Zone 8			7045–8761	286
Zone 9			8762–9815	273
Zone 10			9816–10682	260
Zone 11			10683–11473	247

Table 1: *Specifications for the Seagate Cheetah 4LP and IBM Ultrastar 18ES.*

(a) Compute the cylinder and surface number for LBN 1,874,600 on the Seagate Cheetah 4LP.

Zone 1: $1344 * 195 * 8 = 2,096,640$ *So this LBN is in the first zone. So our cylinder will be* $1,874,600/(195 * 8) = 1,201$ *The remainder is* $1,874,600 - (1,201 * 195 * 8) = 1,874,600 - 1,873,560 = 1,040$. *So our surface will be:* $1,040/195 = 5$ *remainder* 65. *So our coordinate is:*

Cylinder 1,201, Surface 5 (Track offset 65)

(b) Compute the cylinder and surface number for LBN 1,874,600 on the IBM Ultrastar 18ES.

Zone 1: $378 * 390 * 5 = 737,100$ *Zone 2:* $1263 - 378 + 1) * 374 * 5 = 1656820$ *Total:* $737,100 + 1,656,820 = 2,393,920$

So the LBN is in Zone 2. So our cylinder offset in zone 2 is $(1,874,600 - 737,100)/(374 * 5) = 608$ *remainder* 540. *Actual cylinder is :* $378 + 608 = 986$. *Surface is* $540/374 = 1$ *remainder* 166. *So our coordinate is:*

Cylinder 986, Surface 1 (Track offset 166)

(c) What would be the expected average rotational latency for a 30000 RPM disk drive?

Expected rotational latency would be $1/(30000/60) = .002$*sec/2 = 1 msec.*

(d) Assuming an average seek time of 5 ms, what would be the average service time for random 1KB requests to a Seagate Cheetah4LP disk? (It is okay to approximate, but state any assumptions in doing so.)

Average service time would be: rotation latency+average seek+transfer time. The transfer time will change depending on the zone, however, even at its greatest value, this will only be $2/131 * 2 = .03$ *msec. This is insignificant compared to the seek and rotation times.*

Rotational latency = $(1/(10033RPM/60))/2 =$ 3msec Seek time 5ms

$2+1=6$ *so 8 msec.*

Problem 3 : More short answer. [24 points]

(a) Most file systems use the cylinder group (a.k.a. allocation group) concept to improve on-disk locality of related data and metadata blocks. If doing so reduces the average seek distance by a factor of two, why should one not expect a 50% reduction in average service time?

Two primary reasons. First, seek distance is not linearly related to seek times. Second, seek time is not the only factor in service time; both rotational latency and transfer time are also important.

(b) Is any update ordering still required for correctness when using write-ahead logging to provide metadata integrity? Explain.

Yes, the log updates must be transferred to disk before the "actual" filesytem updates. This is needed so the filesystem can redo committed updates and rollback unfinished updates during crash recovery. However, the filesystem no longer needs to assure any kind of ordering between metadata and data updates to the "actual" filesystem structures.

(c) Imagine an inode structure that uses 10 direct blocks, 2 indirect blocks, and one double indirect block. With an 8 KB block size and 32-bit unsigned integers for block pointers, what is the largest file size?

Pointers per block: $8 * 1024/4 = 2048$

Direct blocks: $10 * 8 * 1024 = 81,920$

Indirect blocks: $2 * 2048 * 8 * 1024 = 33,554,432$

Double indirect blocks: $2048 * 2048 * 8 * 1024 = 34,359,738,368$

Total size: $81,920 + 33,554,432 + 34,359,738,368 = 34,393,374,720$*bytes*$\approx$*32GB*

Problem 4 : Instructor trivia. [up to 2 bonus points]

(a) What company does our first guest lecturer work for?

David Anderson works for Seagate.

(b) What should Brandon do for six weeks after completing his Ph.D.?

Hint, Brandon loves climbing and beaches.

(c) Garth (Prof. Gibson) founded a company called PANASAS. What did the acronym ”PANASAS” stand for? (Hint: the first ’A’ is Advanced, and the last two letters are Application Software.)

Pittsburgh Advanced Network Attached Storage Application Software, or Pittsburgh Advanced Network Attached Storage Appliance Systems

(d) Where should Greg (Prof. Ganger) take his family for a few days of active vacation this summer?

Michigan, anyone?