



# 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

# Course Details

- ➔ Course title: Advanced Computer Architecture
- ➔ Course code: 332
- ➔ Syllabus
  - ➔ Basic cluster performance analysis; PageRank algorithm
- ➔ Learning Objectives
  - ➔ be able to perform high-level mean performance analysis of a cluster
  - ➔ understand how PageRank algorithm measures a website's popularity
  - ➔ know how to perform an eigenvalue calculation to calculate a PageRank value

# Books

- ➔ Computer Architecture: A Quantitative Approach. Hennessy and Patterson. 3rd Edition. Morgan Kaufmann 2003.
- ➔ Probability and Statistics with Reliability, Queuing and Computer Science Applications. K.Trivedi. 1st/2nd Edition. Wiley 1980/2002.

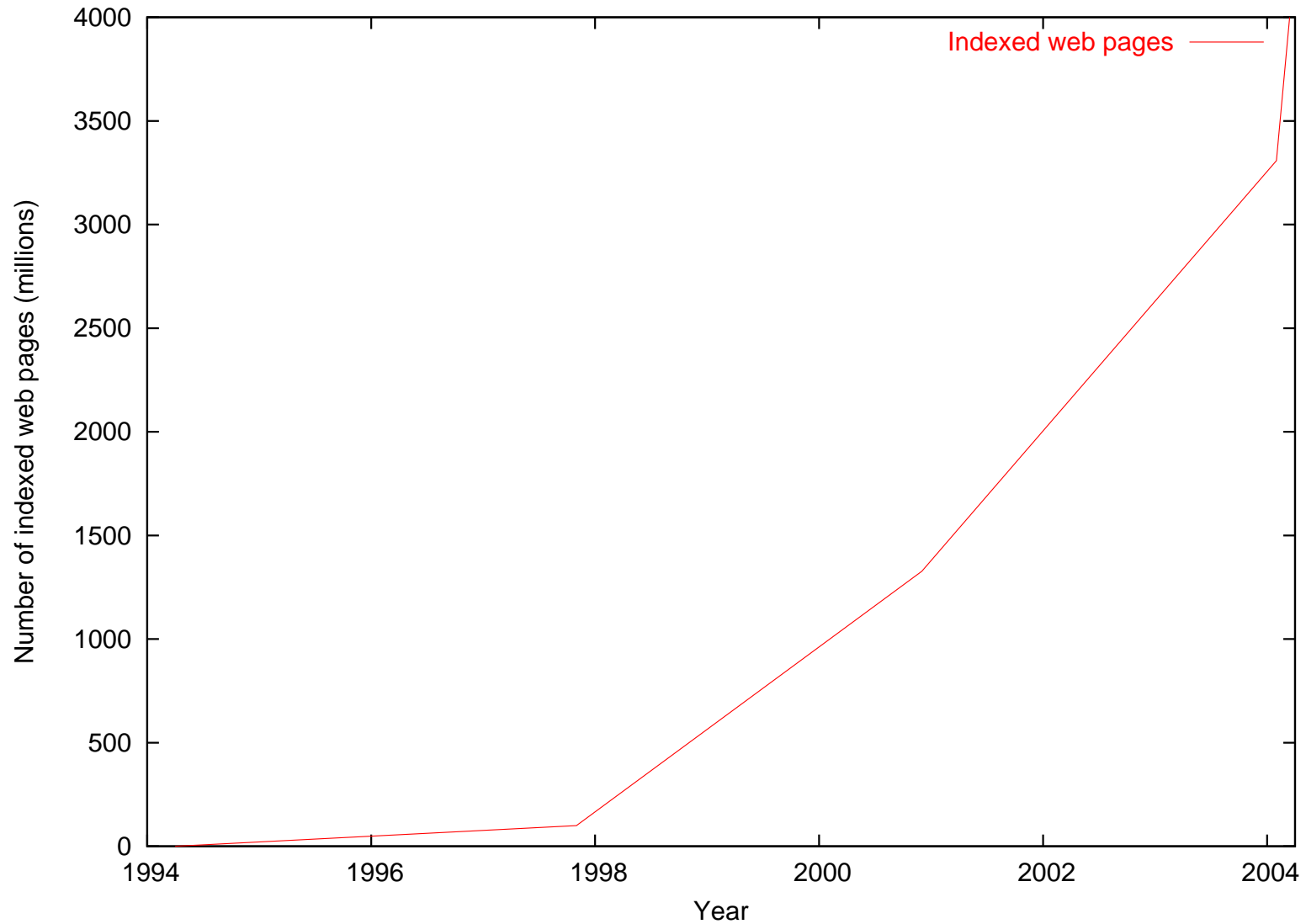
# Challenges for Google

- ➔ Google (or any mainstream internet search engine) has to cope with three major problems:
  - ➔ phenomenal internet growth rate
  - ➔ unstructured information storage
  - ➔ no quality guarantees on web-published data
- ➔ Solves these with:
  - ➔ several enormous cluster computers
  - ➔ the PageRank algorithm

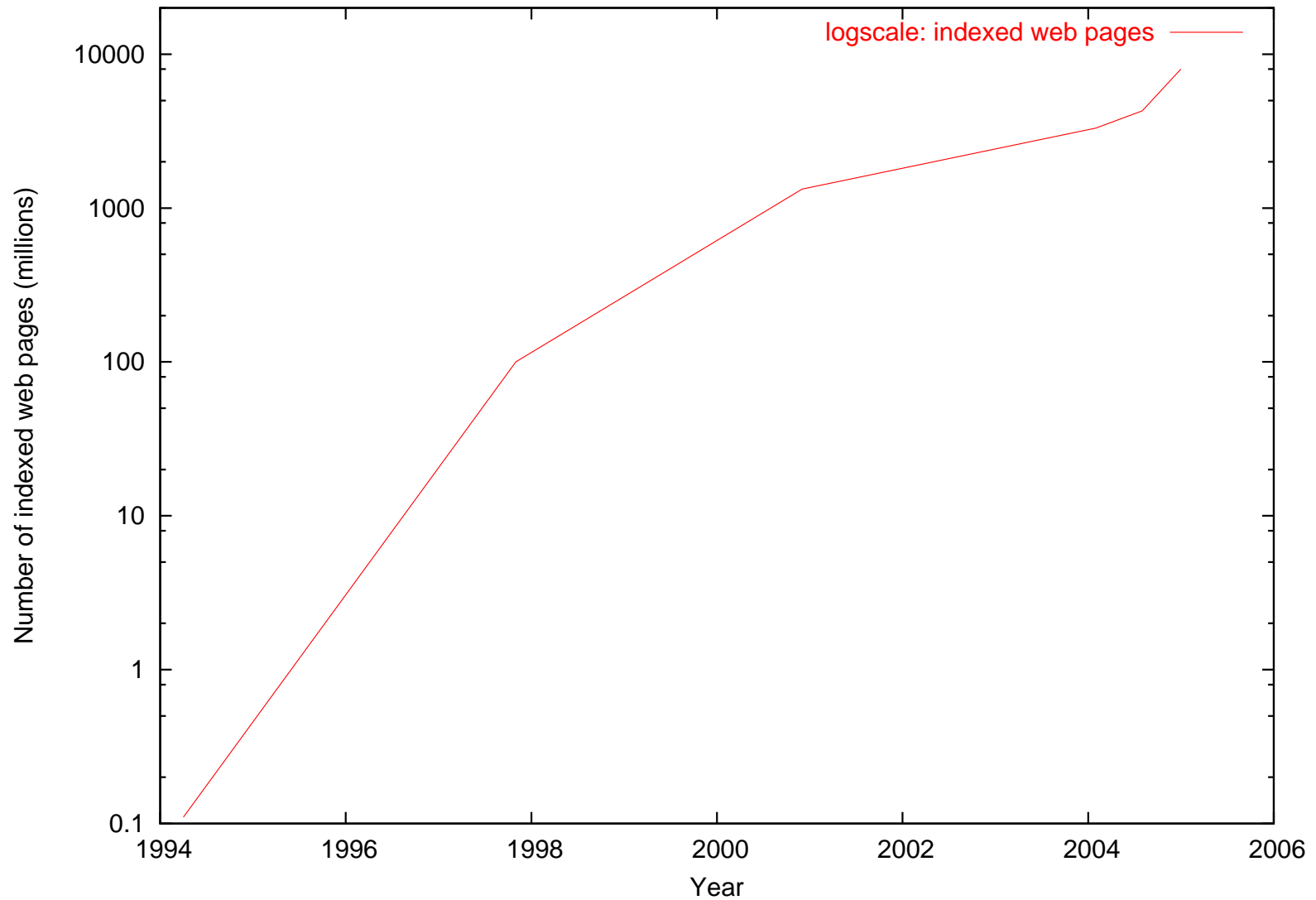
# Growth Rate

- ➔ The web is big... very big
- ➔ Probably  $5 \times 10^9$  to  $6 \times 10^9$  pages (2003)
- ➔ ...and still growing at conservatively  $> 10\%$  per year
- ➔ Sedate compared 1994–98 500% annual growth rate

# Internet Growth



# Internet Growth



# Information storage

- ➔ In contrast to information stored in a traditional database, the internet stores information with:
  - ➔ Ad-hoc data publishing
  - ➔ Semi-random underlying graph structure
  - ➔ Heterogeneous data types
  - ➔ No authoritative index or design