internet services

Badri Nath
Rutgers University
badri@cs.rutgers.edu
What are Internet services?

- What is an internet?
  - Network of networks
- What is the Internet?
  - A global internet based on the IP protocol
- To what does “Internet technology” refer?
  - Architecture, protocols and services
- Services accessed over the net
Systems

- Storage systems
- Networks
- Operating Systems
- Distributed systems
- Application servers
- Databases
- Embedded systems (physical infrastructure)
- Wireless systems
Internet Players

- **Users**, people who use the applications
  - Everyone (mom and pop, kids)
  - get something done (hopefully useful)

- **Service Designers**
  - You: protocol design and implementation
  - Scale, performance, cost, incremental deployment

- **Service Providers/middleware**
  - Administrators and ISPs
  - Management, revenue, deployment

- **Market/business models for the Internet**
  - Consumer to consumer (ebay), Business to consumer (amazon, Orbitz), Business to business (IBM, ARIBA), Consumer to business (hotjobs, monster)
Simple Client server abstraction
Client sends a request, server sends a response
Informational, transactional in nature
Evolution of Internet services

- **Web 1.0**
  - Users as readers of content (read-only)

- **Web 2.0**
  - Users create content (group communication)
  - Read-write

- **Web 3.0**
  - Read, write, execute (in context, personalization), programs

- **Web 4.0**
  - Every living/non-living object connected?
Internet services (whats happening)

- Complicated Sub-services introduced
- In the name of performance (Throughput)
  - Well defined quantitatively (request/sec, byte/sec, transaction/sec)
- What about other metrics? (qualitative maybe)
  - Useability
  - Manageability
  - Diagnosability
  - Configurability
  - Dependability
  - Trustability
## Search

<table>
<thead>
<tr>
<th>Top Ten Search Sites By Unique Visitors</th>
<th>Unique Visitors (000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Web Search</td>
<td>75,281</td>
</tr>
<tr>
<td>Yahoo! Search</td>
<td>68,031</td>
</tr>
<tr>
<td>MSN Search</td>
<td>49,748</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>43,705</td>
</tr>
<tr>
<td>AOL Search</td>
<td>36,092</td>
</tr>
<tr>
<td>Yahoo! Local Search</td>
<td>20,270</td>
</tr>
<tr>
<td>MySpace Search</td>
<td>8,083</td>
</tr>
<tr>
<td>Infospace Web Search</td>
<td>5,942</td>
</tr>
<tr>
<td>LookSmart</td>
<td>4,402</td>
</tr>
<tr>
<td>Lycos Network Search</td>
<td>5,249</td>
</tr>
</tbody>
</table>

Source: comScore Media Metrix qSearch data, October 2005.
## Search Traffic queries 2008

<table>
<thead>
<tr>
<th>Core Search Entity</th>
<th>July 2008 (%)</th>
<th>July 2008 (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total core search</td>
<td>100.0</td>
<td>12B</td>
</tr>
<tr>
<td>Google sites</td>
<td>61.9</td>
<td>7.44B</td>
</tr>
<tr>
<td>Yahoo sites</td>
<td>20.5</td>
<td>2.4B</td>
</tr>
<tr>
<td>Microsoft sites</td>
<td>8.9</td>
<td>1.08B</td>
</tr>
<tr>
<td>Ask Network</td>
<td>4.5</td>
<td>0.54B</td>
</tr>
<tr>
<td>AOL LLC</td>
<td>4.2</td>
<td>0.52</td>
</tr>
</tbody>
</table>

**Notes:**
Internet Traffic

Consumers Leading Way –
58% CAGR in IP Traffic, 2005E-2011E

Global IP Traffic

Source: Cisco Systems, Global IP Traffic Forecast and Methodology; Mobility segment (0.1% of traffic in 2007) not displayed
## Unique Visitors (Million-Month)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Property</th>
<th>4/07</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Google</td>
<td>528</td>
</tr>
<tr>
<td>2</td>
<td>Microsoft</td>
<td>520</td>
</tr>
<tr>
<td>3</td>
<td>Yahoo!</td>
<td>468</td>
</tr>
<tr>
<td>4</td>
<td>Time Warner</td>
<td>267</td>
</tr>
<tr>
<td>5</td>
<td>eBay</td>
<td>248</td>
</tr>
<tr>
<td>6</td>
<td>Wikipedia</td>
<td>209</td>
</tr>
<tr>
<td>7</td>
<td>YouTube</td>
<td>163</td>
</tr>
<tr>
<td>8</td>
<td>Amazon</td>
<td>137</td>
</tr>
<tr>
<td>9</td>
<td>CNET</td>
<td>120</td>
</tr>
<tr>
<td>10</td>
<td>Ask</td>
<td>111</td>
</tr>
<tr>
<td>11</td>
<td>Apple</td>
<td>115</td>
</tr>
<tr>
<td>12</td>
<td>MySpace</td>
<td>107</td>
</tr>
<tr>
<td>13</td>
<td>Adobe</td>
<td>101</td>
</tr>
<tr>
<td>14</td>
<td>Lycos</td>
<td>79</td>
</tr>
<tr>
<td>15</td>
<td>Viacom</td>
<td>81</td>
</tr>
</tbody>
</table>
Internet users

Geographic Distribution of Internet Users (MM)

North America | Europe | Asia/Pacific | Rest of World | Latin America

2000E: 390
2001E: 491
2002E: 621
2003E: 724
2004E: 851
2005E: 977
2006E: 1,122
2007E: 1,302
## Impact of Web 2.0

### Alexa Global Traffic Rankings

<table>
<thead>
<tr>
<th>Rank</th>
<th>Web site</th>
<th>Rank</th>
<th>Web site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yahoo.com</td>
<td>1</td>
<td>yahoo.com</td>
</tr>
<tr>
<td>2</td>
<td>msn.com</td>
<td>2</td>
<td>youtube.com</td>
</tr>
<tr>
<td>3</td>
<td>google.com</td>
<td>3</td>
<td>live.com</td>
</tr>
<tr>
<td>4</td>
<td>ebay.com</td>
<td>4</td>
<td>google.com</td>
</tr>
<tr>
<td>5</td>
<td>amazon.com</td>
<td>5</td>
<td>myspace.com</td>
</tr>
<tr>
<td>6</td>
<td>microsoft.com</td>
<td>6</td>
<td>facebook.com</td>
</tr>
<tr>
<td>7</td>
<td>myspace.com</td>
<td>7</td>
<td>msn.com</td>
</tr>
<tr>
<td>8</td>
<td>google.co.uk</td>
<td>8</td>
<td>hi5.com</td>
</tr>
<tr>
<td>9</td>
<td>aol.com</td>
<td>9</td>
<td>wikipedia.org</td>
</tr>
<tr>
<td>10</td>
<td>go.com</td>
<td>10</td>
<td>orkut.com</td>
</tr>
</tbody>
</table>

Traffic rank is based on three months of aggregated historical traffic data from Alexa Toolbar users and is a combined measure of page views / users (geometric mean of the two quantities averaged over time).
Web tools and age

Social Networking – Connectivity Changing…Is E-Mail Becoming Archaic?

Worldwide Share of Online Time (1)

- All Other: 6%
- Communications: 35%
- Social Connections: 14%
- Shopping & Travel: 16%
- Entertainment & Leisure: 8%
- Work, Business & Education: 22%

Category didn’t exist 3 years ago

Connecting – Younger Users Via Social Networks + Older Users Via E-Mail? (2)

<table>
<thead>
<tr>
<th>Social Network</th>
<th>Age 15-24</th>
<th>Age 44+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo! Mail</td>
<td>16%</td>
<td>37%</td>
</tr>
<tr>
<td>Facebook</td>
<td>38%</td>
<td>23%</td>
</tr>
</tbody>
</table>

Source: (1) comScore ‘Digital World – State of the Internet’ 3/06; (2) comScore global 1/08
# Monthly unique visitors

<table>
<thead>
<tr>
<th>Users</th>
<th>Monthly Unique Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="#">YouTube</a></td>
<td>258MM&lt;sup&gt;(1)&lt;/sup&gt;</td>
</tr>
<tr>
<td><a href="#">Facebook</a></td>
<td>101MM&lt;sup&gt;(1)&lt;/sup&gt;</td>
</tr>
<tr>
<td><a href="#">Skype</a></td>
<td>276MM&lt;sup&gt;(2)&lt;/sup&gt;</td>
</tr>
<tr>
<td><a href="#">PayPal</a></td>
<td>57MM&lt;sup&gt;(2)&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Acc to media metrix (MM)
Source of Information

Internet – Importance as Source of Information

% of Users Age 17+ Responding Important / Very Important

- Internet: 80%
- Television: 68%
- Radio: 63%
- Newspaper: 63%
- Personal Source: 73%
Mobile users vs Internet users

- Internet Users – 1.3B
  - C2007E (1)
  - N. America: 16%
  - Europe: 26%
  - Asia Pacific: 43%
  - ROW: 7%

- Mobile Subscribers – 3.3B
  - C2007E (2)
  - N. America: 8%
  - Europe: 29%
  - LatAm: 11%
  - ROW: 12%
  - Asia Pacific: 40%

Source: (1) ITU, Euromonitor, Morgan Stanley Research; (2) Informa, Morgan Stanley Research
## Mobile Subscriber Base

<table>
<thead>
<tr>
<th>Country</th>
<th>2006 Mobile Subscriptions (000s)</th>
<th>2006 Growth</th>
<th>Mobile Subscriptions Penetration</th>
<th>Worldwide Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>445,754</td>
<td>18%</td>
<td>34%</td>
<td>17%</td>
</tr>
<tr>
<td>USA</td>
<td>232,793</td>
<td>12</td>
<td>78</td>
<td>9</td>
</tr>
<tr>
<td>Russia</td>
<td>151,937</td>
<td>20</td>
<td>106</td>
<td>6</td>
</tr>
<tr>
<td>India</td>
<td>137,369</td>
<td>82</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Brazil</td>
<td>100,661</td>
<td>17</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>Japan</td>
<td>94,936</td>
<td>5</td>
<td>74</td>
<td>4</td>
</tr>
<tr>
<td>Germany</td>
<td>81,242</td>
<td>9</td>
<td>99</td>
<td>3</td>
</tr>
<tr>
<td>Italy</td>
<td>77,605</td>
<td>11</td>
<td>133</td>
<td>3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>69,557</td>
<td>7</td>
<td>115</td>
<td>3</td>
</tr>
<tr>
<td>Indonesia</td>
<td>58,654</td>
<td>47</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>Mexico</td>
<td>56,765</td>
<td>22</td>
<td>55</td>
<td>2</td>
</tr>
<tr>
<td>Turkey</td>
<td>51,659</td>
<td>19</td>
<td>71</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>51,442</td>
<td>7</td>
<td>84</td>
<td>2</td>
</tr>
<tr>
<td>Pakistan</td>
<td>48,543</td>
<td>124</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>Spain</td>
<td>46,339</td>
<td>10</td>
<td>105</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,693,087</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Morgan Stanley Research TMT database, Informa
Multi-Tiered Internet Services/farm

Tier 1:
Web servers

Tier 2:
App servers

Tier 3:
Database server

Tier 4:
Storage networks
Storage services

- No longer a disk or raid
- Storage area network (SAN)
  - Different FS connected to a disk pool via a dedicated high speed network
- Network attached storage (NAS)
  - A special purpose server that offers heterogeneous File sharing
  - But ....
- The cost of storage management is twice the cost of storage
NAS vs SAN

- SAN (Storage Area Network)
- NAS (network Attached Storage)
More on NAS v. SAN

- **NAS**
  - Wires: TCP/IP
  - Protocol: NFS, CIFS

- **SAN**
  - Wires: Fiber Channel
  - Protocol: Encapsulated SCSI

How about home network storage?
Cost of hardware

- Storage cost is decreasing (storage size is increasing)
  - $0.5 Per gigabyte
- Hardware cost is decreasing
  - $200 to $400 PC
- Cost of B/W is decreasing
  - All you can eat for 19.99
- But human cost of managing storage!!
Storage is cheap

- 4GB in < 1.5 cubic inches!

From: Morgan Stanley presentation on internet trends
Managing storage

- Organizations already have petabytes of data
- How to organize, manage, backup, analyze?
Cost trends

Storage areal density CGR continues at 100% per year to >100 Gbit/in². The price of storage is decreasing rapidly, and is now significantly cheaper than paper.

Surajit Chaudhuri– VLDB tutorial notes
Management cost

The High Cost of I/T Management

For example: the cost to manage storage is typically twice the cost of the actual storage system.


1984:
- $2 million System
- $1 million Storage Administration
- $1 million

2000:
- $2 million Storage Administration
- $3 million

---


Surajit Chaudhuri—VLDB tutorial notes
The Internet

- Configuration, fault diagnosis is still manual
- Need a separate plane (knowledge plane) for improving soft metrics
- Aggregates global information, observations, assertions, requirements, constraints, and goals.

What can it do -- "soft metrics"

- Diagnosability

- Answer "Why" questions
  - "Why can’t I get to Yahoo?"
  - Why can’t I access my mail server?

- Monitoring

- Collect network data

- How do you communicate to other nodes what you are seeing?
Configurability

- Network entities manually configured and managed
- High-level network configuration
- Assertions to low level instructions
- Policy ➔ implementation, enforcement
Can we design for?

- Diagnosis -- “Why?” questions
- Monitoring
- Auto-configuration
Human Costs Dominate in Database, Too

81% is “People Cost”

Source: The Aberdeen Group, 1998
http://irelay.bvk.co.yu/progres/aberdeen/aberdeen.htm
Everything is complicated
Wireless networking

- Cannot afford one person per 10 access points
- How to manage
- How do you something is wrong
- Does Ping always work?
- Is a wireless network easier to manage than a wired network?
Wireless networking

- Cheap devices have expensive state

Just Phone

Phone + PDA + Camera
Operating systems

- What happens when system hangs
- System hang time = human hang time
- Diagnostics
- Restoration
- Remote reboot
- Loss of work
Embedded systems

- Sensor networks
- Long term survivable networks
- Deployed in remote and unattended regions
- Cannot afford human RTT (round trip times)
- Trustability an important issue
Sensor Networks

- Ability to measure (sense) attribute values
- Communicate these measured values
- New information architecture

Enabling ubiquitous sensing using RFID, Roy Want, IEEE Computer magazine.
RFID vs barcodes

- Tags can be read as opposed to scan
- Barcodes are scanned
  - Scan is a human activity
  - Scan is a serial process
- Reading is parallel
- What about management of RFID networks
- Placement of readers, diagnostics
- Measurement attacks
Killer app for localization
Course Goals

- Understand techniques for scaling internet services
- Study components of internet services
  - Storage
  - Operating systems
  - Services
  - Databases
  - Wireless
  - Embedded systems
- Evaluate features and design systems that scale using any metric(s)
Course Goals

- System/services design from the viewpoint of scale
- Macro level analysis
  - Overall system design (cluster, power, programming, access)
- Micro-level analysis
  - Feature set
  - What feature needs to be there?
Course Goals

- Understand the basic principles of storage services, operating services, networking services, database services, and embedded system services
- Study new metrics for internet services design
- How to do research
  - How to determine what is important
  - What are the trends
  - What are the economics, technology that is driving innovation
Course Goals (scalability, useability, manageability)

- Metrics tree

- Threats
  - What causes the metric value to be affected

- Attributes
  - What are characteristics

- Means
  - How do we achieve it?

Fundamental concepts of dependability – Avizienis, Laprie, Randell
Administrative

- CS553 webpage
- www.cs.rutgers.edu/~badri/553.html
- Prerequisites: computer networks, databases
- TA: Pravin Shankar
  (Spravin@cs.rutgers.edu)
- Project on OpenMoko phones
Grading

- 20% Paper presentation
- 15% Paper reviews/quiz
- 40% Programming projects
- 25% Final (based on papers presented)
Reading paper

- Understand the basic idea
- What kind of paper?
  - Performance, vision, new direction/protocol paper
- Summarize key idea
- +ve aspects of the paper
  - New, breakthrough, incremental,
- -ve aspects of the paper
  - Assumptions (valid?), scaling issues (does it scale), new metrics, can we evaluate based on new “ilities”, implementation (has it been implemented), measurements (problems?)
- Questions you wish you knew the answer.
Internet service

- Service: a set of actions that can be used to accomplish a goal
- Service Providers, Service consumers agree upon or have an understanding of the goal
- Internet Service: service over the net
  - Ebay service: Buy and sell goods
  - On-line Banking service: offers banking related services
  - Amazon: sells goods
  - E-loan: sells mortgages
  - Google: search web pages
Definitions

- Architecture of a typical internet service
  - General architecture is
    - 1(LB) + 3 (web, app, db) + 1(storage) tiers
  - Can be just 3 tiers (LB + AppServer + storage)
  - Hasn’t changed much since late 1990s

- Main components of the architecture
  - Users, network, Redirector/Load balancer, webs server, app server, db, storage
Users

- User population growing globally
  - Ebay, 69 M users, 1 B hits a day
  - Yahoo, 400 M users
  - Hotmail, 150 M users
  - Google, 0.2 B queries per day
  - Amazon, 50 M users
Load balancers (LB)

- Requests from users classified and directed/routed
- Differentiation based on source, destination addresses
  - User address, ISP etc Level-3 redirector or rotating DNS
L4, L7 load balancers

- Level 4, understands TCP connections
  - Differentiation based on port numbers
    - Web requested routed differently than SMTP request
- Level 7, understands application level information
  - Differentiation based on application-level info
    - Can parse URLs
    - Look at parameter of GET

SMTP

WEB

Camera.Yahoo.com
maps.Yahoo.com
launch.Yahoo.com
Load balancers

- **DNS rotation**
  - Static, if redirected node has failed
  - Takes a long time to switch to another node

- **Level 3 switch**
  - Fast, can detect failed nodes
  - Bottleneck link into the switch, single point of failure
  - Difficult to balance load for global users

- **Level 4, 7 switch**
  - Need to parse port numbers, URLs at line speed
  - Good TCP-level load partitioning
  - Bottleneck link into the switch, single point of failure
  - Can use from-end applications to do smarter balancing

Lessons from Giant-Scale Services – Internet Computing – July, ‘01
- Eric Brewer
Servers

- **Web servers**
  - Process HTML requests, render HTML content, process user input

- **App servers**
  - Process Business logic, business rules, security, authentication
  - Prepare data for web servers

- **DB servers**
  - Execute DBMS queries, retrieve user+app data

- **Storage servers**
  - SAN, NAS, combo, Disk farms
  - Maintain persistent data pertaining to service
Background: WWW & Internet Services

- Tremendous growth in network size since early 1990s
  - WWW, search
  - Commercialization
  - Cheap PCs

- Services’ popularity
  - Integral part; e.g., Yahoo, Ebay, Google Gmail, Amazon, Buy.com, weather.com, etc.
  - Growing client base: Google handles ~300 million requests/day
### Table 1: Example Clusters for Giant-Scale Services

<table>
<thead>
<tr>
<th>Service</th>
<th>Nodes</th>
<th>Queries</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOL web cache</td>
<td>&gt;200</td>
<td>2.5B/day</td>
<td>4-CPU DEC 4100s</td>
</tr>
<tr>
<td>Inktomi Search Engine</td>
<td>800</td>
<td>&gt;40M/day</td>
<td>2-CPU Sun Workstations</td>
</tr>
<tr>
<td>Geocities</td>
<td>&gt;300</td>
<td>&gt;25M/day</td>
<td>PC Based</td>
</tr>
<tr>
<td>Anonymous web-based e-mail</td>
<td>&gt;200</td>
<td>25M/day</td>
<td>FreeBSD PCs</td>
</tr>
</tbody>
</table>

Table A. Example clusters for giant-scale services.

<table>
<thead>
<tr>
<th>Service</th>
<th>Nodes</th>
<th>Queries</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOL Web cache</td>
<td>&gt;1,000</td>
<td>10B/day</td>
<td>4-CPU DEC 4100s</td>
</tr>
<tr>
<td>Inktomi search engine</td>
<td>&gt;1,000</td>
<td>&gt;80M/day</td>
<td>2-CPU Sun Workstations</td>
</tr>
<tr>
<td>Geocities</td>
<td>&gt;300</td>
<td>&gt;25M/day</td>
<td>PC Based</td>
</tr>
<tr>
<td>Anonymous Web-based e-mail</td>
<td>&gt;5,000</td>
<td>&gt;1B/day</td>
<td>FreeBSD PCs</td>
</tr>
</tbody>
</table>
Ebay Stats from 2003

- 69 Million Registered Users
- 16 Million Items across 28,000 categories
- 2002: Gross Sales = $14.87 Billion
- Global community
- 1 Billion hits per day
- Over 1200 URLs
Scale

- Hotmail: 170M users [NY Times]
- Google: 1B hits/day ➔ > 100,000 nodes,
  - Google FS serves a PByte
- Ebay: > 1

Ebay registered users

Source: eBay Inc. estimates
## Scale

### Ebay traffic

<table>
<thead>
<tr>
<th></th>
<th>1999</th>
<th>2002</th>
<th>2005*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outbound Traffic (at peak)</td>
<td>410 Mbps</td>
<td>&gt; 2 Gbps</td>
<td>&gt; 5 Gbps</td>
</tr>
<tr>
<td>Dynamic Page Views (per day)</td>
<td>380M</td>
<td>1B</td>
<td></td>
</tr>
<tr>
<td>Searches (per day)</td>
<td>11M</td>
<td>50M</td>
<td>150M</td>
</tr>
</tbody>
</table>

A Billion Hits a Day – JavaOne ’03  
- Deepak Alur, Rajmohan Krishnamurthy and Arnold Goldberg  
EBay booms as member numbers grow – ENN – Apr, 04  
- Ciaran Buckley
Maintenance

- Gmail maintenance cost -+$2/GB [NY Times]

Ebay maintenance
- More than 15 major features go live-to-site every quarter
- Nearly 30K lines of code change per week
- 4 times/day, 5 days/week HTML rolls to the site
- Internationalized releases phased in 3 weeks
- Instantaneous feedback from millions of users
Service availability is in the range of 2 to 3 nines.
Availability

- **Availability** is the percentage of time when system is operational.
- \[ A = \frac{MTBF}{MTBF + MTTR} \]
- Availability is typically specified in nines notation.
- For example, 3-nines availability corresponds to 99.9% availability.
- A 5-nines availability corresponds to 99.999% availability.
- **Availability** vs **Downtime**
  - 90% (1-nine) 36.5 days/year
  - 99% (2-nines) 3.65 days/year
  - 99.9% (3-nines) 8.76 hours/year
  - 99.99% (4-nines) 52 minutes/year
  - 99.999% (5-nines) 5 minutes/year
  - 99.9999% (6-nines) 31 seconds/year
Service growth

- Collection of apps, h/w, data, storage → farm
  - Farm grows in two ways
    - Cloning
      - Resources are replicated
    - Partitioning
      - Additional resources are added, but load is divided

Scalability terminology, Bill Devlin, Jim gray, Bill Laing, George Spix, Microsoft, December 1999
Cloning (Reliable array of cloned services)

- **Shared-nothing RACS**
  - Each clone has identical storage
    - Highly available
    - Writes on all copies
    - Good for read-mostly

- **Shared-disk a.k.a cluster**
  - Clones shared common disk pool
  - Storage should be fault-tolerant
  - Update traffic can become bottleneck
    - Use SAN
Partitioning

- PACK: Add resources but partition data

Data automatically partitioned by application middleware
Failures affect result in partial service failure
Even with Fault tolerant storage, H/W, S/W affects availability
PACKS

- Shared disk packs
- Two or more nodes per pack
  - Can tolerate S/W, H/W faults
- Packs have access to all storage
  - One Pack can takeover when primary fails
- Data still partitioned
PACKS

- Shared nothing
- Packs serve different partitions
  - Active-active
    - All nodes are active
    - Failure within a pack, another node takes over
  - Active-passive
    - Secondary-nodes in hot standby mode
    - Takes over only on failure

A-Z A-M N-Z

Active-active

Active-passive

A-Z

A-M

N-Z
The Google Cluster Architecture

- Barroso, Dean and Holzle, IEEE Micro 2003
GOOGLE ARCHITECTURE OVERVIEW

- Reliability
  - Software is the king
  - H/W is commodity, can be swapped
  - so that commodity PCs can be used
    - to build a cluster at a low price

- Design
  - for best aggregate throughput
  - rather than peak server response time

- Building a reliable computing infrastructure
  - from clusters of unreliable commodity PCs
    - RAID for disks; PC clusters is to server farms
SERVING A GOOGLE QUERY

- When user enters a query
  - e.g. www.google.com/search?q=ieee+society
  - User browser
    - Domain Name System (DNS) lookup
    - to map to a particular IP address

- Multiple Google clusters distributed worldwide
  - each cluster with a few thousand machines
  - to handle query traffic
Global Google Name Servers

- Over 30 Google clusters around the world.
  - DNS based & geo location driven load-balancing:
    - Domain Name: GOOGLE.COM
      , 4 name servers

- GOOGLE.COM NAME SERVERS

- Name Server   IP   Location
  - ns1.google.com 216.239.32.10
    Mountain View, CA, US
  - ns2.google.com 216.239.34.10
    Mountain View, CA, US
  - ns3.google.com 216.239.36.10
    Mountain View, CA, US
  - ns4.google.com 216.239.38.10
    Mountain View, CA, US
Cluster Architectures

- Cluster
  - Collection of independent computers using switched network to provide a common service
- Main frames, multiprocessor servers offer services at a higher price point.
- Adv of clusters: Scale, cost, repair, failover
- Most web portals (Amazon, eBay, etc) use clusters
DESIGN PRINCIPLES OF GOOGLE CLUSTERS

- Software level reliability
  - No fault-tolerant hardware features; e.g.
    - redundant power supplies
    - A redundant array of inexpensive disks (RAID)
  - instead tolerate failures in software

- Use replication
  - for better request throughput and availability

- Price/performance beats peak performance
  - CPUs giving the best performance per unit price
  - Not the CPUs with best absolute performance

- Using commodity PCs
  - reduces the cost of computation
Google rack

- Google’s racks
  - consist of up to 80 x86-based servers
  - Server components similar to mid-range desktop PC
    - except for larger disk drives
- Ranging
  - from single processor 533-MHz Intel-Celeron based servers
  - to dual 1.4-GHz Intel Pentium III servers
- Servers on each rack interconnected via 100 Mbps Ethernet
- All racks interconnected via a gigabit switch
THE POWER PROBLEM

- A mid-range server with dual 1.4-GHz Pentium III processors
  - 90 W of DC power
    - 55 W for the two CPUs
    - 10 W for a disk drive
    - 25 W for DRAM and motherboard
  - Typical efficiency of an ATX power supply -> 75%
    - means 120 W of AC power per server
    - roughly 10 (120*80) kW per rack
A rack
- 25 ft$^2$ of space
- Corresponding power density: 400 W/ ft$^2$
- With higher-end processors: 700 W/ft$^2$

Typical power density for commercial data centers: between 70 and 150 W/ft$^2$
- Much lower than that required for PC clusters
- Special cooling or additional space required
  - to decrease power density to a tolerable level
Server Farms/datacenter

Google Data Center – The Dalles, Oregon