Scalable Internet Servers and Load Balancing

CS 257/457
Dept. of Computer Science, University of Rochester

Internet Services and Servers

- Internet Services
  - Services hosted by computer systems, accessible to online users through Internet.
- Services on the Internet
  - Online keyword search engine: Google.
  - Web email service: hotmail.
  - News service: CNN.
  - Other portal services: Yahoo!, AOL, MSN.
- Internet Servers
  - Computer systems that host Internet services.

Internet Services are Applications

An Example: How does Google work? (Part I)

- First, we need to get all these Web pages out there - crawling.
- Then we need to reformat them to make them easy to search - indexing.
- As part of indexing, we need to give each page a ID.

Java:
  Page #123  Page #357  ...  ...
Mexico:
  Page #124  Page #468  ...  ...

An Example: How does Google work? (Part II)

So what is the big deal?

- Scalability: How about searching over 2 billion Web pages (with an index size of several Terabytes)?
- Throughput: How about serving 150 million search queries per day?
- Response time: Come on! I have been waiting here for two seconds. Where is the result?
- Reliability: With 1,000 servers and 4,000 disks in your machine room, something is gonna break every day!
Technique 1: Partitioning
The index database is split into many partitions ⇒ better scalability

Java: Page #123 | Page #357 | ... ...
Mexico: Page #124 | Page #468 | ... ...
Food: Page #124 | Page #377 | ... ...
Medicine: Page #12 | Page #468 | ... ...

Technique 2: Replication
- Multiple servers to provide the same service
- More throughput: if each Web server can answer 10 requests/second, then ten Web servers can answer 100 requests/second (Well, at least in theory)
- Better reliability
- Faster response??
- Challenges:
  - Figure out who is least loaded
  - Figure out who is dying

Revisit: How does Google work?

Load Balancing on Internet Servers
Technique 1 - DNS Rotation

Discussions on DNS Rotation
- Problems
  - DNS Caching
  - Rigid load balancing policy - can't balance based on the number of active requests
  - Slow or no feedback
    - no adjustment in response to runtime load changes
    - slow or no adjustment in response to failures
  - Is there anything good about it?
  - Require almost no change on the existing Internet architecture

Load Balancing on Internet Servers
Technique 2 - Cooperative Offloading
Discussions on Cooperative Offloading

- Advantages:
  - More flexible policy is possible
  - Be more responsive to runtime workload and server failures

- Problems:
  - Need a lot more software
  - Longer delay

Discussions on Load Balancing Switch

- Different with cooperative offloading
  - We are messsing around with TCP/IP kernel
  - better performance - no extra application-level processing

- Changes required on all party involved
  - DNS server??
  - Web server??
  - client??
  - switch??????

Load Balancing Policies in LB Switches

- Simple rotation
- Least number of active requests
- Shortest response time

Modern LB Switch Construction

Much like a router: switching at wire speed

Summary

- Scalable Internet servers
- partitioning
- replication
- Load balancing on Internet servers
  - DNS rotation
  - cooperative offloading
  - LB switches
- Is there anything missing?