CMU 18–746 **Storage Systems** **23 April 2008**

Spring 2008 **Exam 2**

Name: _____

Instructions

There are four problems 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!

Problem 1 : Short answer. [56 points]

(a) Some application programs rely on temporary files to hold intermediate state during their computations. To avoid the need to "clean up" after a program crash, some programmers will use a sequence of steps like: (1) tmpfile = create(/tmp/tmpfile); (2) unlink(tmpfile); (3) compute(); (4) close(tmpfile). If /tmp is a UNIX local file system, these steps provide the program with a file to use during compute() that will automatically be reclaimed by the system if the program crashes (or exits). Explain what problem such a program will encounter if /tmp is an NFS file system.

(b) Most storage networks are said to use asymmetric protocols. For example, a SCSI target can issue a DISCONNECT or a RECONNECT against a SCSI initiator, while the reverse is not true. Why are asymmetric protocols used for storage networks?

(c) RAID level 5 fixed a bottleneck in RAID level 4. Why might the performance of a log-structured file system with a 100% read cache hit rate be the same for a disk array configured to use RAID-4 as for the same disk array configured to use RAID-5?

(d) Imagine a file system that computes and stores a checksum of each file in a separate "checksum database". Suppose that you read a file from a disk array, compute its checksum, and find that it does not match the value in the checksum database. Can use of RAID-5 guarantee the ability to correct whatever error has occurred? Explain.

(e) Once upon a time, Greg wrote a simulator for his research. Sometimes, the simulator would take many hours to complete. To help him monitor progress, he modified it to open a file (cleverly named "progress") when it starts and to write a line of text describing its progress every hour. When a friend offered to share his powerful compute server, Greg ran the simulator on the compute server, using a distributed file system to provide input files and a place to store the "progress" file. But, when he checked the "progress" file from his desktop computer three hours later, he found it empty even though the simulator appeared to be running. Later, when the program completed, it produced the proper output and the "progress" file had all of the expected output. What distributed file system was being used? Explain.

(f) CompanyX has created a file system that supports efficient snapshots based on copy-on-write and exposed in directories named /.snapshot-YYYY-MM-DD-HH-MM-SS. With their file system, they provide a backup tool that runs as an application using standard POSIX interfaces. CompanyX's lead engineer believes that the incremental backup feature of the tool could be made much more efficient (fewer disk reads and less data stored into the backup) by allowing it to access internal structures of the file system implementation. Explain how the engineer might achieve their claim.

(g) Fred has designed a direct-access storage system in which clients are provided with cryptographic capabilities that are checked by data servers before any client read or write request is serviced. One requirement given by the marketing team is that it must be possible to revoke a client's access to a given file's data immediately upon a change to that file's permissions. Fred's first thought is to implement this requirement by using a callback-like approach of having the file manager send a "throw away capability X" message to any affected client, when necessary. Identify and explain a problem with this approach.

Problem 2 : Disk Arrays. [19 points]

You are given a disk array with 4 disks. Assume that each disk has an average 7ms seek time, a 6ms full revolution time (i.e., 6ms to rotate 360 degrees), and a transfer rate of 80MB/s. The array is configured to use RAID-5 with a stripe unit size of 64KB.

(a) Assume a workload consisting of random aligned 4KB requests. Compute the average request service time and the maximum array throughput for each of these cases:

- 100% reads.

- 100% writes.

(b) Using mirroring would provide higher throughput for writes. Give one reason why an administrator might choose to use RAID-5 anyway.

Problem 3 : Distributed storage systems. [25 points]

(a) Direct-access SAN file systems come in two types: block-based and object-based. Which type involves more metadata being sent from the central file manager to clients? Explain.

(b) The PVFS "Trove" Storage layer maintains two different types of storage for files: bstreams and keyvals. Identify what PVFS uses each to store and one implementation difference between them.

(c) Thin provisioning is a marketing term for a "virtual" storage system in which real storage is not allocated to an LBN range in the storage space unless data is actually written to that range. Reads to LBNs that have not been written return zeros without going to disk. The marketing folks tell customers that they can configure a file system to believe it has a single huge disk (e.g., 100TB or more) even if they buy only a few TBs. But, if an old-style file system written for plain old SCSI disks uses such virtual storage, it can run out of space after awhile even when there is less user data in the file system than physical storage capacity. (For example, a scan of the file system might reveal only 1 TB of file data on a system with 4 1TB disks, but the virtual storage system pretending to have 100TB reports that it is out of space.) Why can this problem occur? (Hint: Petal has a special operation to fix this problem.)

Problem 4 : Bonus Questions. [Maximum of three points]

(a) Which guest speaker shared an office with Greg during graduate school?

(b) Which guest speaker earned their Ph.D. at the same school as Garth?

(c) Which guest speaker worked at the same school Brandon is attending for graduate school?

(d) Given 10 slides of material, which instructor would you expect to talk longest? How long?