ENGIN 112: Notes for Module 11 – Memory and Storage

While you may already be familiar with many of the aspects of memory and storage, sometimes when you look at a problem more closely, you realize that there is still much to learn. Further, when studying any subject, it is necessary to agree upon a definition of terms, such that a meaningful technical exchange of information can be had when discussing the topic.

Among the terms that are used in today's discussion about memory and storage, are the following:

- **Bit** – a combination of the terms "binary digit", or "bit". It doesn't hurt that in English, the term "bit" also means a tiny amount, as in "I'd like to have a bit of cheese please!"
- **Byte** – eight bits, in essence leading to $2^8 = 256$ possible combinations.
- **Nibble** – one half of a Byte, or 4-bits.
- **Kilobyte** – the definition of kilobyte can vary. It technical terms, it can mean 1000 bytes, or 1024 (2^{10}) bytes.
- **Kibibyte** – a more technically exact term, indicating 2^{10} bytes. Similar terminology can be extended to a Mebibyte, Gibibyte, Tebibyte, Pebibyte, and Exbibyte!

1. Mechanical methods of storing information

Storing information is the essence of memory and storage. The oldest forms of storing information is using the written word, originally carried out by hand, but then by mechanized means, such as a printing press. More modern forms of replicating and storing the written word are things like the photo copier or the fax machine.

Further, not all information has to be stored as words. For example, sounds of course can also be stored. These days they are held in digital file formats such as MP3. In the first days of recording sound, they were recorded on wax cylinders, and then flat vinyl records. Some forms of sound recordings were not held as sounds at all, but rather rolls of paper with holes in them that could trigger the keys of a piano to play in a predefined sequence (a predecessor to the punch-cards of early computing).

Thus far, all of the methods of recording information thus far have been mechanical in nature. That is, the information is recording by changing the physical characteristics of the medium. This applies to the photocopier as well, which uses static electricity as a method for mechanically depositing toner onto paper.

2. Magnetic methods of storing information

While such a process is useful for very long-term storage (e.g. consider the Dead Sea scrolls or the Rosetta stone!), it is less useful for copying information and communications. For that we rely on magneto-electrical methods for storing information. For the magnetic component of storing information, electrical currents can be used to both create and sense the presence of a

magnetic field (see Oersted's experiments and Ampere's Law). Once a magnetic field is created, it can be stored on media such as a flexible tape (e.g. real-to-real and cassette-tape recording devices), or a solid disk, such as a hard-drive platter, that can support a material that can be magnetized.

You will note that the mechanisms of these more modern forms of memory storage mimic their earlier mechanical counterparts. That is, the magnetic tape is similar to the rolls of paper that were used to make player-pianos work, and that magnetic disks, or hard drives, are very similar to how vinyl records work.

Further, early forms of computer memory used to be constructed out of magnetic toroids (donut-like shapes) that were called **magnetic cores**. The magnetic state of such cores (clockwise vs. counterclockwise magnetic fields) could be changed by running strong currents through the cores and then later read using weaker currents. Since early computers had very limited memory, and the state of the computer was basically a result of the configuration of this core memory, as in a **state machine**, if an error occurred in the computation, the entire contents of the memory could be mechanically recorded to paper (or other such medium), for later analysis, in a process known as a **core dump**, a term in use even today.

In modern times, the relatively slow, but semi-permanent form of magnetic storage for computation has given way to the more transient forms of electrical storage of information. Early forms of electrical methods of memory, was in the capacitive storage of electrons which would have to be refreshed on a periodic basis (e.g. Dynamic Random Access Memory, or **DRAM**). Such storage types are much faster than magnetic storage, but are less permanent.

3. Flash Memory and MOSFETS.

More recently, the transient nature of storing charges capacitively has been vastly improved, through **Flash Memory**. Flash memory is based on the use of MOSFETs, or **Metal-Oxide-Semiconductor Field Effect Transistors**, to store information. Here, rather than just using the control-gate to affect the flow of electrons in the semiconductor, a *floating gate* is used between the control gate and the semiconductor channel. This floating gate is isolated from both the control gate and the channel by an insulating oxide layer, and it is this floating gate that is used for storing the electrons that make this non-volatile form of electrically-based data storage.

An interesting aspect of flash memory is that it incorporates forms of both the mechanical forms of storage (a semi-permanent physical storing of electrons on a floating gate) and electrical, in the control of signals passing through a semiconductor channel. It is because of this mechanical aspect of flash memory that limits the number of read/write cycles that the memory can withstand; an aspect of flash memory that has vastly improved in recent years with technological advances and a degree of redundancy in on-board storage resources.

4. Cloud computing and the location of memory

These days we are experiencing a revolution in the way that information is stored and that computation is executed. That is, rather than bringing resources to a local setting where computations are performed, we are instead bringing communications and processing to the location of the data. This is the process known as **cloud computing**, which allows for vast amounts of data to be processed using relatively simple algorithms. Cloud computing helps avoid the need for transferring data and allows for data to be reduced in volume, whereas only the interpreted results of the data need to be communicated, in the form of plots, statistics or derived calculations.

In class (and in the homework), we have studied one type of cloud storage, via Amazon Web Services, or AWS. AWS has various levels of storage types, some were storage resources are immediately available via the cloud (S3, or "Simple Storage Service"), or various other forms of long-term or large-volume storage, known by terms such as Glacier, Snowball, and Snowmobile. Such forms of storage and retrieval can be compared against physical forms of storage that we are more familiar with, such as hard drives, solid state drives, and USB drives (the latter two of these are forms of flash memory).

The benefits and drawbacks of these different storage types can be compared via some basic metrics. These are related to:

1.) the cost of storage ($/GB),
2.) the time to access the storage, and
3.) the time to upload/download memory to a local system where computation can be applied by the user.

In all, the choice of doing the computation on the cloud, versus storing data and processing it locally, requires an awareness of these resources and the tradeoffs that can be measured using the metrics detailed above and the ultimate needs and resources of the end user.