CSCI 1380

Day 6
Load Balancing

**global LB**
map users to Datacenters

**Cluster LB**
Once in a datacenter, maps users to server

Load Balancing for Edge

1. Latency (closest location)
2. Content-based
3. # of conn (capacity)
4. Health (availability)
5. Policy (GDPR)

DNS: protocol for mapping users to Edge Pops/CDNs
(alternative is BGP)

**DNS v. BGP**
- FB
- Azure
- Verizon
DNS

What is google.com?

Google.com Name server

Like zipcodes, IP addr

create location info

when a user connects
to a DNS svr

get user IP

use user IP to
determine user loc

Pick CDN

closest to user loc

Special Global LB logic

Problems with DNS

1. Single of failure
2. Limited capacity
3. Limited storage
4. Overload

CDN1
10.10.10.10

CDN2
7.7.3.8

Practical DNS

1. Distributed infra
2. Distributed control

ISP

Root

.edu
.com
.org

9.8.7.6

Cookies of resv'd names

query google.com

Mail, google.com

To:

YouTube, google.com

IP

IP
DNS caching & Global LB

with caching improve performance by reducing/eliminating lookups

But creates a problem → company can't change your cache & update policies

each DNS response has a TTL (Time To Live) which determines the time period a cache entry is valid for

BGP (ANYCAST)

1) Every CDN has same address

2) network "automatically" routes to one location (hopefully the closest location)
DNS server update takes less time (simple replication)

BGP convergence takes time (run the whole BGP)
Datacenter / cluster / LB

1. Which DC has the data or service? Send the user there...

2. Within the data center, many copies/replicas of your data or the application

   - Techniques
     1. Round robin
     2. Random

   - Don't maintain state
     (requires a mapping for stateful services)

   - Goal: evenly distribute request across servers

   - Next class
     1. Modulo LB
     2. Consistent hashing
LB (Load Balancers)

1. Global (maps users to CDNs)
   (*) DNS versus BGRP
   (*) How to scale (distribution & caching)
   (*) Complex policies (location, policy, content)

2. Cluster LB
   (*) Distribute load across servers
   (*) Random / Round Robin

What happens when you connect? When do you use CDNs versus DC?

always interact
w/ CDN
CDN
if missing info/app
response
# (>100)

go to datacenter
DC

#
(10-20)