CMU 18–746	**Storage Systems**	**25 Feb 2009**
Spring 2010	**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 $\boxed{\text{box around your answer}}$.

Problem 1 : Short answer. [56 points]

(a) Ext-2 does not allow creation of hard links to directories. What is one problem that is avoided by disallowing such hard links?

This is done in order to avoid the creation of cycles in the directory hierarchy.

(b) Arif is using a file system with write-ahead logging enabled. The system crashes immediately after he appends a megabyte of data to a file. After the system is rebooted and the log replayed, Arif notices that the file size is 1 megabyte larger, as he hoped, but that not all of the data he wrote during the append is in the file. What is the most likely cause of this problem?

EXT-3 was most likely being used in 'writeback' mode. In this mode, only metadata contents are journaled; file contents are not. This mode yields faster performance at the cost of lack of synchronization between data and metadata.

We received many answers that claimed that the system must have crashed before the data portion was written to the log, and hence only the metadata portion was replayed during recovery. This cannot happen—if data and metadata were both being journaled, both would be written to the log within the same transaction. Atomicity would guarantee that either both metadata and data would be updated when replaying the log, or neither would be updated.

For further information about EXT-3s journaling modes, we encourage you to read Wikipedia's excellent EXT-3 entry. However, specific knowledge about EXT-3 was not necessary to answer this problem.

(c) A disk has writes outstanding to the following cylinders: 2, 500, 7, 5, 1000, 3 and has just finished writing data to cylinder 600 and read data from cylinder 578 before that. List the order in which these writes (identified by cylinder number) will be serviced by each the following disk scheduling algorithms.

- FCFS:
- SSTF:
- SCAN (LOOK):
- C-SCAN (C-LOOK):

- *FCFS: 2, 500, 7, 5, 1000, 3*
- *SSTF: 500, 7, 5, 3, 2, 1000*
- *SCAN: 1000, 500, 7, 5, 3, 2*
- *C-SCAN: 1000, 2, 3, 5, 7, 500*

NOTE: SCAN and LOOK even though they are different - give the same result. Same applies to C-SCAN and C-LOOK

(d) When using a POSIX-based filesystem, how can an application quickly read byte 123,456?

lseek(fd, 123,456, SEEK_SET) and then issue a read call

(e) Suppose a disk can service 100 I/Os per second. Given a workload that issues requests at an exponentially distributed rate w/mean 20 I/Os per second, what is the utilization of the disk? What is the average time that a request spends in the disk queue?

$$
\begin{aligned}
Utilization &= \frac{20}{100} = 0.2 && (1)\\
&= 0.2 && (2)\\
AverageTimeServer &= \frac{1}{100} && (3)\\
&= 10ms && (4)\\
Average\ time\ in\ queue &= AverageTimeServer * \frac{util}{1 - util} && (5)
\end{aligned}
$$

$$
\begin{aligned}
&= 10ms * \frac{0.2}{0.8} && (6)\\
&= 2.5ms && (7)
\end{aligned}
$$

(f) In a large datacenter with 3000 machines, 100 failures are observed over a period of 2000 days. What is the resulting MTBF?

$$
\begin{aligned}
MTBF &= \frac{Number\ of\ days}{\frac{failures}{device}} && (8)\\
&= \frac{2000}{\frac{100}{3000}} && (9)\\
&= 60,000\ days && (10)
\end{aligned}
$$

Problem 2 : FSCK. [20 points]

Greg decides that he is going to implement his own version of myfsck for ext-2 during his spare time. Help Greg finish his implementation.

For this problem, recall the functionality of the four passes of myfsck:

- **Verify and fix directory pointers**: Verify for each directory: that the first directory entry is "." and self-references, and that the second directory entry is ".." and references its parent inode. If an error is found, correct the entry.
- **Insert unreferenced inodes into lost+found**: Check to make sure all allocated inodes are referenced in a directory entry somewhere. If an unreferenced inode is found, place it in the /lost+found directory.
- **Verify and fix inode link counts**: Count the number of directory entries that point to each inode (e.g., the number of hard links) and compare that to the inode link counter. If a discrepancy is found, update the inode link counter.
- **Verify the block allocation bitmap**: Walk the directory tree and verify that the block bitmap is correct. If a block that should (or should not) be marked in the bitmap is found, correct the bitmap.

(a) Greg first attempts to implement all of the passes of myfsck within a single depth-first directory traversal. Give one reason why this will not work.

Lost items can only be identified after the depth-first traversal. If these items are directories, the need to be traversed in a seperate traversal to fix "." and ".." errors, verify the link counts for their subdirectories (Pass 3), and identify allocated blocks (Pass 4).

Note that we received a lot of different answers for this question. Answers that did not explicitly mention lost items did not receive full credit.

(b) List the steps necessary to add an inode to lost+found, while keeping the filesystem consistent. Note the steps you list may include functionality from other passes of myfsck.

The steps are listed below. In order to receive full credit, students had to list the steps necessary to keep the filesystem consistent.

- *The item must be added to lost+found's directory. The item can be either added to a hole in the directory entry (dentry) list, or can be added as the last item in the directory. Holes can be found by finding dentries whose record length is larger than necessary. When adding the item to lost+found, it is important to set the type field so as to reflect the file type of the lost item (e.g., directory, file, etc.).*
- *If the lost item is a directory, its ".." directory entry must be modified to point at the lost+found directory.*

- *If the lost item is a directory, The link count of the lost+found directory must be incremented by one. It is also a good idea to check the link count of the lost item to make sure it is correct — it's link count should be 2 + the number of subdirectories it contains.*
- *The block allocation bitmap should be traversed to verify that the blocks pointed to by the item are marked as allocated. If the item is a directory, this check should be performed for all subdirectories and files*

(c) Greg decides to implement pass 2 of myfsck by using the following algorithm:

```
for (i = 1; i < num_total_inodes; i++) {
  if (inode is allocated) {
    traverse directory tree to see if it is referenced
    if (!referenced \&\& inode's link count > 0) {
      add inode to lost+found
    }
  }
}
```

This algorithm may insert more items in lost+found than strictly necessary. Explain a scenario in which this could happen.

Consider the case in which a directory which contains subdirectories is lost. If the inode number of any of those subdirectories are smaller than that of the outer directory, both the those subdirectories and the outer directory will be inserted into lost+found. As a result, these subdirectories will have been hardlinked in two places — in its outer directory and in lost+found. If the files contained in the lost outer directory or any of its subdirectories have lower inode numbers than their respective parent directories, then these items will also be inserted into lost+found. In both cases, it is only necessary add the outermost lost directory to lost+found, as its subdirectories and files will no longer be lost as soon as it is added.

Note that inserting a lost directory and its subdirectories to lost+found will result in a violation of ext-2s consistency requirements. The subdirectories will be hardlinked both in lost+found and in the now-found outer directory—ext-2 does not allow directories to be hardlinked multiple times in this manner. See the answer to Question 1(a) to understand why this is the case.

(d) Greg wants to make sure he understands link counts properly. He analyzes a directory (not the root directory) with 10 files and 2 subdirectories. What should its link count be?

Recall that the link count specifies the number of times the directory's inode is pointed to by a directory entry. Note that soft links are files that contain the path to the file or directory they reference and hence do not factor into the link count.

- **“.”**: *The directory contains a “.” directory entry that points back to itself.*
- **“..” dentries of the directory’s subdirectories**: *There are two; each adds a link count of one.*
- **“The entry in the parent that lists the directory**: *The parent must contain a dentry that points to the directories inode, or else the directory would not be found in the directory hieararchy.*

Directories cannot be hard linked in other locations (See the answer to question 1(a) for why).

Problem 3 : Homework related Question. [24 points]

(a) Figure 1 shows the results obtained by running a variant of the skippy algorithm. Label it with the rotational latency and head/cylinder switches. How would the graph change if the rotational speed of the disk were increase?

See Figure 1. This graph was created by timing writes to adjacent sectors on the disk. Since the distance between sectors written is less than the STM value, almost of these writes require a complete revolution of the disk before completing. This is what is shown by the 0-slope line in the graph. Its Y-intercept is simply the rotational latency of the disk. When adjacent sectors are seperated by track or cylinder boundaries however, the track or cylinder skew allows the write to complete without incurring any rotational latency. These cases are denoted by the sparse datapoints near the bottom of the graph. Their height represent head/cylinder switch times.

If the rotation speed of the disk were to increase, then the constant 0-slope line would move down, as the time needed to complete one revolution would decrease. The head/cylinder switch times would not change because these values are dependent on mechanical properties of the disk-head arm, not the disk's rotational speed.

(b) A raw device interface, which bypasses the filesystem cache, was used when running the Skippy experiments to obtain the graphs given in the homework and in the exam. Why was this necessary?

Skippy experiments are microbenchmarks that are used to measure the low-level disk characteristics. If the filesystem cache were not bypassed, reads and writes may not be sent to the disk (or serviced) during the required measurement interval.

(c) Joe has 10 disks available for his disk array, but needs help deciding whether to use RAID-5 or mirroring. Some of his considerations include: (1) no concerns about capacity (he doesn't expect to need even half of the total space), (2) a customer demand for tolerating at least one disk failure, (3) a workload consisting of reading and writing large files sequentially in their entirity. Make a recommendation and justify it.

It depends :). Here are two sample answers.

- *I recommend mirroring because, depending on which disks fail, mirroring can tolerate up to 5 failures.*

- *I recommend RAID-5 because it supports one disk failure (the minimum required by Joe) and will yield better aggregate throughput for streaming large files. 9 out of the 10 disks can be used to stream a single stripe of a large file, whereas only 2 disks can be used to stream data for any one file if the mirroring setup is used.*

We received many interesting answers for this question and considered each carefully in assigning credit.

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

(a) List one paper you have read so far for which Greg is an author.

(b) What is Greg's term for "poor performance" this semester (although, it never seems to change...)?

(c) Where are the Winter Olympics being held this year? Do you think the Olympics should have been moved to Pittsburgh, given all the snowfall here over the past few weeks?

(d) Was Garth happy or sad about the outcome of the U.S.A. vs. Canada ice hockey game? Why?

(e) Pick one instructor or TA and identify the winter olympic sport for which you think he or she would have the best chance of winning a gold medal. Why?

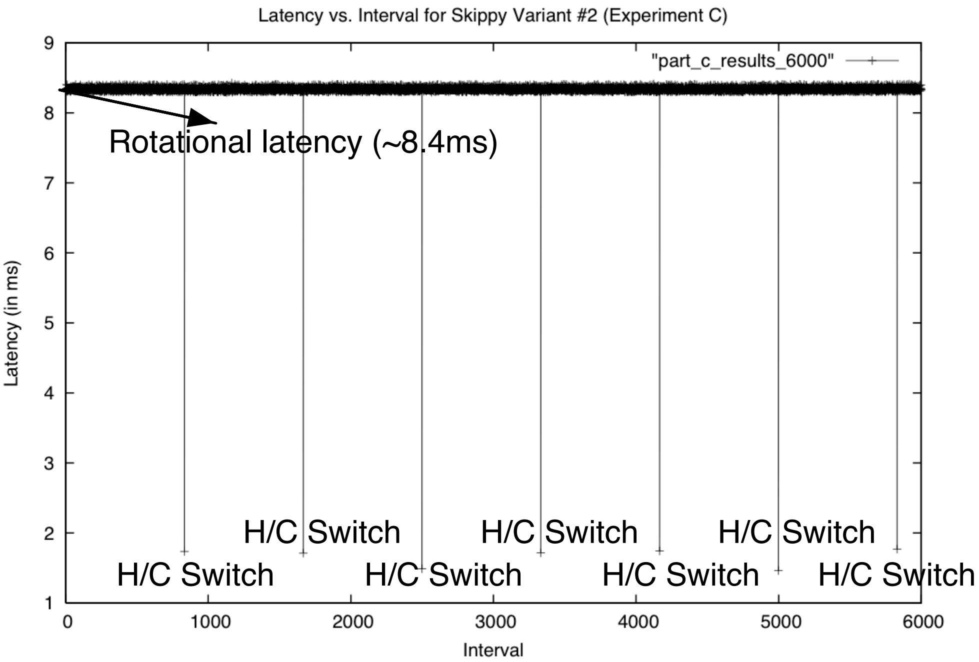

Figure 1: *Skippy results*