Physics of Non-Volatile Digital Data Storage Technologies

Tenzin Rigden

Carleton College, Department of Physics, Northfield, MN 55057

(Dated: April 24, 2015)

Data storage technologies have become increasingly important as more and more of our lives are being stored digitally in our phones and computers, and the need for higher density and faster data storage technologies increases. In this paper, I will talk about modern techniques such as optical disks, hard disk drives, and flash storage that are currently being used. In addition, I will talk about currently in development technologies such as holographic data storage and probe based storage that promises much higher densities than currently offered.

I. INTRODUCTION

Computers and phones have become such a ubiquitous part of our lives that it is becoming more and more difficult to live without them. Just as important as these devices themselves are their contents. Your pictures, music, videos, papers, and more are all stored digitally on the devices. As more and more of our lives are recorded digitally, it has become increasingly important to find new ways to either expand our current data storage technologies or find new ones. On a fundamental level, a computer stores everything in a series of 1s or 0s called bits. In this binary system, 1 represents a "true" state while a 0 represents a "false" state. By using this binary system, it is possible to represent numbers as a sequence of 1s and 0s where each digit starting from the right represent an increasing power of 2 starting with 2^0. For example, the number 13 can be represented as 1101 so we get, from the right, $1*2^0 + 0*2^1 + 1*2^2 + 1*2^3$ which adds up to 13. Since we can represent numbers, we can use these numbers to represent other characters. One method is to use the American Standard Code for Information Interchange (ASCII) encoding system which was first published in 1963. ASCII encodes 128 characters that include letters, both lower case and capital, numbers, and other special characters into 7 bit integers. For example, the capital letter T corresponds to the 84th character and thus in binary can be represented as 01010100. A series of these binary numbers can be used to represent text.

Now that we know we can use binary to represent text, we can look at how we can physically store these 1s and 0s. One of the earlier methods was to use punch cards with holes and not holes representing 1s and 0s respectively. The early punch cards used 36 bits words because they

were used in calculators and they wanted to be able to represent 10 decimal places. However this method is not very dense, data storage wise, and is limited by the number of holes that could be fit on a piece of paper.

In this paper, I will talk about more modern techniques that are currently being used for non-volatile memory, which means that data will not be lost if the device does not have power. I will begin with optical data storage, specifically optical disks. Then I will continue on to magnetic data storage specifcally talking about hard disk drives and their advancements. After that will be flash memory storage which promises much faster access to data. Lastly, I will talk about a state of the art technique currently being developed called holographic data storage which promises to greatly increase storage density.

II. OPTICAL DISK STORAGE

Optical disks are a form of media storage that uses the grooves in the disk to store data. They have been commonplace in the United States since their invention in 1958 [1]. While optical disks aren't typically the preference by consumers for storing data directly from the computer, they are frequently used to archive or transfer data between machines. For optical disks, data is recorded and read using a laser beam. To record data onto a writable disc, a laser beam is shone on the reflective surface creating a pit, a tiny indentation, with a depth of the wavelength of the laser divided by 4. As seen in Fig. 1, to read data, a very low power laser is shone on the disk and the reflected signal is converted to an electrical signal using a scanning photo detector. A binary value of 1 is recorded when there is a change in the surface depth due to switching to a pit to land or vice versa, and a 0 is interperted as when there is no change in the surface depth. When the laser is shone over an intersection between a pit and a land, there will be light reflected from the pit and the land resulting in two beams [1]. These two beams will then destructively interefere back at the photo-diode. To see how they interfere, we can look at it quantitatively.

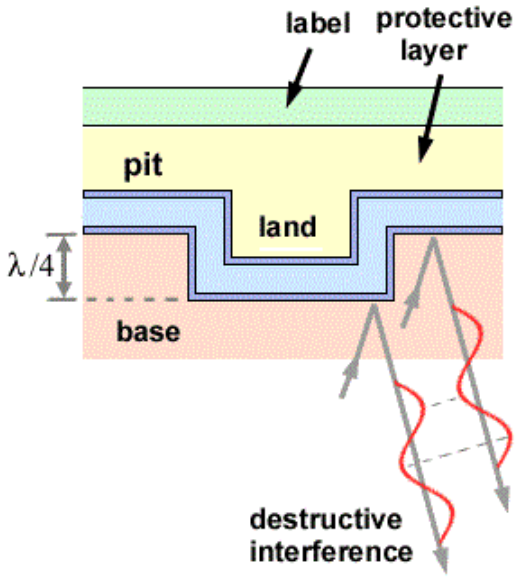

FIG. 1. This is a cross-sectional view of a CD-ROM where there a pit signifies a $\lambda/4$ depth meaning when a laser shines over a change from land to pit or vice versa, destructive interference will take place signifying a bit value of $1 http://www.hk-phy.org/articles/cdrom/cdrom_work_e.gif$.

A. Destructive Interference

When two waves arrive at the same point the superpostion of those waves creates a new wave. Interference depends on the superposition of the waves and certain conditions. In our example, since we have a laser beam effectively being split into two beams, we have coherent light (there is a constant phase difference between two beams), and both beams have the same frequency. Then, we can define the electric field of the two beams as

$$\begin{aligned} \vec{E}_1 &= \vec{E}_{01} cos(ks_1 - \omega t + \phi_1) \\ \vec{E}_2 &= \vec{E}_{02} cos(ks_1 - \omega t + \phi_1), \end{aligned} \tag{1}$$

where $\vec{E}_{01}$ and $\vec{E}_{02}$ are the amplitudes for the wave, k is defined as $\frac{2\pi}{\lambda}$, s_1 and s_2 are defined as distance traveled by each beam, and ϕ_1 and ϕ_2 are the phases of the beams at t = 0 at their source [2]. Since we have the same source for both of our beams, the amplitudes and the phase at t = 0 are the same. Next, the superposition of these two waves at a point P is

$$\vec{E}_P = \vec{E}_1 + \vec{E}_2. \tag{2}$$

The measurement of the effect of this wave on our eyes or a detector depends on the energy of the light beam. The irradiance, I, is the time average of the square of the wave amplitude and is is used to measure effects of waves [2]. The time average of the irradiance is done by the detector; our eye for example has an averaging time of 1/30 of a second. The irradiance is defined as

$$I = \epsilon_0 c < \vec{E}_P \cdot \vec{E}_P >, \tag{3}$$

where ϵ_0 is the vacuum permittivity and c is the speed of light [2]. By subbing in Eq. 2 for $\vec{E}_P$ and simplifying we get

$$I = \epsilon_0 c < \vec{E}_1 \cdot \vec{E}_1 + \vec{E}_2 \cdot \vec{E}_2 + 2\vec{E}_1 \cdot \vec{E}_2 > . \tag{4}$$

The first two terms correspond to the irradiances of the individual beams while the third term corresponds to the interaction between the waves and is the interference term, I_{12} [2]. Then, I can be rewritten as

$$I = I_1 + I_2 + I_{12}. \tag{5}$$

The interference term, I_{12}, by using a trigonemtric identity and simplifying, can be written as

$$I_{12} = 2\sqrt{I_1 I_2} < cos\delta >, \tag{6}$$

where δ is the phase difference and is defined as

$$\delta = k(s_2 - s_1) + \phi_2 - \phi_1. \tag{7}$$

Finally, we can write I as

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} < cos\delta > . \tag{8}$$

Destructive interference will yield a minimum value when $cos\delta$ is -1. In our case, since the beams have the same source, the laser, ϕ_1 and ϕ_2 are the same and their difference is 0, so δ will be 0 when the path length difference between the beams is $\frac{(2m+1)\lambda}{2}$ where m is an integer. Returning to our CD, since the pit has a depth of $\lambda/4$, the light reflected off of the pit will have a path length difference of $\lambda/2$ relative to the light reflected off of the land. This means that δ is π so $cos\delta$ is -1. Since the sources for the beams are the same, $I_1 = I_2$, so our I term is now

$$I = I_1 + I_1 - 2\sqrt{I_1 I_1} = 0. \tag{9}$$

This means at the change between a land and a pit, the two reflected beams will completely destructively interfere and no light will be reflected to the photo diode, which is then recorded as a 1. When there is no change in the surface depth, the light simply reflects back with no phase difference and is recorded as a 0 by the photo-diode [1].

B. Disk Types and Layers

Optical Disks come in a few different styles: Compact Disk ROM (CD-ROM), CD-Recordable (CD-R), and Digital Versatile Disk (DVD). These disc technologies are written sequentially in a continous spiral track eminating from the middle out, the distance between each track is called the pitch. A CD-ROM is a disc that's only meant to be readable and is typically used for things such as installation of programs. Since it is only written once in the factory, a glass master disk is created using a high intensity laser. Liquid polycarbonate is inserted into the master disk to create a copy which is then covered with a reflective layer and then a protective layer over that. A CD-R is slightly different because it needs to be writable once outside of the factory. This is done by changing the structure of the disk to have a protective layer followed by a reflective layer, a painted layer, and a transparent layer as seen in Fig. 2. The painted layer is initially transparent and permantly becomes dark to simulate a pit when impacted by a high-intensity laser. This can then be read by a lower intensity laser that does not cause the painted layer to turn dark.

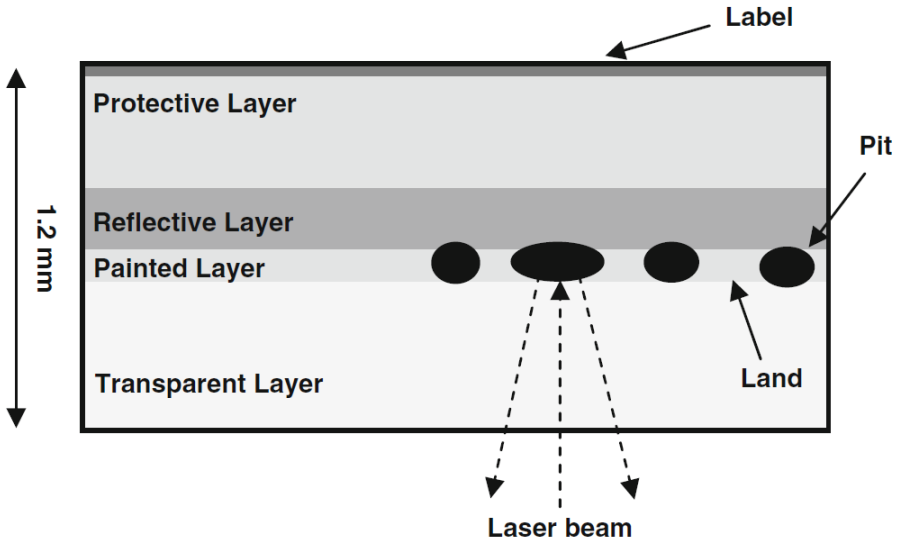

FIG. 2. In this cross sectional view of a CD-R, the painted layer, if shone on by a high intensity laser will develop dark spots simulating pits [1].

C. Rayleigh's Criteron due to Diffraction

While CDs are useful and easy to carry around, their size was limited to only 700 MB until the early '90s when demand for higher storage capacities in the 1990s meant that the CD had to change. Since the 1 and 0 data values are stored in the pits and lands, the way to increase storage density was to decrease their size and also place tracks closer to each other by decreasing their pitch. However, these values can only be decreased to a certain point due to diffraction. Diffraction is the phenomenon where light that propagates through an opening interferes with itself causing a diffraction pattern. For lasers which use a circular aperture, the opening through which light travels, the diffraction pattern produced (as seen in Fig. 4) is called an Airy Disk, with the bright circles representing maxima and the dark circles representing minima. This shape can be derived by starting with the electric field at point p through a circual aperture which can be found by

$$E_p = \frac{E_A}{r_0} e^{i(kr_0 - \omega t)} \iint\limits_{Area} e^{isksin\theta} dA, \tag{10}$$

where E_A is a constant factor that determines the strength of the electric field in the aperture, k is the same as before $\frac{2\pi}{\lambda}$, dA is an elemental area of the aperture as seen in Fig 3, s is the radial distance from the center of the aperture to the dA, θ is the angle of the optical path relative to the axis orthogonal to the aperture, and r_0 is the optical path length to the point P [2].

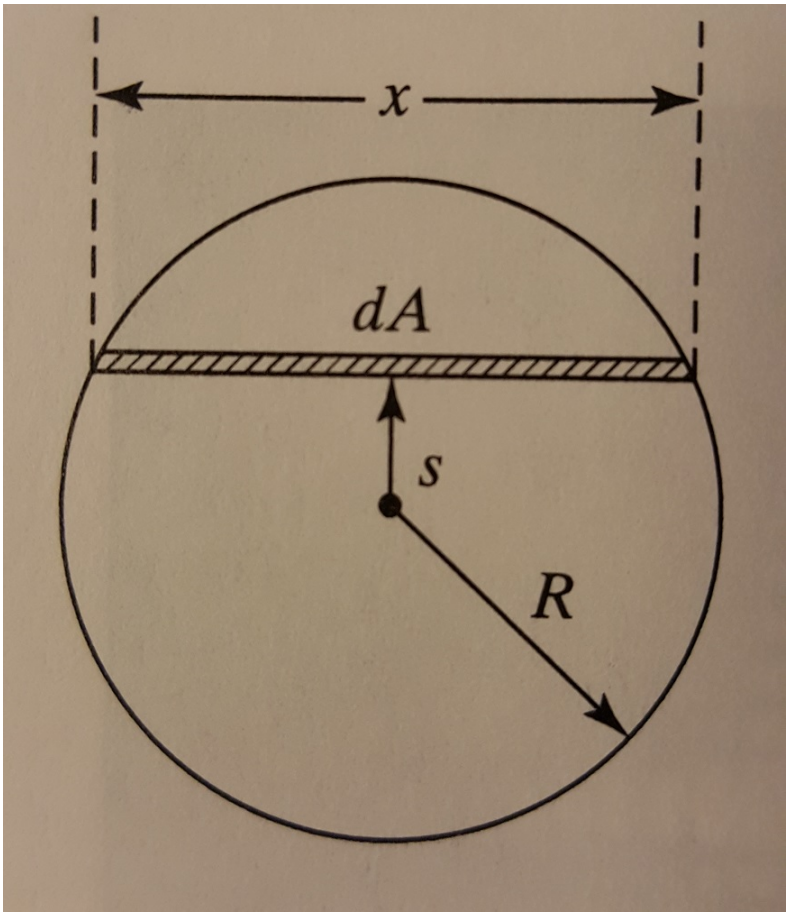

FIG. 3. [**?**]

We can write dA as xds where x is $2\sqrt{R^2 - s^2}$ and R is the aperture radius. Eq 10 then becomes

$$E_p = 2\frac{E_A}{r_0}e^{i(kr_0-\omega t)}\int_{-R}^{R} e^{isksin\theta}\sqrt{R^2 - s^2}ds. \tag{11}$$

Next by making the following substitutions $v = \frac{s}{R}$ and $\gamma = kRsin\theta$, E_P can be rewritten as

$$E_p = 2\frac{E_A R^2}{r_0}e^{i(kr_0-\omega t)}\int_{-1}^{1} e^{i\gamma v}\sqrt{1 - v^2}dv. \tag{12}$$

The integral on the right side is known to be

$$\int_{-1}^{1} e^{i\gamma v}\sqrt{1 - v^2}dv = \frac{\pi J_1(\gamma)}{\gamma}, \tag{13}$$

where $J_1(\gamma)$ is the first-order Bessel function of the first kind and can be represented by the infinite series [2]

$$J_1(\gamma) = \frac{\gamma}{2} - \frac{(\gamma/2)^3}{1^2 * 2} + \frac{(\gamma/2)^5}{1^2 * 2^2 * 3} - . \tag{14}$$

Finally, the irradiance at point P can be written as

$$II_0(\frac{2J_1(\gamma)}{\gamma})^2, \tag{15}$$

where $\gamma = 1/2kDsin\theta$, and I_0 is made up of all the other constants and is the irradiance at the principle maximum [2]. This function is what causes the light dark patterns in the Airy disk. In addition, the irradiance of the 2nd maximum as a ratio of I_0 is 0.0175. Because the principle maxiumum is so much larger than the 2nd, two point sources are considered resolvable when the principle maximum of one source overlaps with the first minimum of the other as in Fig. 5 b. This is known as Rayleigh's Criterion. To create an equation for this criterion, we will need to know when the irradiance first becomes 0 and that distance from the center is the minimum resolvable distance. The irradiance first becomes 0 when $\gamma = 3.832$ so

$$\gamma = (\frac{k}{2})Dsin\theta = 3.832. \tag{16}$$

Using the small angle approximation and multiplying by the distance to the screen being projected onto, f, we get the minimum resolvable distance, d, as

$$d = f\frac{1.22\lambda}{D}, \tag{17}$$

$$d = \frac{1.22\lambda}{2NA}, \tag{18}$$

where d is the minimum resolvable distance, λ is the wavelength of the laser, and NA is the numerical aperture of the lens (memory mass storage). If all other optical instruments are perfect, this is the smallest resolvable distance from a laser. CDs use a laser wavelength of 780 nm along with a numerical aperture of .45 which yields a minimum resolvable distance of 1.06 microns. The actual resolvable distance for the laser in a CD isn't as perfect and is 1.6 microns. To decrease the minimum resolvable distance, and consequently decrease pit size and pitch, the numerical aperture must be increased and the wavelength of light decreased. This is what DVDs do.

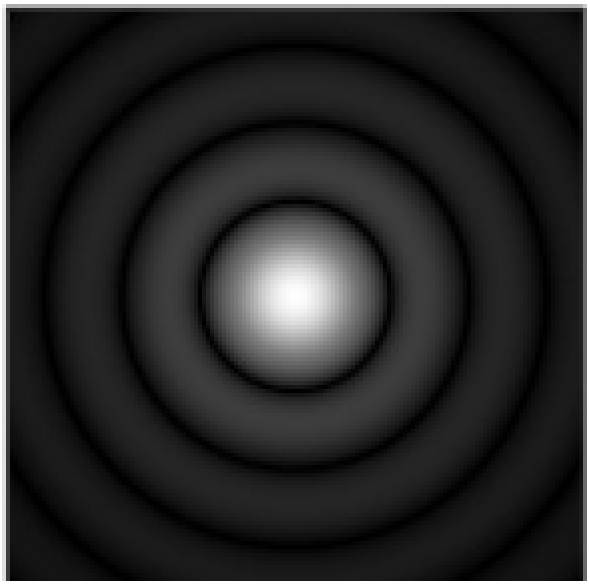

FIG. 4. The airy disk created by a diffraction limited circular aperture lens laser [3].

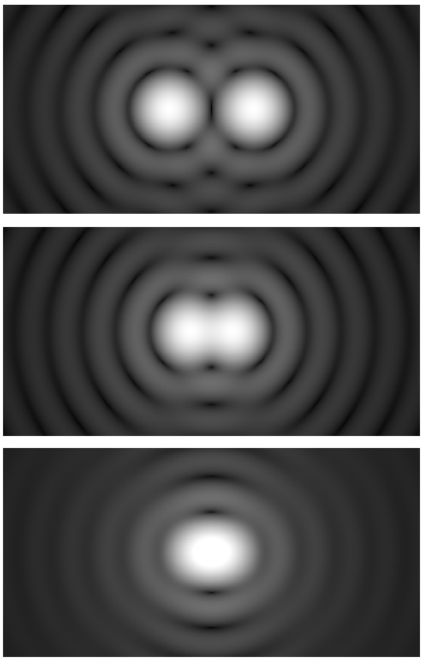

FIG. 5. The top image shows two resolable images, the middle image shows Rayleigh's critereon where the first fringe appears over the max of ther other image, and the bottom image shows them being not resolvable [4].

DVDs use a laser of wavelength 650nm and numerical aperture of .6. This yields a real world minimum resolable distance of 1.1 microns and increases the storage density to 2.2GB/in^2 from

0.9Gb/in^2 of CDs [1]. Another innovation that allowed storage of up to 8.5GB was the ability to have two different recording layer: a semi reflective layer in the middle and a fully reflective layer at the other end. The different layers can be read by changing the focus of the laser. Additionally, instead of having a protective layer on one side of the disk, you can put another disc on it to effectively double the capacity again. However, this requires the manual flipping of the disc when accessing the other half of the data. By being double sided and having a double layer, DVDs can achieve a storage of 17GB [1].

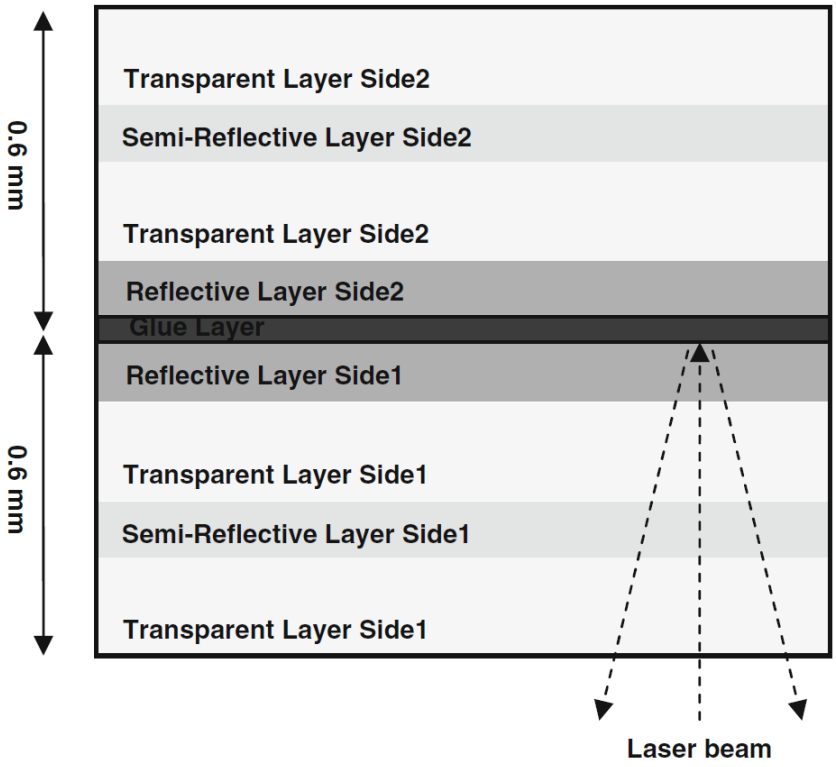

FIG. 6. This is a cross section of a double sided and dual layer DVD. The pit and land information is stored at the two reflective and semi-reflective layers. The layer in which the laser is reading from is selected by focusing the laser to for that layer depth. To access the other side of the DVD, it must be physically removed and placed upside down [1].

A more recent innovation in storage of optical disks was the invention of the Blu-ray disc. Just like how using a smaller wavelength light in DVD compared to the ones in CD allowed the reduction in size of the pits and track size, Blu-ray discs also use a much shorter wavelength light to increase capacity. Blu-ray uses 405nm light which allows both the pit sizes and the track size, space between tracks, to be reduced to 0.15μm and 0.32μm respectively compared to the track size

of 0.74μm for DVDs. This allows for a storage density of 14.7Gb/in^2 to be achieved [1].

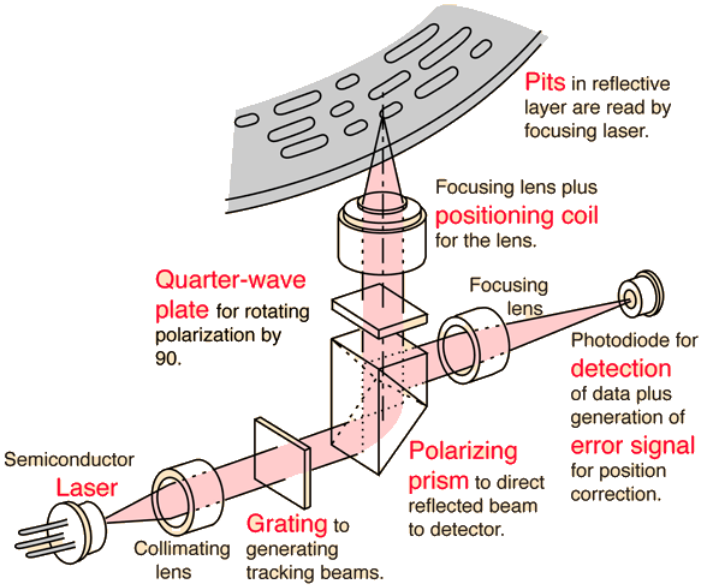

FIG. 7. This is the laser read system set up that corrects for depth errors with the positioning coil and corrects for reading off of the track error [5].

D. Laser Read System

The implentation of the laser system used to read the data from the disk also worth looking into and can be seen in Fig. 7. The first part of this system from the source laser is the grating. The grating creates extra beams known as tracking beams, one on each side of the center beam that is used to read the data. These tracking beams are reflected off of the disk along with the reading beam. Since the tracking beams are symmetric accross the reading beam, any difference in brightness measured in the photodiode is used to re-center the beam along the track.

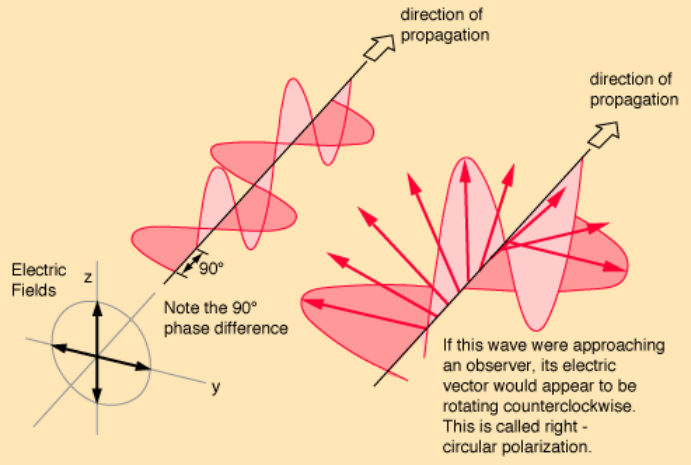

FIG. 8. Circular polarized light on the left with a 90 degree phase difference between the electric field and magnetic field which makes the electric field propagate in circular motion [6].

The next component of the system is the polarizing prism. Before we talk about the prism, we need to talk about polarization in the context of light. Light is an electromagnetic wave meaning it has an electric field component and a magnetic field component that are both orthogonal to the direction of propagation. The polarization of light is the orientation of the electric field. Unpolarized light is light that has random orientations while linearly polarized light is light that has the electric field oscillating in one dimension. A quarter-wave plate is a device that has a fast axis and a slow axis such that when linearly polarized light passes through it, the quarter-wave plate applies a 90 degree phase retardation between the electric field and magnetic field. As seen in Fig. 8, this creates an electric field that moves circularly while propagating.

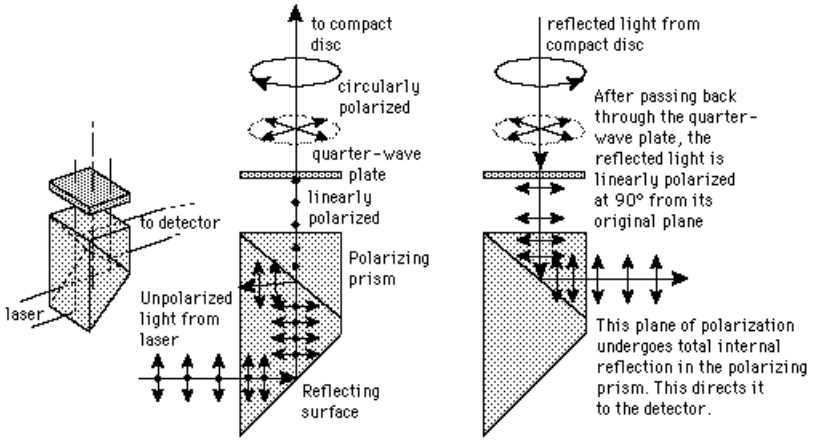

FIG. 9. The polarizing beam splitter with quarter-wave plate that reflects the laser off of the disk and directs it towards the photodiode, not seen [5].

As seen in Fig. 9, the polarizing beam splitter allows one polarization of the unpolarized light through, linearly polarizing it, while reflecting the other polarization.The linearly polarized light then goes through a quarter-wave plate circularly polarizing it. Next the light is reflected off of the disk back to the quarter-wave plate. Since it was reflected, the direction of the circular polarization is flipped, so when it passes back through the quarter-wave plate, it becomes linearly polarized but with an orientation orthogonal to the one it had after it passed through the beam splitter. This means that when it comes back through the beam splitter, it will be reflected towards the photodiode. The last part of the system is the focusing lens and positioning coil. If the laser isn't properly focused onto the surface of the disk, the beam will be larger and the minimum resolvable distance will increase. To account for this, as seen in Fig. 10, the laser light is reflected off the disk and then goes through the focusing cylindrical lens which is then shone on to the 4 component photodiode. Since a cylindrical lens only focuses in one axis, there will only be one distance from the lens to the source image (the disk surface in this case) that will cause the laser beam to be circular. If the image distance is too far or too close, the beam will appear oval shaped and the photodiodes will create an error voltage that drives a coil to reposition the lens so it's properly focused.

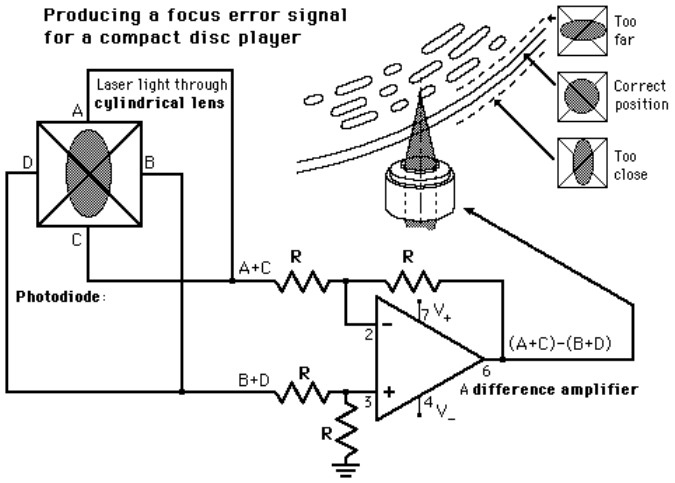

FIG. 10. The positioning coil uses a cylindrical lens that causes the laser's focus point to occur at the disk surface because if it wasn't the reflected image on the diode would not be circular and the positioning coil would move accordingly [5].

III. MAGNETIC STORAGE

A. Background

Another popular technology used to store data is magnetic storage. Though it is available in both tape and disk, I will only talk about disk here. Magnetic storage takes advantage of the ferromagnetic property of certain metals such as iron and cobalt. Ferromagnetic materials are composed of small magnetic domains each with their own magnetic field. An unmagnetized ferromagnet will have these domains in random directions such that the net effect is 0 as seen in Fig. 11. However, if an external magnetic field is applied, these domains will align with the external field and will stay aligned even after the external field is removed as in Fig. 11.

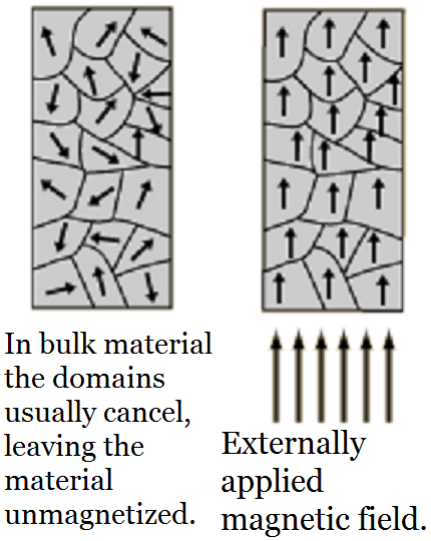

FIG. 11. a) an unmagnetized piece of iron where the domains are randomly oriented. b) the domains are preferentialy aligned down cite http://hyperphysics.phy-astr.gsu.edu/hbase/solids/imgsol/domain.gif [7].)

An important phenomena that ferromagnetic material experience is the hysteresis loop. As seen in Fig.12, when an external magnetic field, B_0, is applied to an unmagnetized piece of metal, a net magnetic field, B exists. However, even when the external magnetic field is removed, a non-zero net magnetic field exists due to the ferromagnetic material aligning.

If we start with an unmagnetized ferromagnetic material with no external magnetic field, point a in Fig. 12, and increase the external magnetic field, B_0, we get to point b where there is a both a nonzero external amgnetic field and total field from the external and material. If the external magnetic field is then decreased to zero, instead of returning to point a, it instead reaches point

c. Here, even though there is no external magnetic field, there is still an internal field from the ferromagnetic material. This phenomenom is what allows magnetic storage to store bits. A bit value of "1" is read when there is a reversal of current measured meaning that the magnetic field between two domains has swapped. A bit value of "0" is read if there is no change in the magnetic field and thus no change in the direction of the current measured.

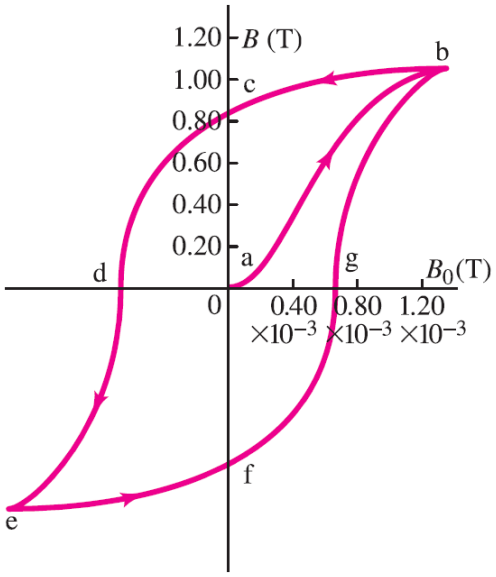

FIG. 12. This displays a hysteresis loop of a ferromagnet. Following from point a, as an external field is applied, the total field reaches point b. As the external field is reduced to 0, since the ferromagnet partially keeps it alignment, a nonzero magnetic field exists even without an external field.

Reading and writing to the material is done by inducing a magnetic field using a head and coil of wire. The mechanics behind this are governed by Maxwell's equations. The two relevant equations are

$$EMF = N\frac{\Delta\phi}{\Delta t}, \tag{19}$$

which is known as Faraday's law of induction and

$$\oint B \cdot dl = \mu_0 I_e nc, \tag{20}$$

which is known as Ampere's law [8]. Faraday's law tells us that a changing magnetic flux will induce an electromotic force (EMF) on the circuit, and Ampere's law tells us how a current through a wire generates a magnetic field.

B. Magnetic Disk Storage

A magnetic disk drive is the most commonly used data storage technology used in consumer products today. First introduced in 1956 by IBM, modern drives consist of mainly of a read/write head and a hard platter with a ferromagnetic layer. Fig. 13 shows how data is recorded and read from the medium. A bit value of 1 is recorded when two adjacent domains have magnetic fields in the opposite direction. This is represented as a spike or dip in the change in magnetic flux curve in Fig. 13. A bit value of 0 is recorded when two domains have magnetic fields is in the same direction. To read record data, a potential is applied in the wire causing a magnetic field to form in the head due to Ampere's law. This generated magnetic field is then used to align the domain to the appropriate direction depending on the bit value desired. To read, a capacitor is attached to the wire. When the read head passes over an intersection between two domains with opposide magnetic fields, bit value 1, there is a spike in the flux which due to Faraday's law applies an EMF onto the wire. This then induces a current which builds up charge on the capacitor. The presence of charge on the capacitor signifies a bit value of 1 while no charge is 0.

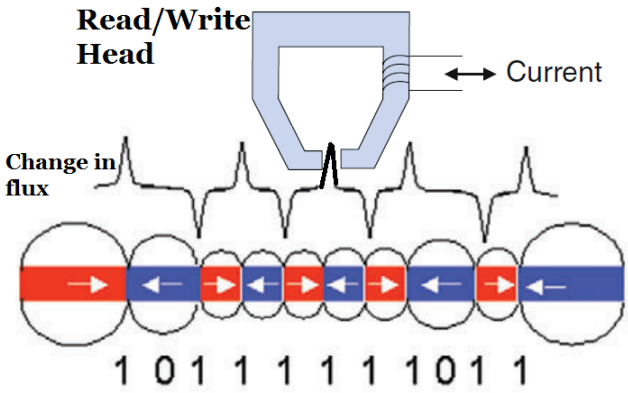

FIG. 13.

Originally, the head was used for both reading and writing but compromises had to be made. The read head wants a thicker gap to be better able to penetrate the medium while the write head wants a smaller gap to get better accuracy. As domains got smaller to increase density, it became more and more difficult for the read/write head to be able to properly read the intended domain due to the weaker magnetic fields sustained by the smaller domains. The discovery of the giant magnetoresistive (GMR) effect in the late 1980s brought in a new type of read head that could read weaker magnetic fields [9]. Magnetoresistance (MR) is the property of a material to change

its resistance due to an external magnetic field [9]. However, that was limited to only a 5

The heart of the GMR can be seen in Fig. 14 where a nonmagnetic layer is sandwhiched by two ferromagnetic layers. One of the ferromagnetic layers is placed next to an antiferromagnetic material, not seen in Fig. 14, that keeps it magnetic field aligned in one direction. The other ferromagnetic layer aligns itself to whatever field it reads from the hard drive. In addition, electrons are traveling through the layers horizontally. If the spin state of the electron is antiparallel to the magnetic field, it scatters and slows down. The probability of scattering depends on the number of quantum states for the electron to scatter into; since there are more states for electrons to scatter into when they are antiparallel, it occurs more [9]. If the two ferromagnetic layers are parallel, one state of the electron spin will pass through without scattering meaning normal resistance, but if the two ferromagnetic layers become antiparallel, both states of the electron spin will scatter in one of the layers causing the resistance to decrease.

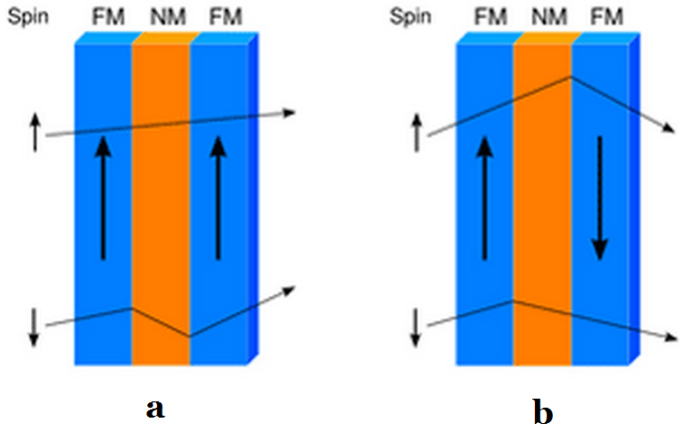

FIG. 14. a) The two ferromagnetic layers are aligned causing the up spin electron to flow with out scattering and thus low resistance. b) The two ferromagnetic layers are antiparallel so both electron spins scatter in one layer causing the resistance to increase [9].

The innovation of GMR read head and a dedicated write head allowed the domain bits to shrink even more. However, a new issue began to arise. The ability to retain stored data in the material is true as long as

$$\frac{K_u V}{k_B T} \geq 60, \tag{21}$$

where K_u is the magnetic anisotropy, V is the volume, k_B is boltzman's constant, and T is temperature. The magnetic anisotropy is proportional to how well it can keep its aligned magnetic field.

This is because after you've aligned all the easy axis of all the domains together, the easier the easy axis is, the harder it is to try to get the magnetic field to align in a different axis. The numerator of the left side represents the energy barrier that must be overcome for the ferromagnetic material to lost its magnetic field. As the domains decrease in size to increase density, the volume decreases and so does the energy barrier. If the volume of the domain decreases to the point where Equation 21 isn't true, then the domain may randomly switch its field direction which would "flip bits"; this is known as the super paramagnetic limit. This creates a limit to the storage density that longitudinal recording, magnetic field parallel to surface of hard drive, can attain [10].

To surpass this limit, the orientation of the recorded magnetic field changed from longitudinal recording to perpendicular recording as seen in Fig. 15. In addition to the change in orientation, the writing head was changed to a mono-pole head and a highly malleable soft underlayer(SUL) was added below the recording material. The permeability of the SUL allowed it to act as a mirror to the write head and directs the magnetic flux to the collector head. This greatly increased the induced magnetic field used to write data**??**. The stronger magnetic field means that materials with higher anisotropy values could be used that were not feasible for the weaker magnetic field in longitudinal recording. Because the anisotropy constant can be increased, the volume of the domain can be further decreased further without worrying about thermal fluctuations. With all the above innovations, current HDDs can achieve a storage density of around 500Gb/in^2, but that continually seems to be increasing.

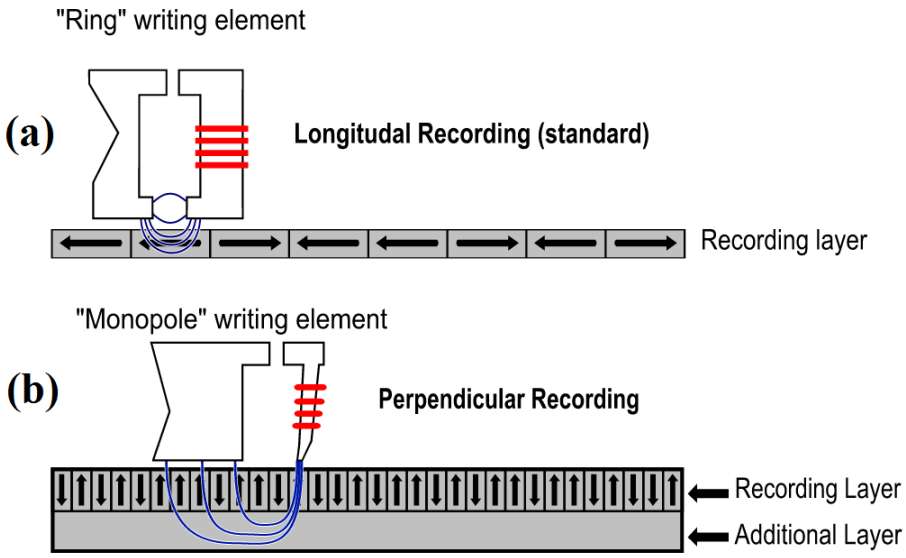

FIG. 15. Perpendicular recording due to the presence of the additional layer, soft underlayer, a much larger magnetic field can be applied to the recording layer. This allows us to use material with higher anisotropy meaning the domain size can be reduced and density increased.

However, this too will have a limit at a smaller domain size where thermal fluctuations will be an issue again. One method still currently in development is called heat-assisted magnetic recording (HAMR). This method is based on the fact that all ferromagnetic materials have a curie temperature where the internal magnetic field the material had is lost and aligns itself to any external magnetic field. In HAMR, when a bit needs to get written to, a laser is used to heat up the desired domain to the curie temperature allowing the magnetic field generated by the write head to align the domain, and then allowing the material to cool back down. This means that the recording material can be changed to one with a much higher anisotropy value allowing the domain sizes to decrease even more, further increasing the storage density to around 1TB/in^2 [11].

IV. FLASH MEMORY

Flash memory is has been becoming more and more popular recently as a replacement for hard drives due to their speed. They are already commonly used in flash drives but are becoming more and more popular. One of the most, if not the most, important device in flash memory and modern computing and as a whole is the transistor. A transistor is able take in a circuit input and act as a switch turning another circuits on or off without any moving parts. This ability to switch off and

on a circuit will prove crucial in saving bits of memory.

A. Semiconductors

To explain how transistors work, first we need to explain what a semiconductor is since a transistor is a semiconductor device. A bandgap diagram is useful in describing semiconductors. Band structure represents the specific energy levels that electrons can occupy due to quantum mechanics. However, only the valence band and the conduction band are relevant when discussing conductivity. The valence band is the highest energy levels that electrons can occupy at absolute zero temperature while the conduction band is where electrons can move freely. As seen in Fig. 16, the conductor is any material that doesn't have a bandgap between the valence band and conduction band meaning all the electrons can move freely. An insulator is any material whose conduction band is separated by a large band gap from its valence band. This means that the valence electrons can not easily move to its conduction band and thus are not free too move. Lastly, a semiconductor is any material whose conduction band is separated from the valence band by a band gap small enough that thermal fluctuations can excite electrons to the conduction band [8].

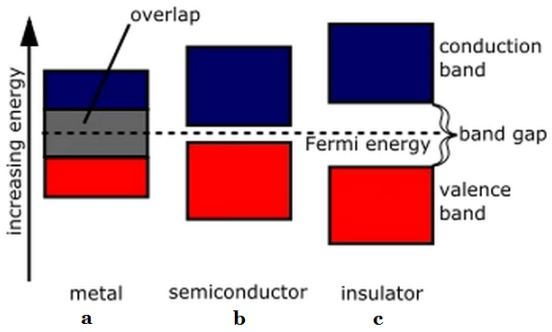

FIG. 16. a) Metals/conductors are materials which have their valence electrons are free to move in the overlapping conduction band. Insulators, c, are materials which have their valence electrons are unable to move in their valence band and is unable to easily access the conduction band. Semiconductors, b, are materials which have their electrons unable to move in their valence band, but is easily exicted into the conduction band where they can carry current.

Most commercial applications of semiconductors use impurity semiconductors. Impurity semiconductors are materials that gain different properties than pure semiconductors by adding an impurity, dopant, to the material. For example, say we have silicon with 4 valence electrons where

each valence electron is shared with another silicon atom to create tetrahedron. Then if we replace one of those silicon atoms with a phosphorus atom with 5 valence electrons, 4 of the 5 valence electrons will be shared with the 4 nearby silicon atoms, but the last one is loosly bound to the phosphorus. At room temperature, thermal energy is enough to release it from this loosly bound state and it becomes a conduction electron. This is called n-type doping. If instead of adding an element with 5 valence electrons and instead added one with only 3, we would then have one less electron than its pure state which is represented as a hole.This is called p-type doping [8].

A simple example of a semiconductor device is the pn junction diode. This device is created when a p-type semiconductor is comes in contact with an n-type semiconductor. Since p-type materials have a higher concentration of holes than the n-type, holes diffuse across to the n-type material while electrons diffuse the other way; this is the diffusion current. The holes and electrons that diffuse then recombine in the n-type and p-type material respectively. This creates a depeltion zone at the intersection of the two materials. The diffusion and recombination also creates a net positive charge on in the p-type material and a net negative charge in the p type material creating an Electric field towards the p-type. The diode is in equilibrium when the drift current and diffusion current are equal.

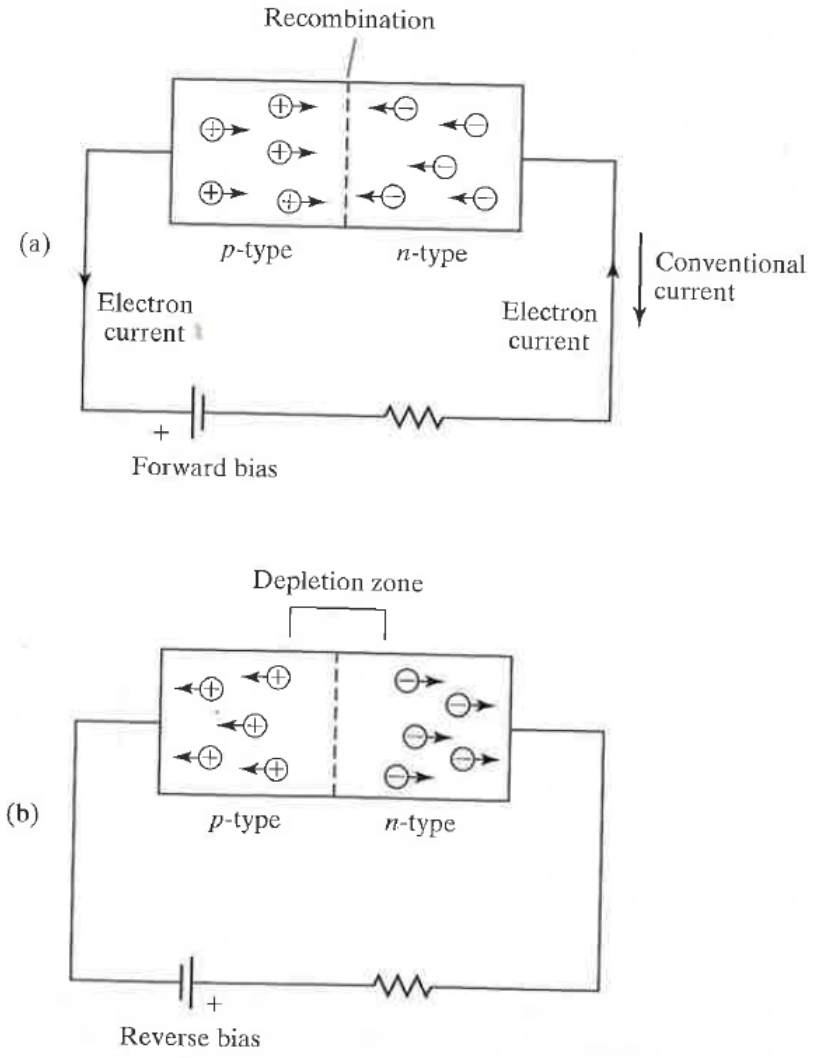

FIG. 17. This is a diagram of a pn junction diode with current in different directions. In the clockwise current in a, the holes and electrons are constantly refreshed due to the applied potential, and the current can continuously flow. In the counterclockwise current, b, the applied potential increases the depeletion zone until the local electric field cancels out the applied potential and stops the flow of current.

The special property of a PN junction is that current is only allowed to flow in one direction. Fig. 17 shows what happens when a clockwise current is applied and when a counterclockwise current is applied. In Fig. 17a, the clockwise current pushes holes from the p type and electrons from the n type towards the interface where they combine and produce heat. So the diode loses electrons and holes at the interface due to recombination, however the current pulls electrons from the p type resupplying the p type with holes. The converse happens in the n type so a current can flow indefinitely. This is known as forward bias. In Fig. 17b the counter-clockwise current directs

the holes in the p type and the electrons in the n type flow away from the interface. The flow of electrons and holes away from the interface continues until the local electric field produced is strong enough to oppose the applied potential, causing the current to stop; this is known as reverse bias [12]

B. Transistors

So now that we know what p-type and n-type doped semiconductors and pn junction diodes are, we can now talk about transistors. While there are many different types of semi conductors, the one that flash memory uses is the metal oxide semiconductor field effect transistor (MOSFET). A MOSFET, as seen in Fig. 18a consists of a p-type silicon with two n-type regions, one on each side as a source and drain, and a metal gate separated from the semiconductor by an insulating oxide layer (typically SiO_2) in the middle. When no voltage is applied to the gate as seen in a, the MOSFET acts as two pn diodes, and so any voltage difference between the source and drain will reverse bias one of the two pn diodes so no current will flow. However, if a positive voltage is applied to the gate, the positive charge attracts electrons and repels holes as in Fig. 18b. If the voltage applied to the gate is above a certain threshold of around 1V, a thin layer of electrons will be formed under the oxide. Also known as the inversion layer, this creates a conducting channel between the source and drain, and the p-type silicon acts as an n-type material at the oxide layer due to the extra electrons allowing current to flow between the source and drain. This change created by the field from the gate is why it is called a field effect transistor. One advantage MOSFETS have over other transistors is that the MOSFET requires nearly no current, so the input power is much less and there will be less issues with self-heating, which can be an issue since they're so tightly packed together[7].

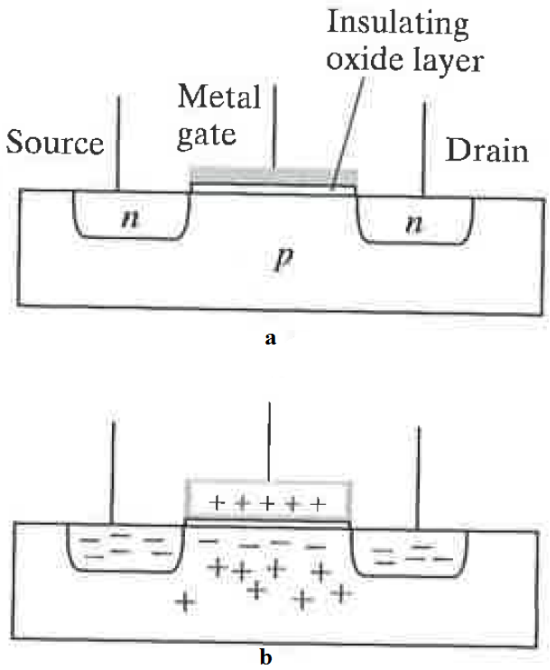

FIG. 18. a) is a normal MOSFET without any applied charge and thus acts an npn semiconductor and doesn't let current flow. b) The positive charge on the metal gate cause electrons to build up along the oxide layer/semiconductor interface between the two n drops. This causes the semiconductor to turn into a n type semiconductor which allows current to flow.

What is actually used to store bits is called the floating gate transistor. It is the same as the MOSFET except instead of only having one gate above the insulating oxide, you have another layer of insulating oxide on the other side of the gate followed by another gate called the control gate. So now the middle gate, also known as the floating gate, is electrically isolated due to the oxide layer on each side. Because of this, the only way to add and remove charge is through Fowler-Nordheim tunneling. This means that when positive charge is stored on this gate, current will pass through and send a 1 and a 0 otherwise. Since the charges on the floating gate can not leave it without tunneling, that means that even when the transistor isn't powered, it will still store its memory.

One method used to increase storage density is to change from using single level cells (SLC) to multi level cells (MLC) and three level cells (TCL). SLC is the standard floating gate where it either outputs a 1 or 0 by dividing the gate into two possible states, current or no current. However, the MLC stores 2 bits instead of one by dividing the gate into 4 possible states. While this has the benefit of doubling the storage density, it allows more room for error since the states are smaller and more difficult to distinguish. The oxide layer that was being used previously degraded

after certain number of uses, and this degradation would become more visible on MLC drives first because they have to be able to disntinguish states more than SLC. However, advancements in both the oxide layer and also dynamic mapping of memory such that no one gate is used more than the other means that this have become less of an issue.

V. FUTURE TECHNOLOGY

While these three types of storage already exist for the most part, we need to look into other techniques to store data as the world creates more and more of it. Two possible techniques are holographic data storage and probe based storage.

A. Holographic Data Storage

Holographic data storage (HDS) has the potential to store over 1Tb/in^2. In the most basic sense, holography is the method of capturing and recreating optical information through interference patterns, and a hologram is a recording of an interference pattern. Fig. 19 shows how to record a hologram; first a coherent light source is split into a reference and object beam using a beam splitter. The object beam is reflected off the object being recorded on to the photographic plate. The reference beam, also being shone on to the photographic plate, interferences with the object beam and that interference pattern is recorded on the plate as a hologram. Destructive interference appears as dark spots and constructive interference appears as bright spots.

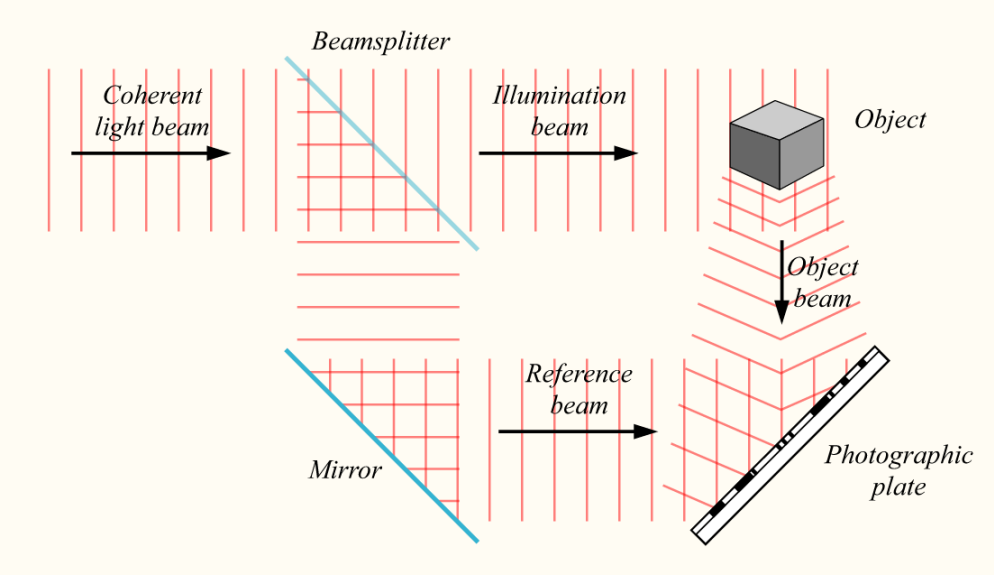

FIG. 19. This is an example of a transmission hologram where a coherent souce of light is split into a reference beam and an object beam. The two beams eventually create an interference pattern at the photographic plate creating a hologram [13].

To recreate the image from the hologram, the reference beam in Fig. 20 is shone back at the hologram at the same angle and by looking at where the object should be through the hologram, a virtual image can be seen of the object. This is due to the reversibility of light which states that light will take the same path if the direction it travels is reversed. So when we shine the reference beam through the hologram and view it with our eyes, our eyes trace the light back the same path it took when the hologram was created with the object beam and see a virtual image.

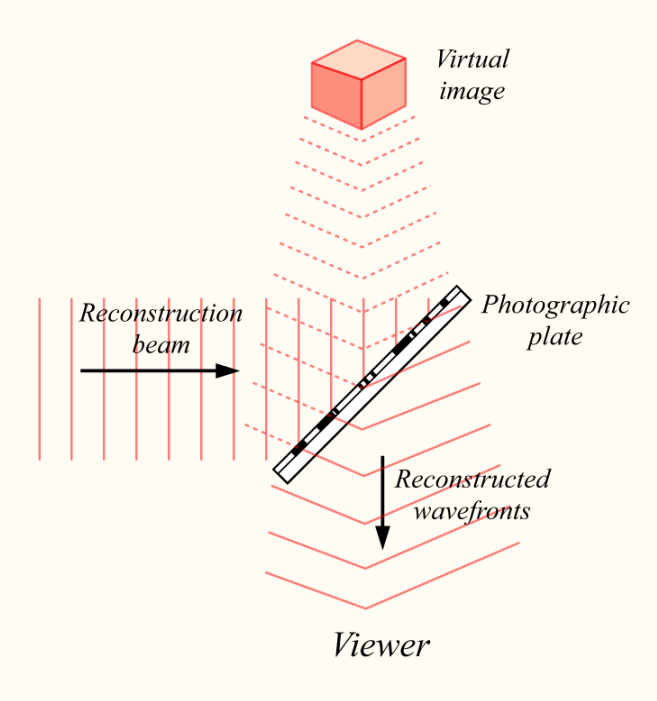

FIG. 20. To recreate the virtual image, the reference beam is shined on the hologram at the same angle, and by looking through the hologram to where the object should be, a virtual image can be seen [13].

Another way to look at this would be quantitatively. First, the reference beam can then be represented a complex electric field,

$$E_R = re^{i(\omega t+\phi)}, \tag{22}$$

where r is the amplitude of the beam and is assummed to be constant over the surface of the film due to the plane wavefront of the reference beam (wavefront is the collection of points that have the same phase and thus same amplitude) [2]. ω is simply the angular frequency, and ϕ is the phase angle that relates to the tilt of the photographic plate relative to the reference beam and can be described by

$$\phi = (\frac{2\pi}{\lambda})\Delta = (\frac{2\pi}{\lambda})xsin\alpha \tag{23}$$

when the top border of the beam hits the plate at x = 0. As seen in Fig. 21,δ is the extra distance traveled by certain parts of the wavefront depending on the angle of the beam relative to the normal, α.

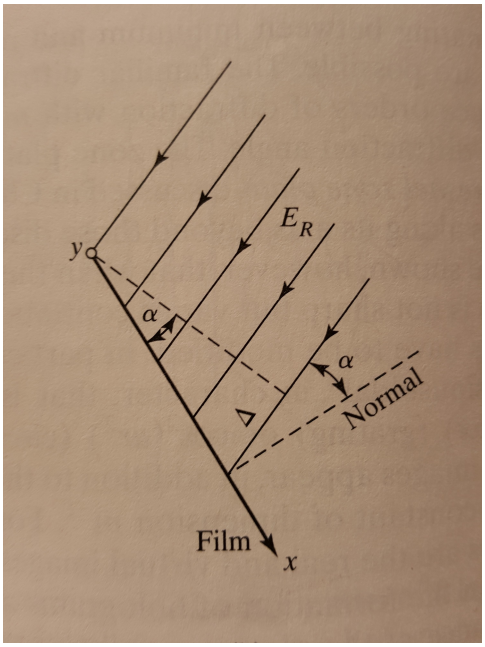

FIG. 21. The intersection of the reference beam and the photographic film [2].

Similary, the subject beam can be represented by,

$$E_S = re^{i(\omega t+\theta)}, \tag{24}$$

where s is the ampltiude of the reflected light off the object, and θ is analogous to ϕ except that it is more complicated due to variations in the phase of the light because it is reflected from different parts of the object [2].

The resultant electric field at the plate is then given by

$$E_P = E_R + E_S. \tag{25}$$

The quantitiy that describes the hologram is known as scaled irradiance which is defined as the magnitude squared of the electric field. Thus, the scaled irradiance at the plate, is given by

$$I_P = |E_P|^2 = (E_R + E_S)(E_R^* + E_S^*). \tag{26}$$

By multiplying the binomials and subbing in for E_R and E_S from Equations 22 and 24, respectively, I_P can be simplified to [2]

$$\begin{aligned} I_P &= r^2 + s^2 + E_S E_R^* + E_R E_S^* \\ I_P &= r^2 + s^2 + rse^{i(\theta-\phi)} + rse^{-i(\theta-\phi)}. \end{aligned} \tag{27}$$

Then by shining the reference beam back through the hologram, we can recreate the image of the object. This can be expressed as

$$E_H = I_F E_R = (r^2 + s^2)E_R + r^2 s e^{i(\omega+\theta)} + r^2 e^{i(2\phi)} e^{i(\omega t-\theta)}. \tag{28}$$

The first term of E_H represents the reference beam that has been amplitude-modulated but not phase-modulated, so it passes straight through the hologram. The second term represents the subject beam that has been amplitude-modulated by a factor of r^2. This term represents the reconstructed wavefront from the subjectand strikes the plate at an angle α. This is the virtual image seen typically viewed and is a virtual image since no light is present at the image, rather it is created due to our eye reversing the light back to the image. The last term represents the an amplitude and phase-modulated subject beam. Since it is phase modulated as well, it appears on the opposite side of the hologram and is a real image.

Moving on, holograms can be created in both thick and thin regimes. The thin regime is unsuitable for high density storage because thin holograms like those written on photographic film confine their diffractive interaction to a single plane and can't be used for multiplexing [13]. The criterea used quantify the thickness or thinness of a hologram is known as the Q parameter. The Q parameter is given by

$$Q = \frac{2\pi\lambda L}{n_0 \Lambda^2}, \tag{29}$$

where λ is the wavelength of the light, L is the thickness of the recording layer, n_0 is the index of refraction of the medium, and Λ is the grating period. A hologram is considered in the thin regine if $Q < 1$ and is considered in the thick regime if $Q > 1$. The thickness of the medium can range from 500μm to a few millimeters [13].

To store digital data in holograms, a spatial light modulator (SLM) is needed. The SLM converts the 1s and 0s of the digital data into light and dark pixels respectively onto a 2D image called a page. The number of bits that can be stored per SLM image can be over one million, but this varys with the SLM's pixel count. So to record digital data, instead of reflecting the object beam off of the object, it is shone through the SLM then onto the storage medium as seen in Fig. 22.

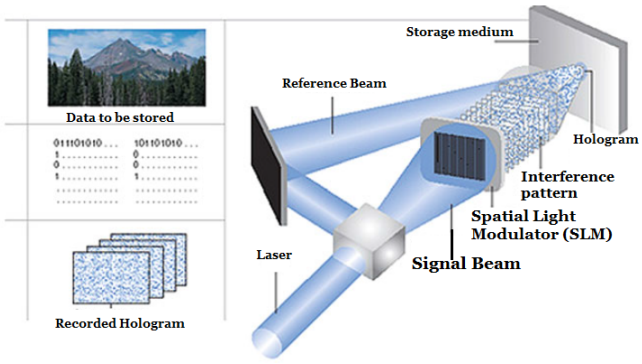

FIG. 22. This implementation of holographic storage uses the same technique as a transmission hologram except instead of reflecting the object beam off of the object, it is shone through the SLM which contains the bit information. Since the SLM page contains many bits at once, it can write more than one bit at once unlike magnetic drives.

Reading the hologram is the same as it would be for a thin hologram. The reference beam is shone at the hologram, but now there is a detector array to convert the SLM image of dark and bright spots into 1s and 0s for the computer. Since a page of bits can be read/written all at once instead of bit by bit, HDS has the ability to have much greater read write speeds over other competing technologies such as magnetic data storage.

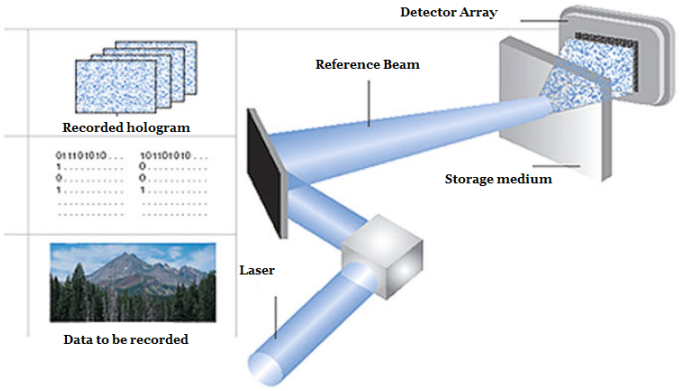

FIG. 23. Reading information is the same as recreating a hologram except now the virtual image is shone onto a detector array which converts the dark and bright spots into 1s and 0s.

The reason why HDS can be so dense is because of its ability to superimpose holograms in the same volume; this is known as multiplexing. HDS offers many different methods for multiplexing. An important parameter in multiplexing is called the diffraction efficiency, η, which is the ratio of

the diffracted optical power divided by the incident optical power and is expressed as

$$\eta = sin^2(\sqrt{\nu^2 + \xi^2})/(1 + \frac{\xi^2}{\nu^2}) \tag{30}$$

where ν and ξ are defined as

$$\nu = \frac{\pi n_1 L}{\lambda(cos^2\theta - \frac{Kcos\phi}{\beta})^{1/2}} \tag{31}$$

and

$$\xi = \frac{[\Delta\theta K sin(\phi - \theta_o) - \Delta\lambda K^2/4\pi n_0]d}{2(cos\theta - \frac{Kcos\phi}{\beta})}. \tag{32}$$

In above equations are for a pure index grating in a transmission hologram where n_1 is the index perturbation, L is the thickness of the recording layer, θ is the angle of the reference beam outside the media, K is the grating number, n_o is the bulk index of the material, ϕ is the grating slant angle inside the material, β is $2\pi(epsilon_o)^(1/2)/\lambda$, where $epsilon_o$ is the bulk dielectric constant, $\Delta\Lambda$ is the discrepancy from the Bragg condition (constructive interference from reflection off of grating) for wavelength, and $\Delta\theta$ is the discrepancy from the Bragg condition for angle. A diffraction efficiency of 0

One category of methods is known as Bragg-Based techniques which rely on the Bragg effect/condition. Since the angle of the reference beam must match exactly the angle it had when recording the hologram, that means that a small change in the angle when recreating the image will cause the image to not appear sharply or at all. We can use this sensitivity by having one page of data from the SLM stored at one angle and another page in the same volume by changing the angle of of entry. More specifically, the spacing between the diffraction max and the first null in the diffraction efficiency, angular selectivity, is given by

$$\Delta\theta = \frac{\lambda cost\theta_s}{Lsin(\theta_r + \theta_s)}, \tag{33}$$

where θ_r and θ_s are the reference and signal(object) angles and L is the media thickness. To minimize noise and interference between different pages, each page would have its Bragg peak at the null position in the diffraction efficiency of every other page by changing the reference angle in by multiples of the angular selectivity. A similar method to add multiplexing is to vary the wavelength of the the source and reference beams. This method is limited due to the small tuning range of lasers. Yet another method is to vary the point of entry of the beams into the medium. While there are more methods of multiplexing, the true benefit of multiplexing occurs when combining multiple multiplexing techniques at once. For example, hybrid wavelength and angular multiplexing systems have been tested before [1].

VI. CONCLUSION

Digital data storage allows us to store our digital information in a variety of different methods. We're already able to store data at incredible densities or speeds, but as more and more information is being saved digitally we will need more and more dense and efficient ways to store our data. Since only two states are needed to store memory, it is very easy to find techniques to store memory like punch cards. However, it is quite difficult to find technique that are dense and efficient. As we've seen, there is continued improvements being made to current data storage techniques such as Blu-Ray discs and the change to perpendicular recording in hard drives. However, these are only short term solutions to our storage needs. Both holographic storage and probe storage are new techniques that each have the potential to store over 1 Tb/in^2, but they each have their own issues that currently keep them in the development stage. As our data needs keep growing, the technology needed to support that will continue to evolve as well.

---

[1] G. Campardo, F. Tiziani and M. Iaculo, *Memory Mass Storage* (Springer Berlin Heidelberg, 2011), URL: `https://books.google.com/books?id=5hxzxjTsTEQC`

ANNOTATION: This book provided information on a significant portion of my paper. It had information on magnetic storage, flash, but I used it mainly for optical disk.

[2] F. L. Pedrotti, L. S. Pedrotti and L. M. Pedrotti, *Introduction to Optics* (Pearson Addison Wesley, 2007)

ANNOTATION: This was used heavily to explain the quantitative calculations in holograms.

[3] *Airy Disk Image*
URL: `http://cdn.cambridgeincolour.com/images/tutorials/airydisk-rings.jpg,ANNOTE={Providedanimageofanairydisk.}`

[4] *Rayleigh's Critereon*
URL: `http://upload.wikimedia.org/wikipedia/commons/a/ae/Airy\_disk\_spacing\_near\_Rayleigh\_criterion.png`

ANNOTATION: Provided an image of raylight's critereon with airy disks.

[5] *Compact Disc Audio*,
URL: `http://hyperphysics.phy-astr.gsu.edu/hbase/audio/cdplay.html\#c1`

ANNOTATION: I used this extensively to learn the different checks and balances in the laser read system.

[6] *Circularly Polarized Light image*,
URL: http://hyperphysics.phy-astr.gsu.edu/hbase/phyopt/imgpho/polcir.gif

ANNOTATION: Provided an image of circularly polarized light.

[7] J. R. Taylor, M. A. Dubson and C. D. Zafiratos, *Modern physics for scientists and engineers* (Prentice-Hall, 2004)

ANNOTATION: I mainly used this to look up pictures of domains and definitions of domains.

[8] E. M. Purcell and D. J. Morin, *Electricity and Magnetism*, 3rd ed. (Cambridge Univeristy Press, 2013)

[9] *Giant magnetoresistance* (2013),
URL: http://physicscentral.com/explore/action/magnetoresistance.cfm

ANNOTATION: I used for both a diagram of a GMR and also to explain the reason why electron spin is effected by magnetic field..

[10] S. N. Piramanayagam, *Perpendicular recording media for hard disk drives*, Journal of Applied Physics **102** (2007) (1), p. 11301

ANNOTATION: This paper provided an overview of how perpendicular magnetic storage is implemented and briefly touched on the super paramagnetism limit.

[11] L. Pan and D. B. Bogy, *Data storage: Heat-assisted magnetic recording*, Nature Photonics **3** (2009) (4), pp. 189–190

ANNOTATION: This provided an overview of HAMR technology specifically why it is able to achieve such high density rates.

[12] K. Stowe, *An Introduction to Thermodynamics and Statistical Mechanics* (Cambridge Univeristy Press, 2007)

ANNOTATION: This was used mainly to learn what happened when a p type and n type semiconductor are brought together.

[13] K. Curtis, L. Dhar, A. Hill, W. Wilson and M. Ayres, *Holographic data storage* (Wiley Online Library, 2010)

ANNOTATION: This book provided an overiew of how holograms are created and also how higher densities can be achived using multiplexing.