How Works

by

David Lazarus

Vladimir Iovu

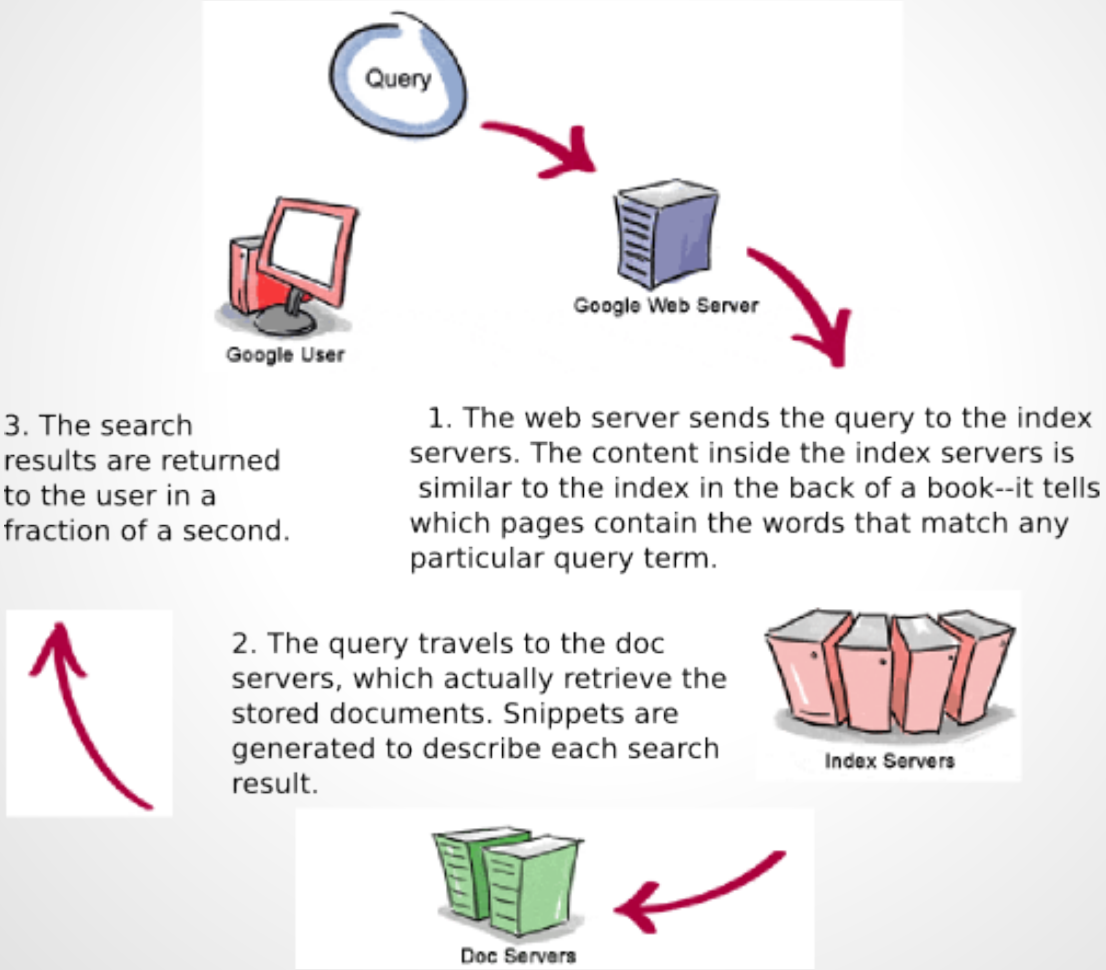
Table of Contents

- Introduction and History
- Data Centers
- Technical Aspects
- Google Search
  - Googlebots
  - Indexing
  - Query Processing
  - Spam
  - Result Delivery
    - PageRank
- Summary

Introduction

- What is Google?
- History
  - Founded in 1996 by Larry Page and Sergey Brin
  - Incorporated in 1998
  - Initial Public Offered in 2004
- Interesting Facts
  - Most visited website in the world
  - Yahoo! relied on Google searches for nearly four years until developing its own search engine technologies in 2004
  - Runs on over 1 million servers
  - Processes over 3 billion search queries a day
  - Revenue of $37.9B in 2011

How Google Works



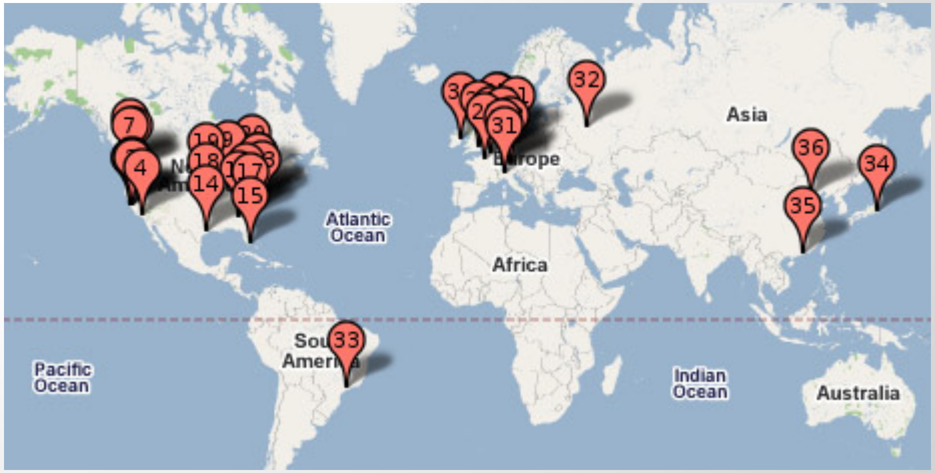
1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book--it tells which pages contain the words that match any particular query term.

2. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.

3. The search results are returned to the user in a fraction of a second.

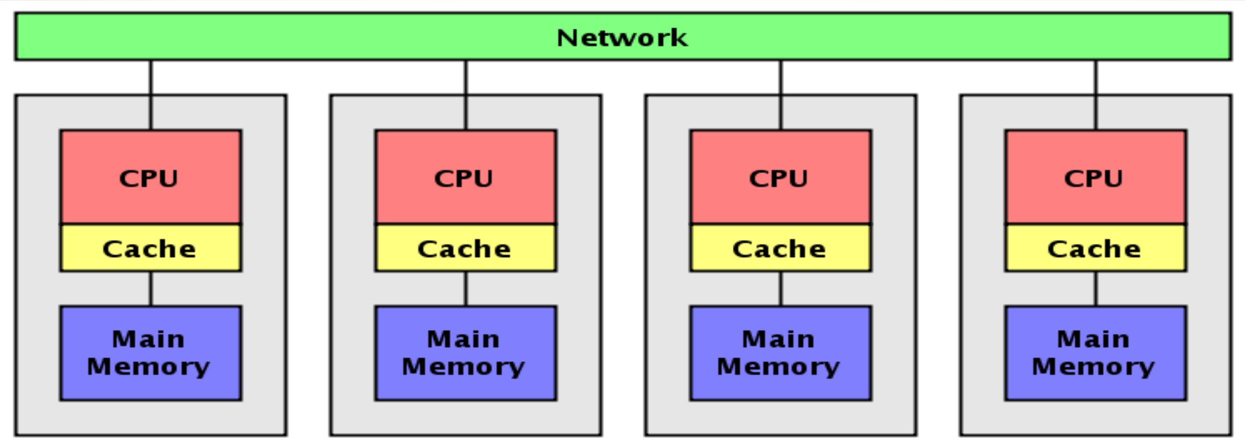
Google's Data Centers

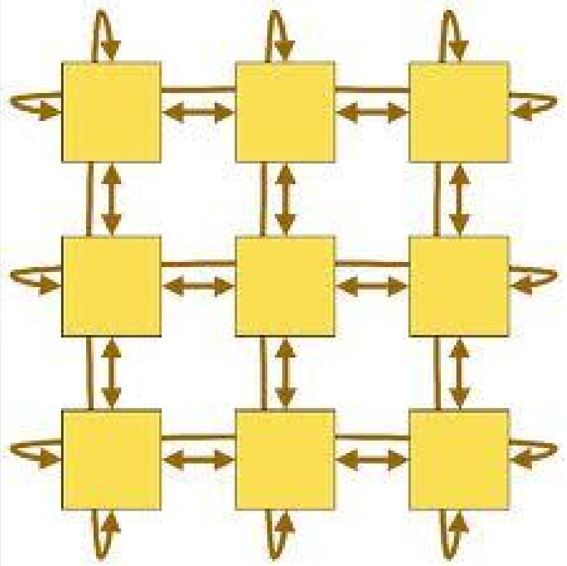
- Very secretive
- 19 locations in the US, 17 around the world
- Up to 500,000 square feet and $600 million each
- Use between 50-100 megawatts of power and often found near water
- Highly energy efficient, aiming for carbon neutrality



Google's Data Centers

How Google Works





Parallel Computing

Google Structure: Part 1 - Googlebot

- Google's Web crawler that finds and fetches web pages
- More like a web browser
- Slows down when sending requests to the servers

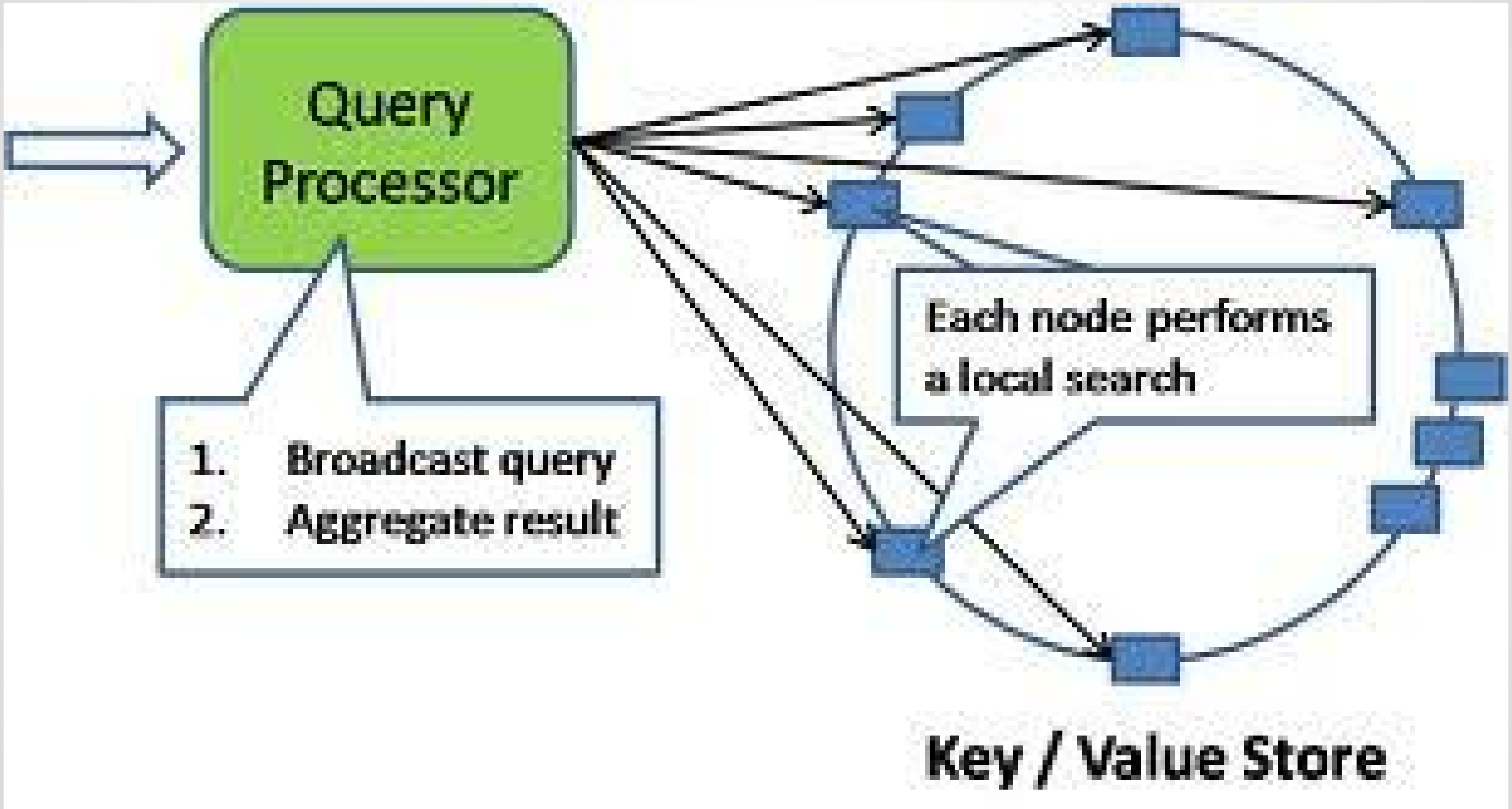
Part 1 Googlebot Continue. . .

- Googlebot finds pages two ways:
  - Through an add URL form (*i.e. google.com/addurl.html*)
  - Through finding links by crawling the web

Google Structure: Part 2 - The Indexer

- Sorts every word on every page
- Stores the resulting index of words in a huge database
- Discards stop words (*i.e. is, on, or, of, as*)
- Converts all letters to lowercase to improve performance

Google Structure: Part 3 - Query Processor



How Google filters spam

- Review and refine algorithms
- 10,000+ remote testers are used
- Solicits spam reports from users
- Pirated works are taken down

Google's PageRank

- Google uses a trademarked algorithm called PageRank, which assigns each Web page a relevancy score.
- A web page's PageRank depends on a few factors:
    - The frequency and location of keywords within the Web page
    - How long the Web page has existed
    - The number of other Web pages that link to the page in question

Summary

- User types in query
- Google sends query out to hundreds of machines
- Finding the right balance between word proximity, page reputation and links pointing to it is the key to finding the top results.
  - PageRank!
- These results are returned to the user, showing a snippet of each page
- All in under half a second

References

http://www.google.com/about/datacenters/

http://ppcblog.com/how-google-works/

http://computer.howstuffworks.com/internet/basics/google.htm

http://www.googleguide.com/google_works.html

How Google Works

by

David Lazarus

Vladimir Iovu