CMU 18-746/15-746 **Storage Systems** **21 April 2010**

Spring 2010 **Exam 2**

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. [48 points]

(a) Joe used FUSE to implement the client-side functionality of a simple distributed file system. On Joe's clients, the kernel-level FUSE module redirects each intercepted file system call to a user-level process that Joe wrote, called Joes_FS_client, using the libfuse library. Joes_FS_client sends each such call to a program running on the server, waits for a response, and provides that response to the FUSE module. If no response is received from the server within 100ms, Joes_FS_client resends the request. After extensively testing with file reads and writes, Joe extends his tests to include file creates and deletes, and is surprised to uncover rare error cases where the create fails with "file exists" or the delete fails with "file does not exist". Assuming that Joe's code works as described (i.e., no unspecified behavior), what is the most likely cause of the errors?

(b) Moe argues that the file system cache, currently kept in RAM, should instead be kept in Flash-based storage because Flash is cheaper per byte than RAM. What is the biggest performance concern you would raise with this design choice? (That is, why might it result in lower performance than using RAM for the file system cache.)

(c) Roe is planning to use 100 5-disk RAID-6 arrays in his disk array system. Assuming his system does not support rebuild, what will be the mean time to data loss for Joe's final disk array system (as a function of $MTBF_{disk}$)?

(d) Toe is shocked that BigTable always uses a 3-level B-tree, since it puts an upper bound on size and uses more internal index nodes than necessary for smaller tables. What is the biggest benefit that the BigTable creators would claim in favor of their design choice?

(e) Poe relies on a distributed log entry collection application. It consists of a process running on each node that periodically checks the local log for new records, opens the shared log file on a file server, appends the new records to the shared log file, and closes it. Explain the most likely reason he could observe some records being lost when using a bug-free NFS server to store the shared log file.

(f) Doe's organization has 1TB of user data and wants a daily backup. Each day, 100GB of user data is updated, and all user data (and modifications) are unique. How much data (maximum) is stored by the back-up system for each of the following backup strategies:

- keep one week of daily full backups:

- keep one weekly full backup and six daily incremental backups:

- keep one week of daily full backups with perfect de-duplication:

Problem 2 : More short answer. [24 points]

(a) Identify (for Koe) a workload (i.e., access pattern and file characteristics) for which performance will be better with AFS than with NFS.

(b) Loe and Woe use separate computers and each run a program for displaying a file's contents that open a file, keep it open, and re-read its contents every second. Loe updates the file (via a different program) and tells Woe, but Woe does not see the new version even after a few seconds. Which of these three distributed file systems could **not** be the one they are using? Sprite, AFS, or NFS. Explain your answer.

(c) Many parallel file systems stripe data across multiple servers in order to improve bandwidth. Foe is convinced that PLFS should be irrelevant, given such striping. Describe a workload for which he is incorrect and explain why the workload creates a performance problem (without PLFS).

Problem 3 : Solid State Disks. [28 points]

(a) Solid state disks (SSD) are built out of NAND flash technology. These parts are comparable in price to magnetic disks but store much less data. Why is the entire storage industry trying to figure out how to use SSDs in their storage systems?

(b) Flash-n-disk (your project 2) stores the contents of a file system split between two devices, a Seagate Barracuda magnetic disk and an Intel SLC SSD. Suppose Lin asked you to test another student's Flash-n-disk by creating 1 million files in the file system and then running "ls -lR" on it. The 'ls' takes about an hour to finish. Lin says she thinks it took too long and asks you two questions:

- approximately how long should the 'ls' have taken?
- what is the most likely mistake made in the student's Flash-n-disk implementation? Explain why it is wrong.

(c) Flash-n-disk uses one magnetic disk for every SSD, but SSDs are much more expensive per byte than a magnetic disk. The MBA student in the back says that we should amortize the SSD cost over more magnetic disks—that is, split the file system over one SSD and five magnetic disks. Then, to make this new Flash-n-disk even better, we decide that really big files, say $>$ 1 MB, will be striped over all the magnetic disks. This introduces new challenges for your implementation, because big files are stored with parts in each of five separate magnetic disks. Identify an implementation problem that this would create in the basic Flash-n-disk design and what would you do about it?

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

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

(b) What is this semester's record for the longest time with a single slide displayed (while lecturing) by an instructor in this class? Which instructor?

(c) How does Greg suggest implementing functionality in a decentralized system, whenever feasible?

(d) Which instructor had more pie thrown at him this semester?

(e) Which instructor or TA is most in need of a vacation? What should they do?