Computer Architecture, Appendix D

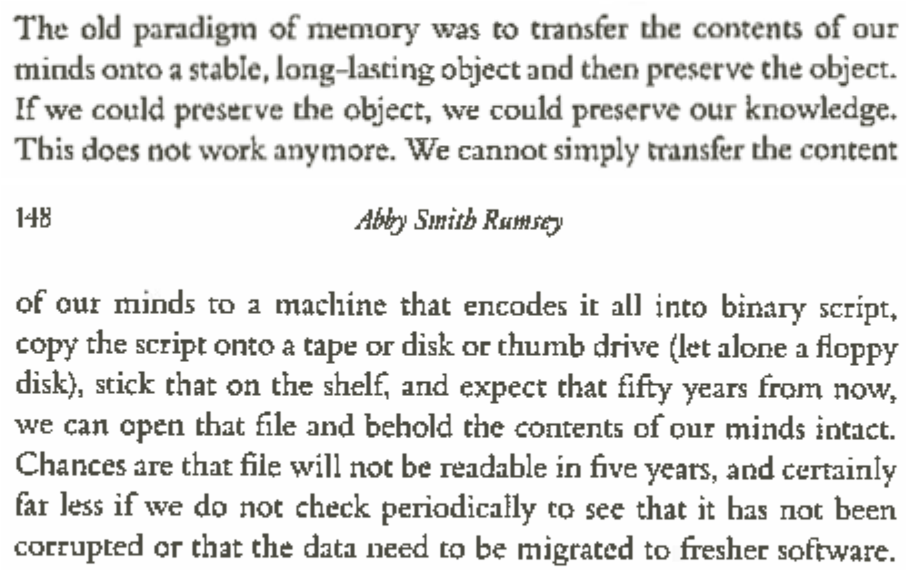
ECE 8405, Fall 2017

Table of Contents

1. Computer Architecture, Appendix D

Storage Systems

> Computer Architecture, A Quantitative Approach, Fifth Edition,
>
> John L. Hennessy and David A. Patterson, 2011.



The old paradigm of memory was to transfer the contents of our minds onto a stable, long-lasting object and then preserve the object. If we could preserve the object, we could preserve our knowledge. This does not work anymore. We cannot simply transfer the content

of our minds to a machine that encodes it all into binary script, copy the script onto a tape or disk or thumb drive (let alone a floppy disk), stick that on the shelf, and expect that fifty years from now, we can open that file and behold the contents of our minds intact. Chances are that file will not be readable in five years, and certainly far less if we do not check periodically to see that it has not been corrupted or that the data need to be migrated to fresher software.

> -- When We are No More: How Digital Memory Will Shape Our Future, by Abby Smith Rumsey, 2016.
> Excerpt: *When distracted ... we fail to build the vital repertoire of knowledge and experience that may be of use to us in the future. And it is the future that is at stake. For memory is not about the past. It is about the future.*

2. Access Time Gap

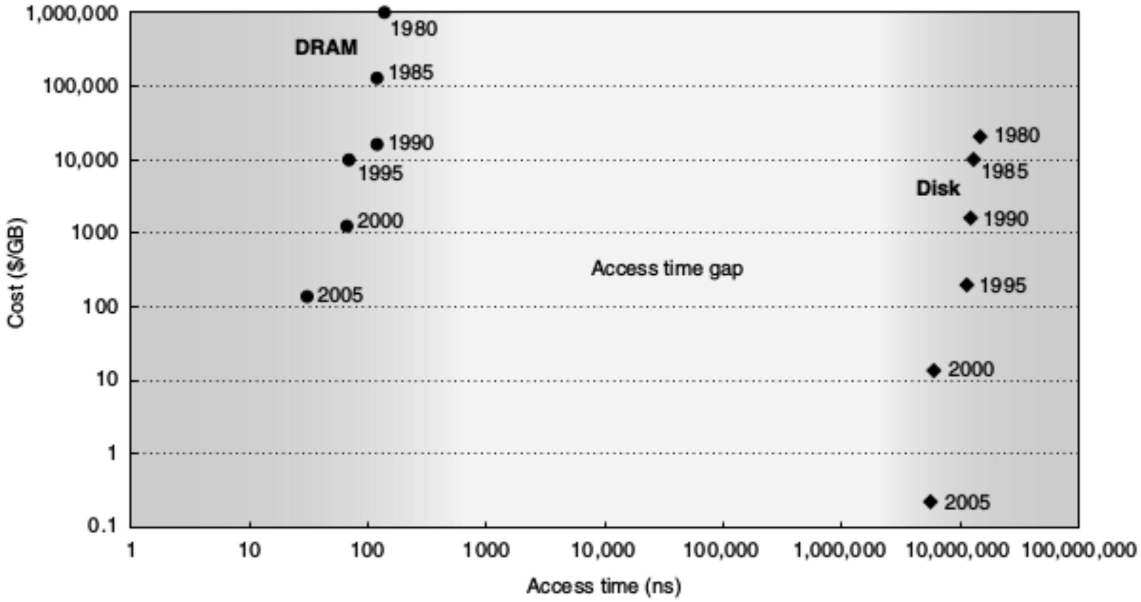


Figure D.1 Cost versus access time for DRAM and magnetic disk in 1980, 1985, 1990, 1995, 2000, and 2005. The two-order-of-magnitude gap in cost and five-order-of-magnitude gap in access times between semiconductor memory and rotating magnetic disks have inspired a host of competing technologies to try to fill them. So far, such attempts have been made obsolete before production by improvements in magnetic disks, DRAMs, or both. Note that between 1990 and 2005 the cost per gigabyte DRAM chips made less improvement, while disk cost made dramatic improvement.

DRAM latency is about 100,000 times less than disk, but costs 30 to 150 times more per gigabyte.

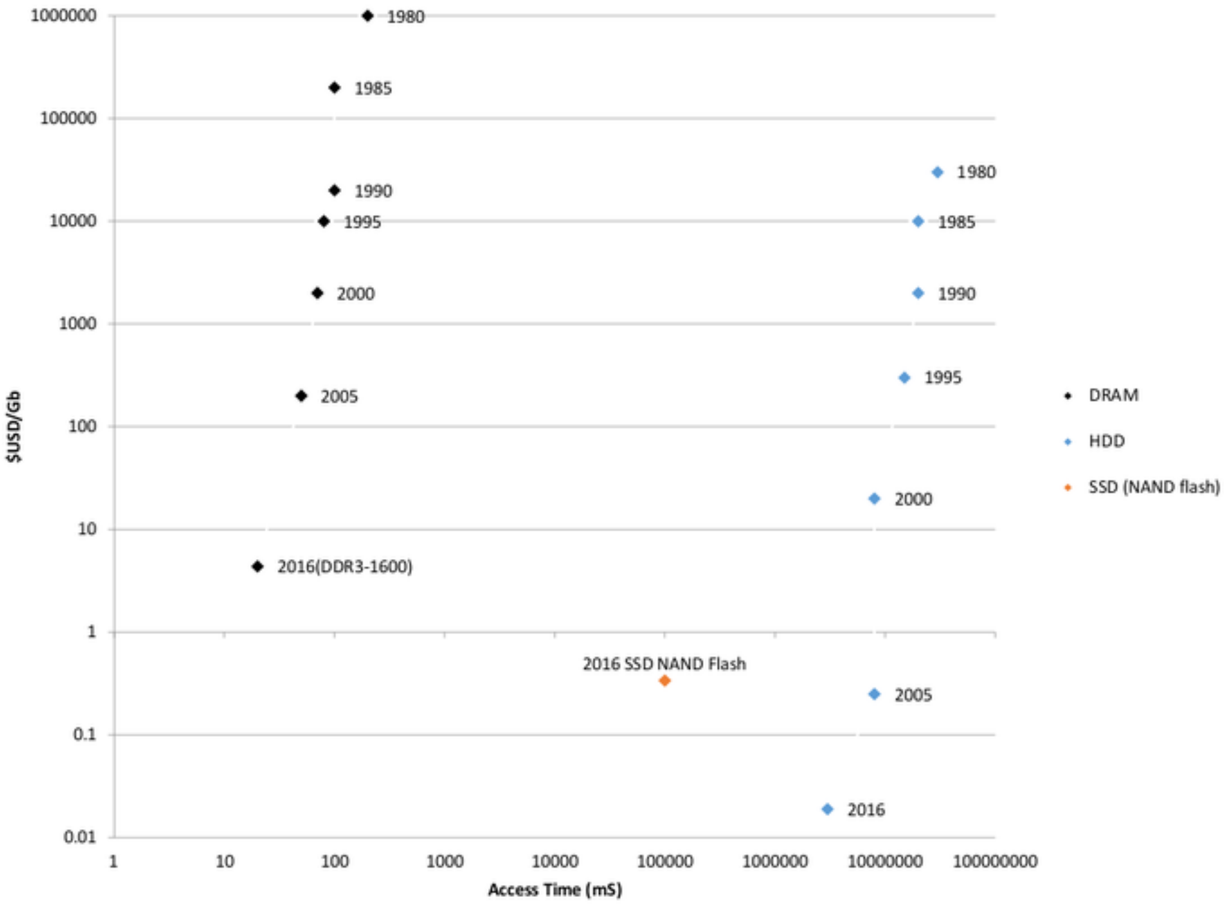
Disk: 600 GB, $400, 200 MB/sec

DRAM: 4 GB, $200, 16,000 MB/sec (80 times faster than disk)

Bandwidth per GB: 12,000 times higher for DRAM

Bandwidth per dollar: 160 times higher

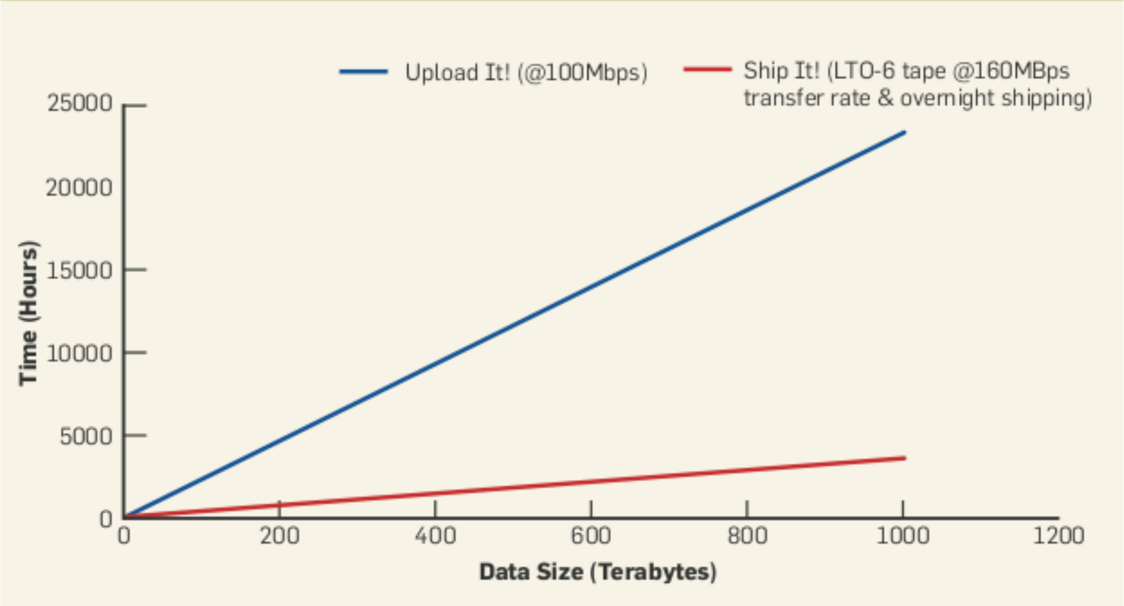
3. Access Time Gap - update



B. Warabak, Dec. 2016

4. Upload or Ship It?

Figure 4. Growth in data transfer time, 100Mbps vs. tapes.

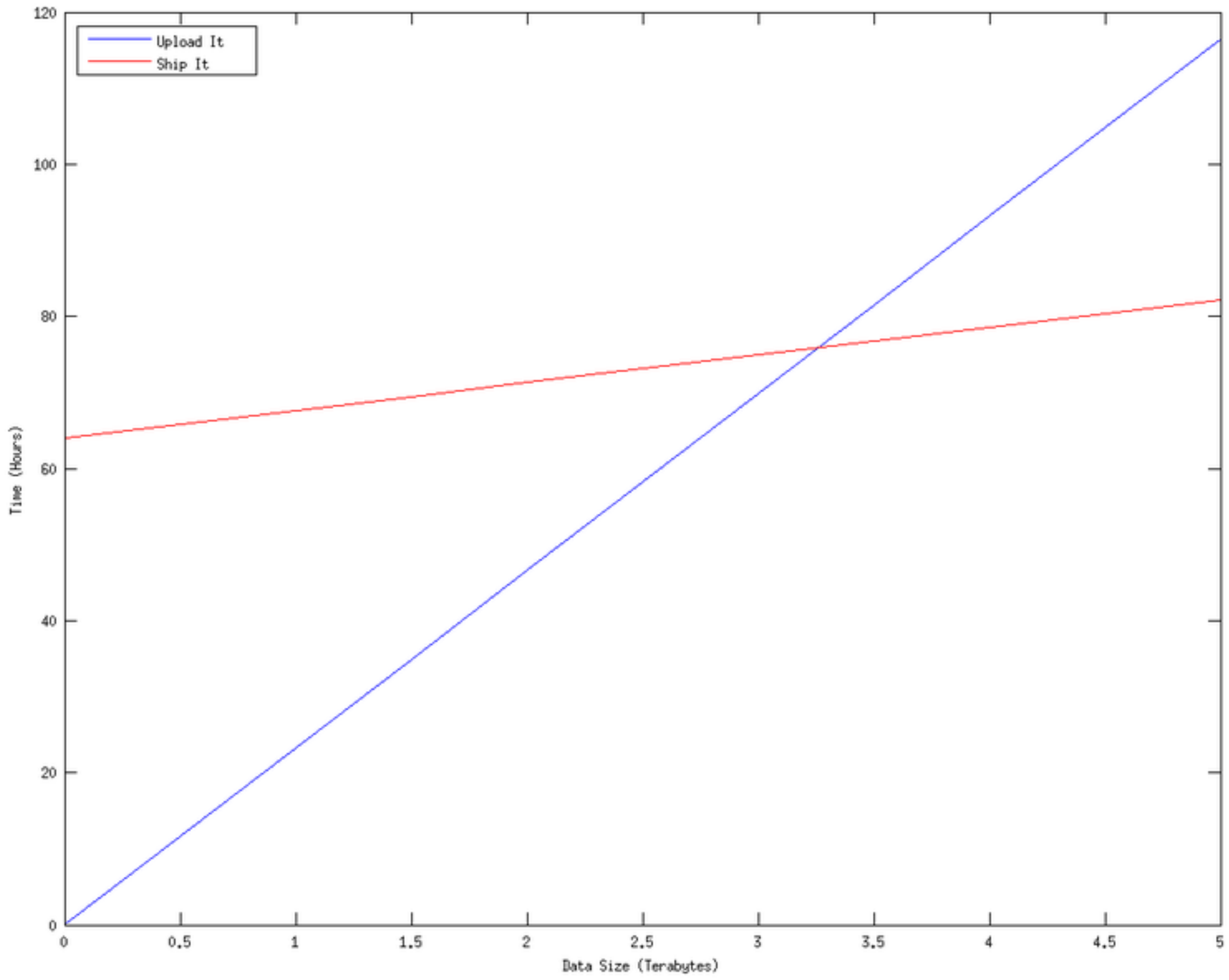


Should You Upload or Ship Big Data to the Cloud?, Sachin Date, CACM, July 2016.

equation (1) (missing from the paper):

```
TimeTransit_hours = 16;  TimeOverhead = 48;  SpeedIn_MB = 160;  SpeedOut_MB = 160;

% ship it
%
TransferTime_hours =   VolumeContent_MB / (3600 * SpeedIn_MB)  + TimeTransit_hours
                     + VolumeContent_MB / (3600 * SpeedOut_MB) + TimeOverhead;

% upload @ 100 Mbps
%
UploadTime_hours = VolumeContent_MB / (3600 * (100/8));
```

5. Upload or Ship It - Zoom

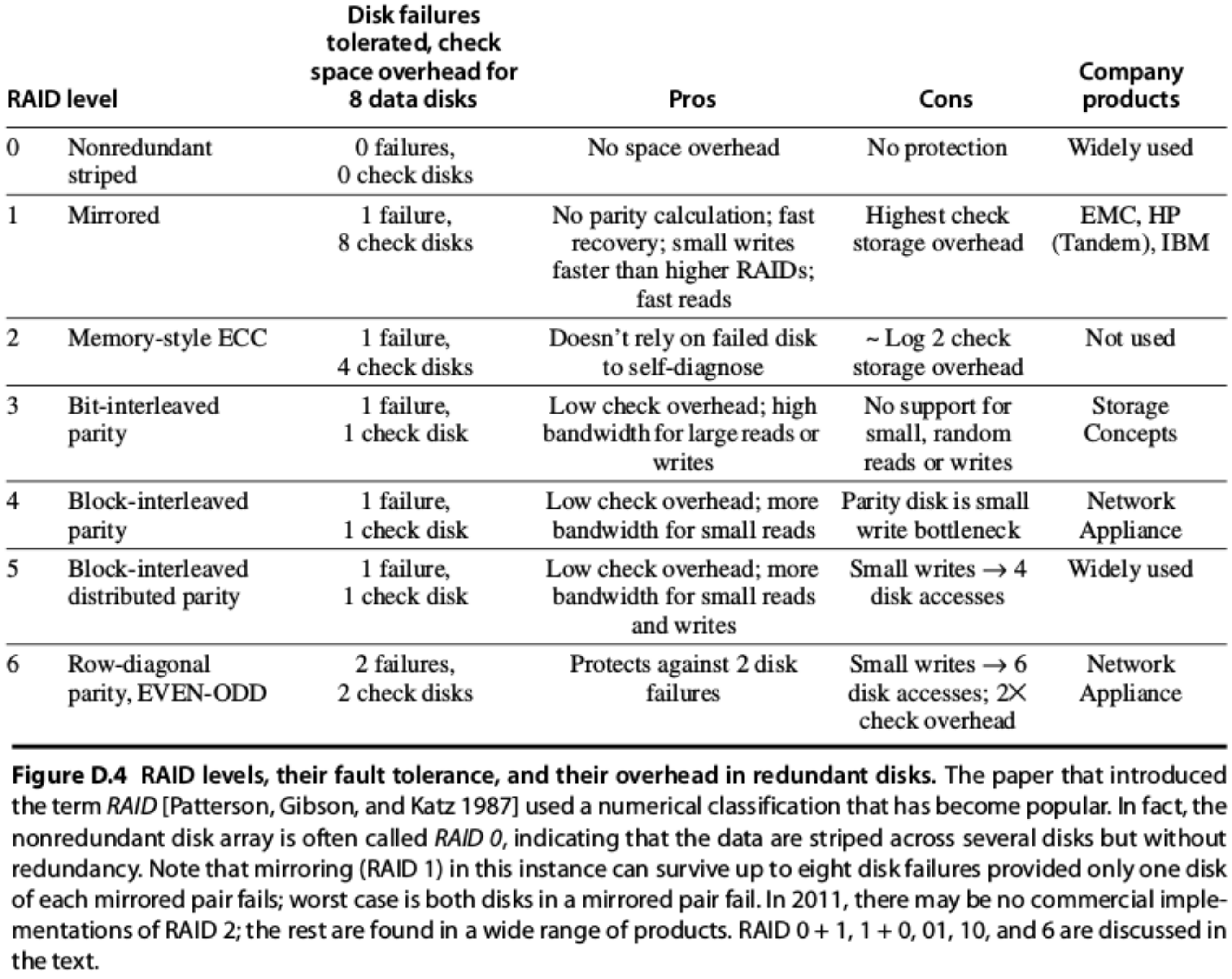


6. Genomic Data

Figure 1. (a) Moore's and (b) Kryder's laws contrasted with genomic sequence data.

Computational Biology in the 21st Century, Bonnie Berger, Noah M. Daniels, and Y. William Yu, CACM, August 2016.

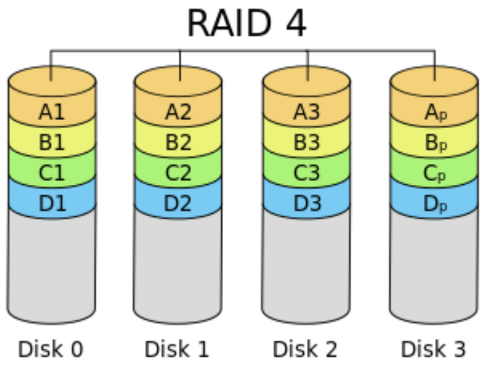
7. RAID

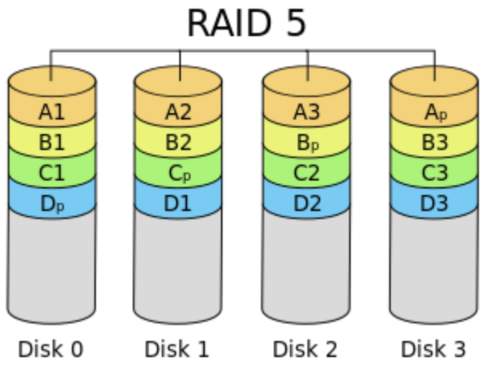


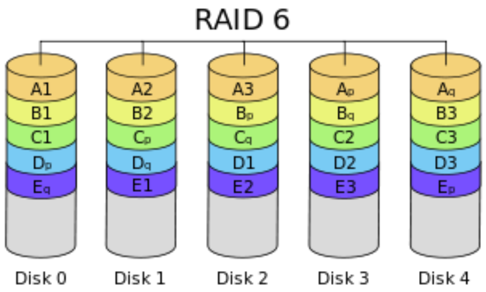
RAID level		Disk failures tolerated, check space overhead for 8 data disks	Pros	Cons	Company products
0	Nonredundant striped	0 failures, 0 check disks	No space overhead	No protection	Widely used
1	Mirrored	1 failure, 8 check disks	No parity calculation; fast recovery; small writes faster than higher RAIDs; fast reads	Highest check storage overhead	EMC, HP (Tandem), IBM
2	Memory-style ECC	1 failure, 4 check disks	Doesn't rely on failed disk to self-diagnose	~ Log 2 check storage overhead	Not used
3	Bit-interleaved parity	1 failure, 1 check disk	Low check overhead; high bandwidth for large reads or writes	No support for small, random reads or writes	Storage Concepts
4	Block-interleaved parity	1 failure, 1 check disk	Low check overhead; more bandwidth for small reads	Parity disk is small write bottleneck	Network Appliance
5	Block-interleaved distributed parity	1 failure, 1 check disk	Low check overhead; more bandwidth for small reads and writes	Small writes → 4 disk accesses	Widely used
6	Row-diagonal parity, EVEN-ODD	2 failures, 2 check disks	Protects against 2 disk failures	Small writes → 6 disk accesses; 2× check overhead	Network Appliance

Figure D.4 RAID levels, their fault tolerance, and their overhead in redundant disks. The paper that introduced the term *RAID* [Patterson, Gibson, and Katz 1987] used a numerical classification that has become popular. In fact, the nonredundant disk array is often called *RAID 0*, indicating that the data are striped across several disks but without redundancy. Note that mirroring (RAID 1) in this instance can survive up to eight disk failures provided only one disk of each mirrored pair fails; worst case is both disks in a mirrored pair fail. In 2011, there may be no commercial implementations of RAID 2; the rest are found in a wide range of products. RAID 0 + 1, 1 + 0, 01, 10, and 6 are discussed in the text.

8. RAID Levels 4, 5, 6







https://en.wikipedia.org/wiki/Standard_RAID_levels

9. RAID Level 6 Example

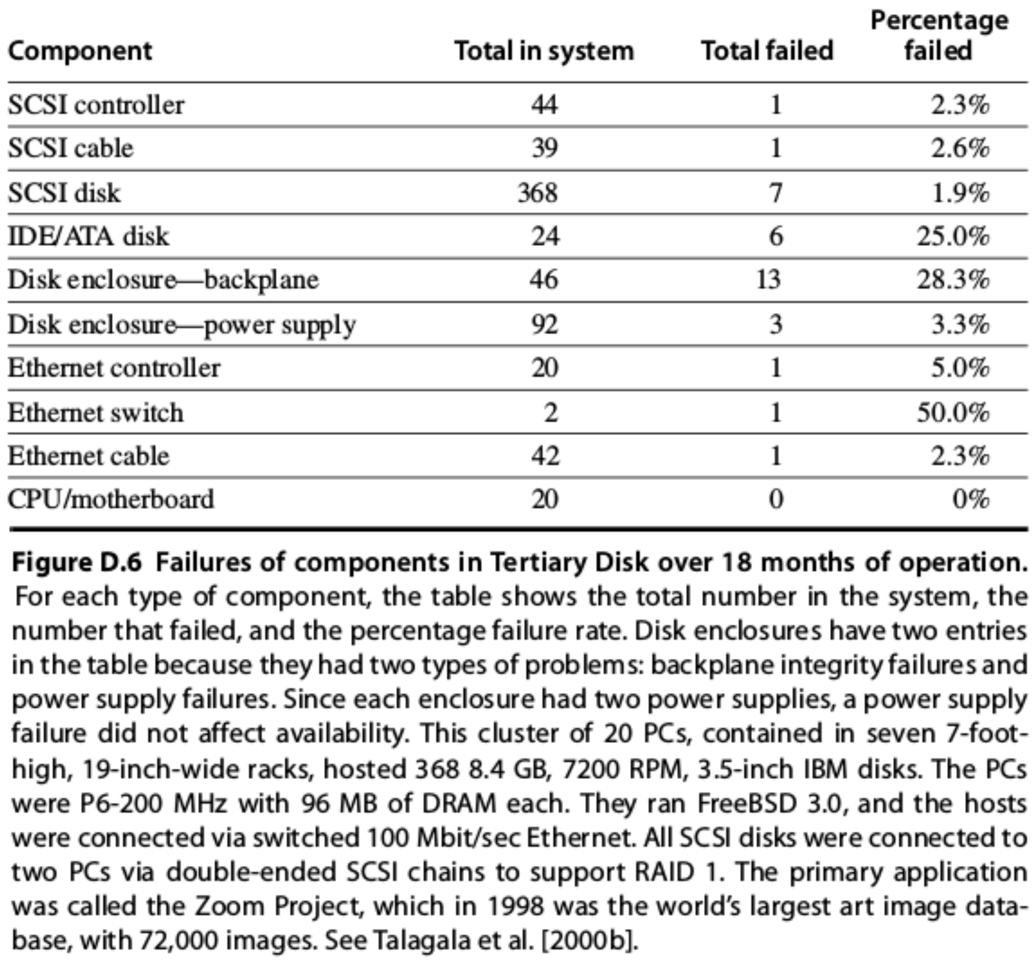
Data disk 0	Data disk 1	Data disk 2	Data disk 3	Row parity	Diagonal parity
0	1	2	3	4	0
1	2	3	4	0	1
2	3	4	0	1	2
3	4	0	1	2	3

Figure D.5 Row diagonal parity for *p* = 5, which protects four data disks from double failures [Corbett et al. 2004]. This figure shows the diagonal groups for which parity is calculated and stored in the diagonal parity disk. Although this shows all the check data in separate disks for row parity and diagonal parity as in RAID 4, there is a rotated version of row-diagonal parity that is analogous to RAID 5. Parameter *p* must be prime and greater than 2; however, you can make *p* larger than the number of data disks by assuming that the missing disks have all zeros and the scheme still works. This trick makes it easy to add disks to an existing system. NetApp picks *p* to be 257, which allows the system to grow to up to 256 data disks.

10. Linux mdadm Example

```
# df
Filesystem     1K-blocks     Used Available Use% Mounted on
/dev/md1        32858920  4738524  27785324  15% /
tmpfs            4024308      336   4023972   1% /dev/shm
/dev/md2        70429036 50379700  16465084  76% /home
/dev/sde1       70430128 57635932   9209892  87% /a
/dev/sdf1       61403764 23268544  35009432  40% /media/SD10
# mdadm --misc --detail /dev/md1
/dev/md1:
        Version : 1.0
  Creation Time : Sun May 27 17:03:43 2012
     Raid Level : raid1
     Array Size : 33516472 (31.96 GiB 34.32 GB)
  Used Dev Size : 33516472 (31.96 GiB 34.32 GB)
   Raid Devices : 2
  Total Devices : 2
    Persistence : Superblock is persistent

  Intent Bitmap : Internal

    Update Time : Tue Aug  9 09:34:48 2016
          State : clean
 Active Devices : 2
Working Devices : 2
 Failed Devices : 0
  Spare Devices : 0

           Name : vecr.ece.villanova.edu:1  (local to host vecr.ece.villanova.edu)
           UUID : 3ee16fb8:8ae32795:73708c46:69d25403
         Events : 6382

    Number   Major   Minor   RaidDevice State
       0       8        2        0      active sync   /dev/sda2
       1       8       18        1      active sync   /dev/sdb2
#
```

11. Failure Measurements Example



Component	Total in system	Total failed	Percentage failed
SCSI controller	44	1	2.3%
SCSI cable	39	1	2.6%
SCSI disk	368	7	1.9%
IDE/ATA disk	24	6	25.0%
Disk enclosure—backplane	46	13	28.3%
Disk enclosure—power supply	92	3	3.3%
Ethernet controller	20	1	5.0%
Ethernet switch	2	1	50.0%
Ethernet cable	42	1	2.3%
CPU/motherboard	20	0	0%

Figure D.6 Failures of components in Tertiary Disk over 18 months of operation. For each type of component, the table shows the total number in the system, the number that failed, and the percentage failure rate. Disk enclosures have two entries in the table because they had two types of problems: backplane integrity failures and power supply failures. Since each enclosure had two power supplies, a power supply failure did not affect availability. This cluster of 20 PCs, contained in seven 7-foot-high, 19-inch-wide racks, hosted 368 8.4 GB, 7200 RPM, 3.5-inch IBM disks. The PCs were P6-200 MHz with 96 MB of DRAM each. They ran FreeBSD 3.0, and the hosts were connected via switched 100 Mbit/sec Ethernet. All SCSI disks were connected to two PCs via double-ended SCSI chains to support RAID 1. The primary application was called the Zoom Project, which in 1998 was the world's largest art image database, with 72,000 images. See Talagala et al. [2000b].