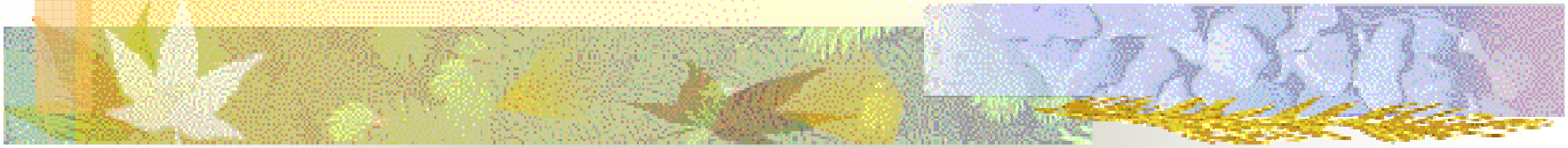
Arquitectura de Google



Universidad de Costa Rica
Escuela de Ciencias de la Computación e Informática
M.Sc. Kryscia Daviana Ramírez Benavides

Introducción

- Google fue fundada el 7 de septiembre de 1998 por Larry Page y Sergey Brin.

- Google se enfocó en:
  - Poner entre las primeras diez páginas lo que el usuario promedio está buscando.
  - Construir un sistema que la mayoría de las personas puedan utilizar sin problemas.
  - Guardar todos los documentos que se encuentren en el rastreo.

Introducción (cont.)

Primera oficina de Google.
Consiguió hacer funcionar varias máquinas totalmente diferentes con un impresionante rendimiento.

Una máquina hecha a medida, con techo de LEGO.
Go lego!!! ⇒ Google!!!

Tomado de http://royal.pingdom.com/2009/03/02/original-google-setup-at-stanford-university/.

Características

- Sus principales ventajas se deben a que es muy rápido, y sus resultados son relevantes y bastante bien ordenados.
- Para jerarquizar sus páginas utiliza diversos factores tales como modelo vectorial, texto de enlaces, *Page Rank*.
- Google analiza más de 100 factores para determinar la relevancia de una página Web.
  - Entre ellos, destacan el texto del enlace (*anchor text*), el tamaño de la fuente y la proximidad.

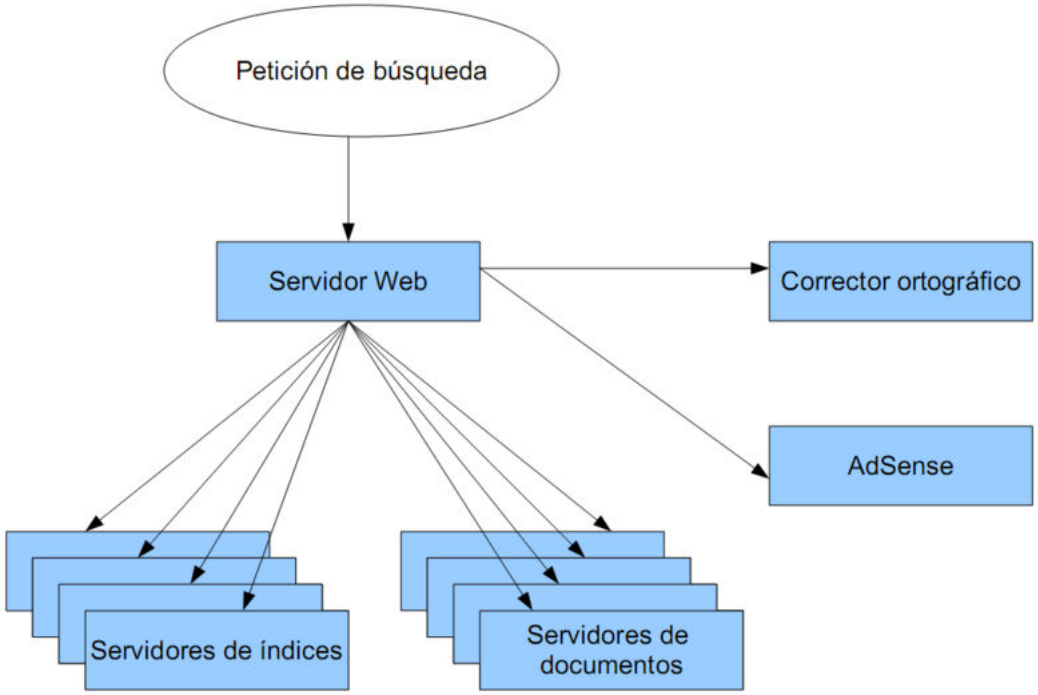
Características (cont.)

- Google indexa más de 3 mil millones de páginas Web, aunque ofrecen más resultados gracias a los “rastreos profundos”.

- Hay varios “rastreadores” (*crawlers*):
  - El general (una vez al mes), que busca en la mayoría de la WWW.
  - El *Fresh*, que rastrea en las páginas que se actualizan frecuentemente.
  - El de noticias, que rastrea cada 10 minutos.

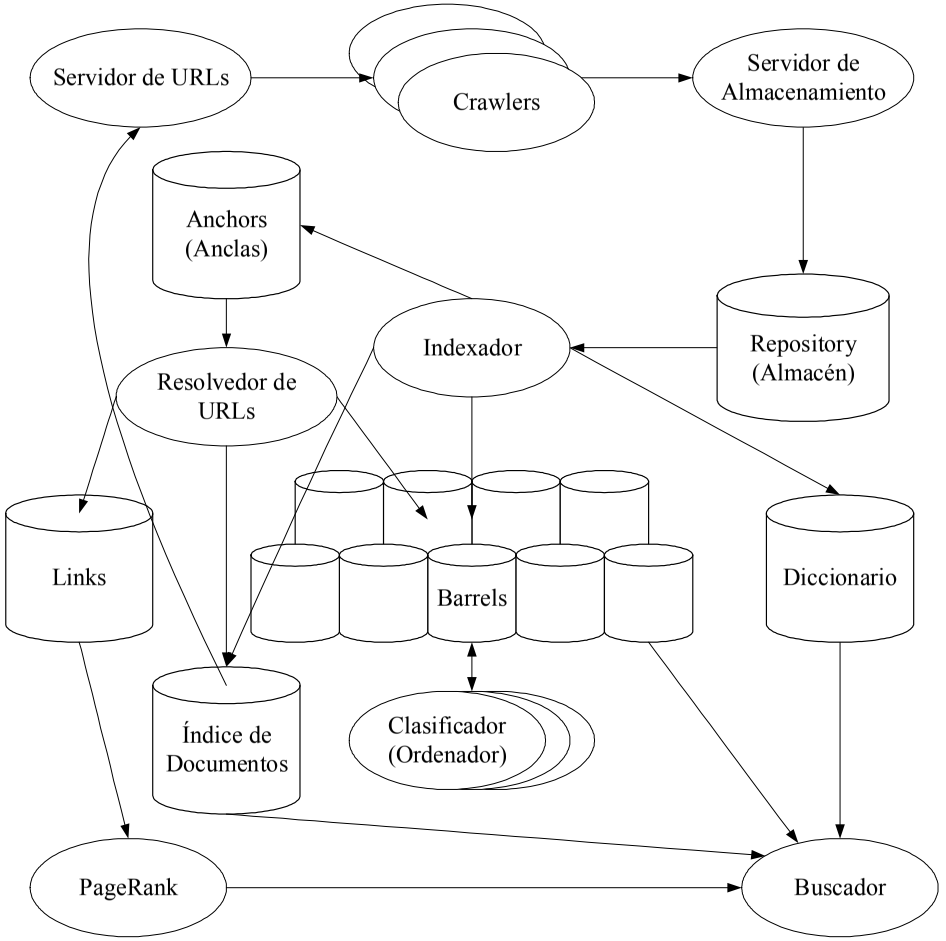
Características (cont.)

- Hay 4 tipos de servidores en el clúster de Google, situados en paralelo del servidor Web:



Tomado de http://es.wikipedia.org/wiki/Plataforma_de_Google.

Arquitectura de Google



Tomado de http://infolab.stanford.edu/~backrub/google.html.

Hardware

- Especificaciones del *hardware* del año 2003:
  - Más de 15.000 servidores con velocidades comprendidas entre el Intel Celeron de 533 MHz y el Pentium III a 1,4 GHz dual (a fecha de 2003). Según Paul Strassman, Google tendría en 2005 unos 200.000 servidores mientras que algunas fuentes indican que el número de servidores podría haber alcanzado los 450.000 en 2006.
  - Uno o más discos duros de 80 GB por servidor (en 2003).
  - Entre 2 y 4 GB de memoria por máquina.

Hardware (cont.)

- El tamaño exacto de los centros de datos que Google utiliza es desconocido, y las cifras oficiales se mantienen poco precisas intencionadamente.

- Según una estimación del año 2000, la granja de servidores de Google estaba compuesta por 6000 procesadores, 12.000 discos duros IDE (dos por máquina).
  - Cada centro tenía una conexión de fibra óptica de 2488 Mbit/s y otra de 622 Mbit/s.
  - Los servidores ejecutan un software llamado Google Web Server.

Hardware (cont.)

- Actualmente Google está desarrollando un supercomputador en un centro de datos en Dallas.

- El proyecto se llama **Proyecto O2** y se espera que incremente sustancialmente la capacidad de su red global actual, permitiendo ejecutar miles de millones de búsquedas al día y un catálogo de otros servicios que cada vez crece más.

Topología de Red

- Se estima que Google mantiene más de 450.000 servidores, ordenados en *racks* de *clusters* en varias ciudades del mundo.

- Es por eso que Google puede ofrecer un servicio más rápido a los usuarios.
  - En el año 2005 Google había indexado 8.000 millones de sitios Web.

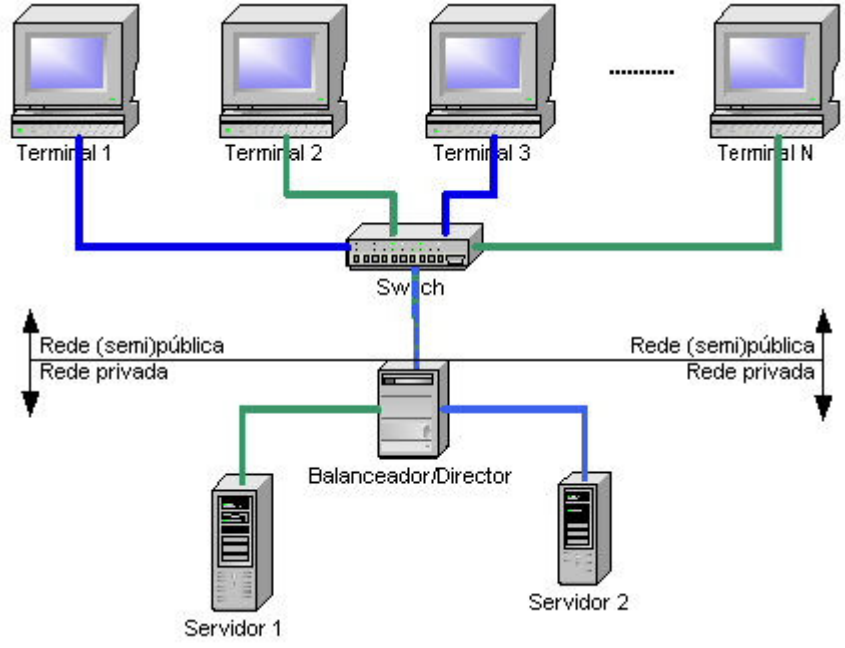
- Cuando se hace conexión a Google, los servidores DNS traducen la dirección *www.google.com* a varias IP's distintas, permitiendo que se distribuya la carga entre varios *clusters*.

Topología de Red (cont.)

- Cada *cluster* tiene miles de servidores.

- Los *racks* de Google están hechos a medida y pueden contener entre 40 y 80 servidores.
  - Cada *rack* tiene una conexión *ethernet* a un *router* local que a su vez se conecta al *router* central utilizando una conexión de 1 Gigabit.
  - Un *rack* es algo así como: 88 dual-CPU 2Ghz servidores Intel Xeon con 2Gbytes de RAM y un disco duro de 80Gbytes.

Topología de Red (cont.)



Arquitectura típica de un balanceador de carga.

Un centro de datos donde se pueden ver varios *racks*.

Tomado de http://es.wikipedia.org/wiki/Plataforma_de_Google.

Operaciones de los Servidores

- La mayoría de operaciones son de solo lectura.
- En la actualización de datos, las consultas se envían a otros servidores, para simplificar los problemas de consistencia.
  - Las consultas se dividen en subconsultas y se envían por diferentes canales en paralelo, reduciendo así el tiempo de latencia.
- En los fallos de hardware se utiliza tecnología RAID.
  - El software también está diseñado para gestionar los fallos.
  - Cuando un servidor se cae, los datos todavía están disponibles en otros servidores.

Implementación

- Los lenguajes de programación utilizados son:
  - La amplia mayoría de los módulos que componen la arquitectura están implementados en C y C++.
  - Ejecución sobre Solaris y Linux.
  - Los *Crawlers* y el Servidor de URLs están implementados en Perl.

Referencias Bibliográficas

- La información fue tomada de:
  - http://es.wikipedia.org/wiki/Google.
  - http://en.wikipedia.org/wiki/Google_platform.
  - http://www.maxglaser.net/arquitectura-original-de-google/.
  - http://www.promocionarweb.com/google/arquitectura.htm.
  - http://www-gist.det.uvigo.es/~martin/nst/google.pdf.
  - http://google.dirson.com/.
  - http://royal.pingdom.com/2009/03/02/original-google-setup-at-stanford-university/.