Abstract and Virtualization

- When you use cloud computing, you are accessing pooled resources using a technique called virtualization. Virtualization assigns a logical name for a physical resource and then provides a pointer to that physical resource when a request is made.

- Virtualization provides a means to manage resources efficiently because the mapping of virtual resources to physical resources can be both dynamic and facile.

- Virtualization is dynamic in that the mapping can be assigned based on rapidly changing conditions, and it is facile because changes to a mapping assignment can be nearly instantaneous.

These are among the different types of virtualization that are characteristic of cloud computing:

- **Access:** A client can request access to a cloud service from any location.

- **Application:** A cloud has multiple application instances and directs requests to an instance based on conditions.

- **CPU:** Computers can be partitioned into a set of virtual machines with each machine being assigned a workload. Alternatively, systems can be virtualized through load-balancing technologies.
- **Storage:** Data is stored across storage devices and often replicated for redundancy.

Abstract and Virtualization

- To enable these characteristics, resources must be highly configurable and flexible. You can define the features in software and hardware that enable this flexibility as conforming to one or more of the following mobility patterns:
- P2V: Physical to Virtual
- V2V: Virtual to Virtual
- V2P: Virtual to Physical
- P2P: Physical to Physical
- D2C: Datacenter to Cloud
- C2C: Cloud to Cloud
- C2D: Cloud to Datacenter
- D2D: Datacenter to Datacenter

Abstract and Virtualization

- According to Gartner, virtualization is a key enabler of the **first four of five key attributes of cloud computing:**
- **Service-based:** A service-based architecture is where clients are abstracted from service providers through service interfaces.
- **Scalable and elastic:** Services can be altered to affect capacity and performance on demand.
- **Shared services:** Resources are pooled in order to create greater efficiencies.
- **Metered usage:** Services are billed on a usage basis.
- **Internet delivery:** The services provided by cloud computing are based on Internet protocols and formats.

Load Balancing and Virtualization

- One characteristic of cloud computing is virtualized network access to a service. No matter where you access the service, you are directed to the available resources. The technology used to distribute service requests to resources is referred to as load balancing.

- Load balancing can be implemented in hardware, as is the case with F5's BigIP servers, or in software, such as the Apache mod_proxy_balancer extension, the Pound load balancer and reverse proxy software, and the Squid proxy and cache daemon.

- Load balancing is an optimization technique; it can be used to increase utilization and throughput, lower latency, reduce response time, and avoid system overload.

The following network resources can be load balanced:

- Network interfaces and services such as DNS, FTP, and HTTP
- Connections through intelligent switches
- Processing through computer system assignment
- Storage resources
- Access to application instances

Load Balancing and Virtualization

- Without load balancing, cloud computing would very difficult to manage. Load balancing provides the necessary redundancy to make an intrinsically unreliable system reliable through managed redi-rection.
- It also provides fault tolerance when coupled with a failover mechanism. Load balancing is nearly always a feature of server farms and computer clusters and for high availability applications.
- A load-balancing system can use different mechanisms to assign service direction. In the simplest load-balancing mechanisms, the load balancer listens to a network port for service requests.
- When a request from a client or service requester arrives, the load balancer uses a scheduling algorithm to assign where the request is sent.
- Scheduling algorithms in use today are **round robin and weighted round robin,** fastest response time, least connections and weighted least connections, and custom assignments based on other factors.

- A session ticket is created by the load balancer so that subsequent related traffic from the client that is part of that session can be properly routed to the same resource.
- Without this session record or persistence, a load balancer would not be able to correctly failover a request from one resource to another.
- Persistence can be enforced using session data stored in a database and replicated across multiple load balancers. Other methods can use the client's browser to store a client-side cookie or through the use of a rewrite engine that modifies the URL.
- A session cookie stored on the client has the least amount of overhead for a load balancer because it allows the load balancer an independent selection of resources.
- The algorithm can be based on a simple round robin system where the next system in a list of systems gets the request.
- Round robin DNS is a common application, where IP addresses are assigned out of a pool of available IP addresses. Google uses round robin DNS,

Advanced load balancing :

- The more sophisticated load balancers are workload managers. They determine the current utilization of the resources in their pool, the response time, the work queue length, connection latency and capacity, and other factors in order to assign tasks to each resource.

- Among the features you find in load balancers are polling resources for their health, the ability to bring standby servers online (priority activation), workload weighting based on a resource's capacity HTTP traffic compression, TCP offload and buffering, security and authentication, and packet shaping using content filtering and priority queuing.

- An Application Delivery Controller (ADC) is a combination load balancer and application server that is a server placed between a firewall or router and a server farm providing Web services.

- An Application Delivery Controller is assigned a virtual IP address (VIP) that it maps to a pool of servers based on application specific criteria.

- An ADC is a combination network and application layer device. You also may come across ADCs referred to as a content switch, multilayer switch, or Web switch.

Vendors, among others, sell ADC systems :

- A10 Networks (http://www.a10networks.com/)
- Barracuda Networks (http://www.barracudanetworks.com/)
- Brocade Communication Systems (http://www.brocade.com/)
- Cisco Systems (http://www.cisco.com/)
- Citrix Systems (http://www.citrix.com/)
- 5 Networks (http://www.f5.com/)
- Nortel Networks (http://www.nortel.com/)
- Coyote Point Systems (http://www.coyotepoint.com/)
- Radware (http://www.radware.com/)

Vendors, among others, sell ADC systems :

- A10 Networks (http://www.a10networks.com/)
- Barracuda Networks (http://www.barracudanetworks.com/)
- Brocade Communication Systems (http://www.brocade.com/)
- Cisco Systems (http://www.cisco.com/)
- Citrix Systems (http://www.citrix.com/)
- 5 Networks (http://www.f5.com/)
- Nortel Networks (http://www.nortel.com/)
- Coyote Point Systems (http://www.coyotepoint.com/)
- Radware (http://www.radware.com/)

Load Balancing and Virtualization

- Services provided by an ADC include data compression, content caching, server health monitoring, security, SSL offload and advanced routing based on current conditions.
- An ADC is considered to be an application accelerator, and the current products in this area are usually focused on two areas of technology: network optimization, and an application or frame work optimization.
- An architectural layer containing ADCs is described as an **Application Delivery Network (ADN),** and is considered to provide WAN optimization services. Often an ADN is comprised of a pair of redundant ADCs.
- The purpose of an ADN is to distribute content to resources based on application specific criteria. ADN provide a caching mechanism to reduce traffic, traffic prioritization and optimization, and other techniques.
- ADN began to be deployed on **Content Delivery Networks (CDN)** in the late 1990s, where it added the ability to optimize applications (application fluency) to those networks.
- Most of the ADC vendors offer commercial ADN solutions.

Load Balancing and Virtualization

- Additional ADN vendors :
- Akamai Technologies (http://www.akamai.com/)
- Blue Coat Systems (http://www.bluecoat.com/)
- CDNetworks (http://www.cdnetworks.com/)
- Crescendo Networks (http://www.crescendonetworks.com/)
- Expand Networks (http://www.expand.com/)
- Juniper Networks (http://www.juniper.net/)
- Google's cloud is a good example of the use of load balancing.

The Google cloud :

- According to the Web site tracking firm Alexa (http://www.alexa.com/topsites), Google is the single most heavily visited site on the Internet; that is, Google gets the most hits.
- The investment Google has made in infrastructure is enormous, and the Google cloud is one of the largest in use today. It is estimated that Google runs over a million servers worldwide, processes a billion search requests, and generates twenty petabytes of data per day.
- Google is understandably reticent to disclose much about its network, because it believes that its infrastructure, system response, and low latency are key to the company's success.
- Google never gives datacenter tours to journalists, doesn't disclose where its datacenters are located, and obfuscates the locations of its datacenters by wrapping them in a corporate veil.
- The discretely named Tetra LLC (limited liability company) owns the land for the Council Bluffs, Iowa, site, and Lapis LLC owns the land for the Lenoir, North Carolina, site.

Load Balancing and Virtualization

- This makes Google infrastructure watching something akin to a sport to many people.
- Google's infrastructure and the basic idea behind how Google distributes its traffic by pooling IP addresses and performing several layers of load balancing.
- Google has many datacenters around the world. As of March 2008, Rich Miller of DataCenterKnowledge.com wrote that Google had at least 12 major installations in the United States and many more around the world.
- Google supports over 30 country specific versions of the Google index, and each localization is supported by one or more datacenters. For example, Paris, London, Moscow, Sao Paolo, Tokyo, Toronto, Hong Kong, Beijing and others support their countries' locale.
- Germany has three centers in Berlin, Frankfurt, and Munich; the Netherlands has two at Groningen and Eemshaven.
- The countries with multiple datacenters store index replicas and support network peering relationships. Network peering helps Google have low latency connections to large Internet hubs run by different network providers.

Load Balancing and Virtualization

- Google's datacenters are sited based on the following factors (roughly in order of importance):
- Availability of cheap and, if possible, renewable energy
- The relative locations of other Google datacenters such that the site provides the lowest latency response between sites
- Location of nearby Internet hubs and peering sites
- A source of cooling water
- The ability to purchase a large area of land surrounding the site
- Speculation on why Google purchases large parcels of land ranges from creating a buffer zone between the datacenter and surrounding roads and towns or possibly to allow for building wind farms when practical.
- Tax concessions from municipalities that lower Google's overhead.

Load Balancing and Virtualization

- Google maintains a pool of hundreds of IP addresses, all of which eventually resolve to its Mountain View, California, headquarters.
- When you initiate a Google search, your query is sent to a DNS server, which then queries Google's DNS servers.
- The Google DNS servers examine the pool of addresses to determine which addresses are geographically closest to the query origin and uses a round robin policy to assign an IP address to that request.
- The request usually goes to the nearest datacenter, and that IP address is for a cluster of Google servers. This DNS assignment acts as a first level of IP virtualization, a pool of network addresses have been load balanced based on geography.
- A Google cluster can contain thousands of servers. Google servers are racks of commodity 1U or 2U servers containing 40 to 80 servers per rack with one switch per rack.
- Each switch is connected to a core gigabit switch. Google servers run a customized version of Linux with applications of several types.

- When the query request arrives at its destination, a Google cluster is sent to a load balancer, which forwards that request to a Squid proxy server and Web cache dameon. This is the second level of IP distribution, based on a measure of the current system loading on proxy servers in the cluster.
- The Squid server checks its cache, and if it finds a match to the query, that match is returned and the query has been satisfied.
- If there is no match in the Squid cache, the query is sent to an individual Google Web Server based on current Web server utilizations, which is the third level of network load balancing, again based on utilization rates.
- It is the Google Web Servers that perform the query against the Google index and then format the results into an HTML page that is returned to the requester.
- This procedure then performs two more levels of load balancing based on utilization rates. Google's secret sauce is its in-memory inverted index and page rank algorithm.
- Google's GoogleBot crawls the Web and collects document information. Some details of the search and store algorithm are known. Google looks at the title and first few hundred words and builds a word index from the result. Indexes are stored on an index server.

Load Balancing and Virtualization

- Some documents are stored as snapshots (PDF, DOC, XLS, and so on), but lots of information is not addressed in the index.
- Each document is given a unique ID (“docid”), and the content of the document is disassembled into segments called shards, subjected to a data compression scheme and stored on a document server.
- The entire index is maintained in system memory partitioned over each instance of the index’s replicas. A page rank is created based on the significant links to that page.
- Queries are divided into word lists, and the Google algorithm examines the words and the relationships of one word to another.
- Those word relationships are mapped against the main index to create a list of documents, a feature called an inverted index. In an inverted index, words are mapped to documents, which can be done very quickly when the index is fully kept in memory.
- The Web server takes the result of a query and composes the Web page from that result. Ads included on the page are from ad servers, which provide Google’s AdSense and AdWords services.

Load Balancing and Virtualization

- The query also is presented to a spelling server to provide suggestions for alternative spellings to include in the search result.
- Certain keywords, data input patterns, and other strings are recognized as having special operational significance.
- For example entering “2 plus 2” initiates Google’s calculator program, and a ten-digit number returns a reverse phone lookup using the phonebook program. These programs are supported by special application servers.
- Google doesn’t use hardware virtualization; it performs server load balancing to distribute the processing load and to get high utilization rates.
- The workload management software transfers the workload from a failed server over to a redundant server, and the failed server is taken offline.
- Multiple instances of various Google applications are running on different hosts, and data is stored on redundant storage systems.