Lecture 9 - Scalability Harvard Web Development Notes:

- A web host should have sftp(secure file transfer protocol) in contrast to ftp(secure file transfer protocol) as sftp is encrypted.
  - User data is important to be encrypted. Ex. Username and password.
- A web host can also be shared where u are sharing resources with a bunch of other users.
- A VPS (virtual private server) is similar to a shared web host, however each user is guaranteed a certain amount of resources thanks to a virtual machine running on the web host server.
  - The hypervisor on the virtual machine is responsible for splitting into virtual compartments
- Most easy way to to scale is called vertical scaling.
  - Increase Resources such as RAM, CPU capacity.
  - There is a ceiling to this due to financial and technological constraints.
- Horizontal scaling is using multiple slower machines/ servers instead of relying on one really good machine.
- Load balancer is used to distribute internet traffic to backend servers in horizontal scaling.
  - The load balancer acts as the public ip and the rest of the servers can operate on a private ip.
  - The load balancer can decide to send a request to a particular server based on:
    1. The least busy server. Allows for optimizing performance.

    2. You could also have dedicated servers for specific uses. One for images, One for text etc.

- The load balancer can also be a DNS server wherein it returns the server addresses 1 … n on each request from a client. This approach is known as round robin. Eventually we will rap back around to server 1.
    1. Round robin has the catch that a heavy request can cause a bunch of users to be a using a singular server while the other servers remain dormant.
    2. Caching in browsers (of the IP address for a particular link) can also cause a disproportionate amount of load on one server as the client continues to send request to one server.

- Trying to store session data on a dedicated set of servers or disks, which are shared across all other servers, allows for all servers to have session data for a client, however, it introduces a flaw. The dedicated disk space, if broken can result in a loss of all session data and there is no point to horizontally scaling.
- RAID (redundant array of independent disks) can be used to combat the issue of storing session data across multiple machines.
- There are several forms of raid:
    - RAID0: The system has two disks. Data is Striped across both hard drives, wherein a write operation happens on both disks, one after the other. This negates the time it takes to write a bunch of bits to the disk. This is good for performance.
    - RAID1: The systems still has two disks, however data is stored simultaneously across both drives.

  - RAID10: A combination of RAID0 and RAID1. You use 4 drives with both striping and redundancy.
  - RAID5: Only one drive is used to maintain redundancy and a bunch of others can be used for actual storage.
- Load balancers can use cookies to remember the session id and thus the server the user is using.
- Using multiple static pages as opposed to code which dynamically generates content has the benefit of being quicker with the downside of using more disk space and being very hard to change.
- Memchache is a piece of software that is used to store data on a dedicated or undedicated serve's RAM or memory.
  - A type of garbage collection can be implemented on the memchache so we don't run out of memory.
- Database use replication, where a master database contents are copied onto multiple slave databases.
- There are master-slave and master-master database setups.
  - For read heavy websites databases can use the slave databases for reading as well.
  - Load balancers can be used to partition user data on data bases based on certain bounds like letter the username starts from.
- High availability refers to replication of server states across a bunch of servers and that is managed by a load balancer.