CMPSCI 691ST Systems **Fall 2011**

Lecture 15

Lecturer: Emery Berger *Scribe: Dan Barowy*

15.1 Redundant Arrays of Inexpensive Disks (RAID)

15.1.1 Background

RAID was motivated more by performance than by reliability. Many small, expensive disks used together have better performance characteristics than a single large expensive disk ("SLED"). Reliability is often *worse* with multiple disks, however RAID mitigates this problem with its variety of parity strategies. Error-correction codes allow for both detection and recovery when failures happen.

The mathematical model for RAID is intentionally simplified. Disk failures are assumed to be independent; in reality, they are not.

In RAID, the controller hides the physical disks, exposing a logical disk to the operating system. Failures of single physical disks are thus hidden by this abstraction layer. One downside is that the operating system's scheduling for this logical disk is essentially ignored by the RAID controller, including the UNIX command to flush data to disk ("sync") which is supposed to be guaranteed to write to the disk.

15.1.2 Alternatives for Reliability

15.1.2.1 Filesystem Techniques

Journaling keeps filesystem metadata consistent when power failures happen; there are no guarantees for other data on the disk, though. Journaling also does not protect you against disk failures.

ZFS provides some protection against data corruption through the use of hashes.

Filesystems are notoriously difficult to get right, so filesystem-level techniques are not common.

15.1.2.2 Solid-State Disks

Unclear whether these are more reliable than HDDs.

SSDs are traditionally made with NAND flash. Future technologies may use phase-change memory, however this is extremely costly at the moment.

SSDs have better random-access time than moving-head magnetic disks; it is essentially constant. Traditional HDDs have very high throughput for sequential reads but terrible seek performance.

An engineering issue with SSDs is that blocks have a limited number of write cycles before they deteriorate and can no longer be used. Thus disk writes need to be uniformly distributed across the disk. This technique is called "wear leveling". Blocks are also remapped to keep load uniform.

Writing is expensive in SSDs. SSDs are relatively power-hungry given that they have no moving parts– on par with traditional HDDs.

15.1.2.3 Backup

In practice, people just backup their data. Making copies of the data is the strategy here. Commonly people copy to disk or the cloud. However, tape is still popular in some places, and tape is very cheap and reliable. The advantage of RAID is that the system remains online when failures happen, however it is not a substitute for a backup strategy.

15.1.3 Why Don't RAIDs use SSDs?

Latency would be awesomely low.

However, two big issues: cost and power.

But it turns out that RAID is something of a niche. Why don't we have RAID in laptops? Physical space, cost, power are issues with mobile computers, and using RAID to create very large logical disks for this application is not needed.

Even data centers do not use RAID. Redundancy is handled at a higher level, and the smallest serviceable unit is the entire machine.

RAID is very popular in medium-sized installations, e.g., small businesses with single points of failure like a MS Exchange server.

15.2 Digression: "I want more RAM"

Why don't computers allow you to have more RAM? Emery's machine swaps all the time, and he thinks that if more RAM were an option, people would use it. "Build it and they will come". Good phrase, but Field of Dreams is a terrible movie.