CMU 18–746 **Storage Systems** **22 April 2009**

Spring 2010 **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) A UNIX system administrator runs a web server on a machine that gets all of its files from an NFS server. Using a different client machine, the administrator removed permissions for a number of the web server files, and doing so caused the web server to crash. The administrator is confused, because the exact same permission removal did not cause a crash the previous year, when the web server's files were all stored in a local file system on the same machine as the web server. Briefly explain why the use of NFS storage (instead of the local file system) could result in such a problem.

Answer: The web server likely had files open that became inaccessible when the permissions were changed, since the NFS server did not know about the files being open.
Explanation: In a local file system, an open call is checked against the ACL. Once the file is successfully opened, reads/writes pass through without such checks. However, in NFS every read/write is checked against the appropriate ACL. For the given case, the web server might have a few files open which were accessible on the local file system but were suddenly inaccessible over NFS.

(b) The basic SCSI bus has 8 data lines and can support up to 8 devices. Why can it not support more?

The SCSI arbitration phase requires each device to raise a particular data line based on its SCSI ID. Since, there are only 8 data lines, only 8 devices can be supported.

(c) A storage system engineer noticed something interesting when experimenting with different disk array configurations: RAID level 5 performs better than RAID level 4 (given the same set of disks) for a workload consisting entirely of small random reads. Briefly explain why.

If we have N disks in the parity group, a RAID 4 configuration will distribute the small reads over N-1 disks whereas a RAID 5 configuration will distribute the small reads over N disks (i.e., one more).

(d) In most parallel file systems, reliability is provided by RAID "underneath" (i.e., within) each server. Why does this approach require expensive dual-attached RAID controllers to provide high reliability?

RAID protects against failures of disks. But, other components within a server might fail causing those disks to be inaccessible. Protection against such server failures is usually provided by having dual-attached RAID controllers connected to two servers.

(e) A system administrator has 100 disks at his disposal. He is considering configuring them as either ten 10-disk RAID-5 arrays or one 100-disk RAID-5 array. Quantify the improved reliability that would come from the configuration with 10× more capacity being used for redundancy. (Assume a MTBF of 1000 days for each disk.)

$MTTF_{RAID5} = \frac{MTTF(disk)^2}{N \times (G-1) \times MTTR(disk)}$

Hence, $MTTF_{10arrays} = \frac{1000^2}{100 \times 9 \times MTTR(disk)}$ *and* $MTTF_{1array} = \frac{1000^2}{100 \times 99 \times MTTR(disk)}$

Hence, after cancelling common factors, $\frac{MTTF_{10arrays}}{MTTF_{1array}} = \frac{99}{9} = 11$

(f) Joe uses a high-end storage server that includes RAID support for reliability, LFS-based mechanisms for improved write performance, and snapshots for on-line backup. Unfortunately, he has almost completely filled the storage capacity. To fix the problem, he identified and deleted a number of large, old files. But, when he checked the capacity utilization right after deleting the files, he found that there was no additional free space. Briefly explain why.

Most likely: The old files are preserved as part of the snapshots kept in the system.
Another possibility: The LFS cleaner has not yet reclaimed the blocks freed by deleting the files.

(g) Janet needs to access some files in an AFS directory for her work. After authenticating properly to her client machine and to Kerberos, she finds that the system allows her to access some of the files in the directory and not others. Give one possible cause for this behavior.

Answer: the UNIX permissions for some of the files may not allow her access.
Explanation: To access a file in the directory following conditions should be satisfied:

- *She should have appropriate AFS permissions for the enclosing directory. Since she can access some files within the directory, this condition is satisfied.*
- *The standard unix permissions should allow access according to per-file permissions. This condition is probably not satisfied for the files that she can not access.*

Problem 2 : Disk Arrays. [20 points]

You are given a disk array with 5 disks. Assume that each disk has an average 6ms seek time, a 6ms full revolution time (i.e., 6ms to rotate 360 degrees), synchronized spindles, and a transfer rate of 100MB/s. The array is configured to use RAID-4.

(a) Assume a workload consisting of random aligned 100KB write requests. Compute the average request service time and the maximum array throughput for each of these stripe unit size options:

- 100KB. *Here, every write will involve a read-modify-write cycle.*
  Read cycle: Average service time = 6ms (seek) + 3ms (rotational latency) + 100KB / 100MBps (transfer time) = 10ms
  Write cycle: Average service time = 6ms (full rotation) = 6ms
  Hence, average service time = 10ms + 6ms = 16ms
  Maximum array throughput = 1/16ms = 62.5 requests/sec (Only one such request can be serviced at a time because of the single parity disk.)
- 25KB. *Since requests are aligned at 100KB every write can directly update the parity.*
  Hence, average service time = 6ms (seek) + 3ms (rotational latency) + (100KB / 4) / 100MBps (transfer time) = 9.25ms
  Maximum array throughput = 1/9.25ms = 108.11 requests/sec

(b) Suggest one simple configuration change (with no new hardware) that could significantly improve the throughput of the 100KB case.

Use RAID-5 instead of RAID-4. (Doing so allows multiple writes in parallel.)

Problem 3 : Distributed storage systems. [24 points]

(a) Accessing even one block of a file in most parallel file systems (e.g., PanFS or Lustre) involves RPCs to two different servers. Identify the two servers and the primary functions of the two RPCs.

The first RPC is sent to the metadata server (a.k.a. file manager) to retrieve locations and other info for file data. The second RPC is sent to the data server(s) (possibly an object storage server) to fetch the actual data.

(b) PLFS is designed especially to make parallel checkpointing fast. Briefly describe a workload for which PLFS would perform poorly.

A workload involving small random reads.

(c) An application that has an open AFS file does not see changes made to that file after the open() call. Given this AFS session semantic, why does the AFS server need to send a "callback break" message to the client system running that application, if the file changes?

The "callback break" message is sent to the client to make sure that any subsequent open() calls on the client trigger fetching of the new file from the server rather than using the cached copy.

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

(a) Which instructor will be moving to the new Gates building in August?

Garth

(b) Which instructor watches University of Michigan football games with his sons?

Greg

(c) What kind of camp (i.e., what theme) should Prof. Ganger's kids attend this summer?

Ice fishing

(d) What color was George Washington's white horse?

white