

# Advanced Computer Architecture: *A Google Search Engine*

Jeremy Bradley

Room 372. Office hour - Thursdays at 3pm. Email: [jb@doc.ic.ac.uk](mailto:jb@doc.ic.ac.uk)

Course notes: <http://www.doc.ic.ac.uk/~jb/>

Department of Computing, Imperial College London

Produced with prosper and L<sup>A</sup>T<sub>E</sub>X

# Properties of a search engine

- ➔ Different time zones  $\Rightarrow$  24h availability
- ➔ Response time  $< 0.5s$  per query (including network latency)
- ➔ Multiple interconnected sites
- ➔ Multiple redundant high bandwidth connections to internet

# Potential Query Load

- ➔ Handle 1000s of queries per second
  - ➔ 1994: WWW receives 1 per minute
  - ➔ 1997: Altavista handles 200 per second
  - ➔ 2000: Google handles 1000 per second
  - ➔ 2003: Google handles 1700 per second?
- ➔ Keep up-to-date picture of web sites
- ➔ Need to do a web crawl at least every 4 weeks

# Extra services

- ➔ Acts as an archive service caching heterogeneous document types
- ➔ Provides a translation engine to provide on-the-fly translation of multilingual documents
- ➔ Second guess user input errors e.g. spelling mistakes (reduce future queries)

# Bandwidth issue

- ➔ 2000 queries per second (average 20K per results page)
  - ➔ 312Mbit s<sup>-1</sup> downlink
- ➔ Web crawl: 3,308,000,000 documents every 4 weeks (average 50K per page)
  - ➔ = 154 terabytes ( $1 \times 10^{12}$  bytes)
  - ➔ Requires 533Mbit s<sup>-1</sup> uplink

# Bandwidth issue II

- ➔ Assume 7Tbyte index ( $\sim 2\text{k}$  index material per page)
  - ➔ Index duplication to 3 sites once per week
  - ➔ Requires  $291\text{Mbit s}^{-1}$
- ➔ Total  $1136\text{Mbit s}^{-1}$
- ➔ Compare with total of  $211\text{Mbit s}^{-1}$  in 2000 (Hennessy 2003)

# Snapshot from Dec. 2000

Information/diagram in pp. 857–860, [Hennessy]:

- ➔ 6000 processors
- ➔ 12000 hard disks = 1 petabyte storage  
( $1 \times 10^{15}$  bytes)
- ➔ 4 independent sites
  - ➔ 2 on West coast of US
  - ➔ 2 in West Virginia

# Snapshot II

- ➔ Search index (of order terabytes) is replicated across all sites
- ➔ Search index and local version of internet cache stored at each site
- ➔ Each site connected to internet with OC48 connection ( $2488\text{Mbit s}^{-1}$  link)
- ➔ OC12 link ( $622\text{Mbit s}^{-1}$ ) connects pairs of local sites

# Google Cluster Architecture

Each cluster location has:

- ➔ 2 BigIron 8000 switches both connect
  - ➔ OC48 and OC12 external lines to...
  - ➔ 128  $1\text{Gbit s}^{-1}$  possible ethernet interfaces
- ➔ Each rack is connected twice to each BigIron switch i.e. 4 interface cards
  - ➔ Max of 64 racks per location

# A single rack...

- ➔ Has room for 80 PCs and 2 modular HP switch
- ➔ HP switch has:
  - ➔ 5 *blades* each with 8  $100\text{Mbit s}^{-1}$  ports = 40 possible connections
  - ➔ 2 *blades* each with 2  $1\text{Gbit s}^{-1}$  ports to the BigIron switches
- ➔ 3 inch gap between each column of 40 PCs to provide a chimney for exhaust heat

# The heat/power issue

- 55W output heat per PC
- 70W per switch
- ⇒ 4.5kW per rack i.e. 2 boiling kettles
- ⇒ 288kW for 64 racks
- ⇒ 0.3MW per site

# Each PC has...

- ➔ 2 40-80Gb ATA/IDE hard disks
- ➔ 256Mb RAM
- ➔ 533MHz Celeron - 800MHz Pentium III processor
- ➔ Runs Redhat Linux OS
- ➔ Is upgraded for disk capacity/processor/DRAM every 2-3 months

# The Cost Issue

- ➔ 80 PCs across 64 racks = 5120 machines
- ⇒ every \$100 of extra cost at the PC level gives \$512,000 per site
- ⇒ \$2,000,000 extra cost across 4 sites
- ➔ In March 2000:
  - ➔ 800MHz PIIIs were \$800
  - ➔ 533MHz Celerons were \$200
- ➔ i.e. 50% faster for 4 times the cost
- ➔ ...or \$3,000,000 per site of upgrade cost
- ➔ By November 2000, 800MHz PIIIs were \$200

# The Site Cost

- ➔ \$1500 per PC (x 5120)
  - ➔ \$1500 per HP switch (x 128)
  - ➔ \$1600 per rack (x 64)
  - ➔ \$100,000 per BigIron switch (x 2)
- = \$8.17 million per site
- + \$5.12 million of upgrades per year to keep site effective (assume  $2 \times \$500$  per PC per year)