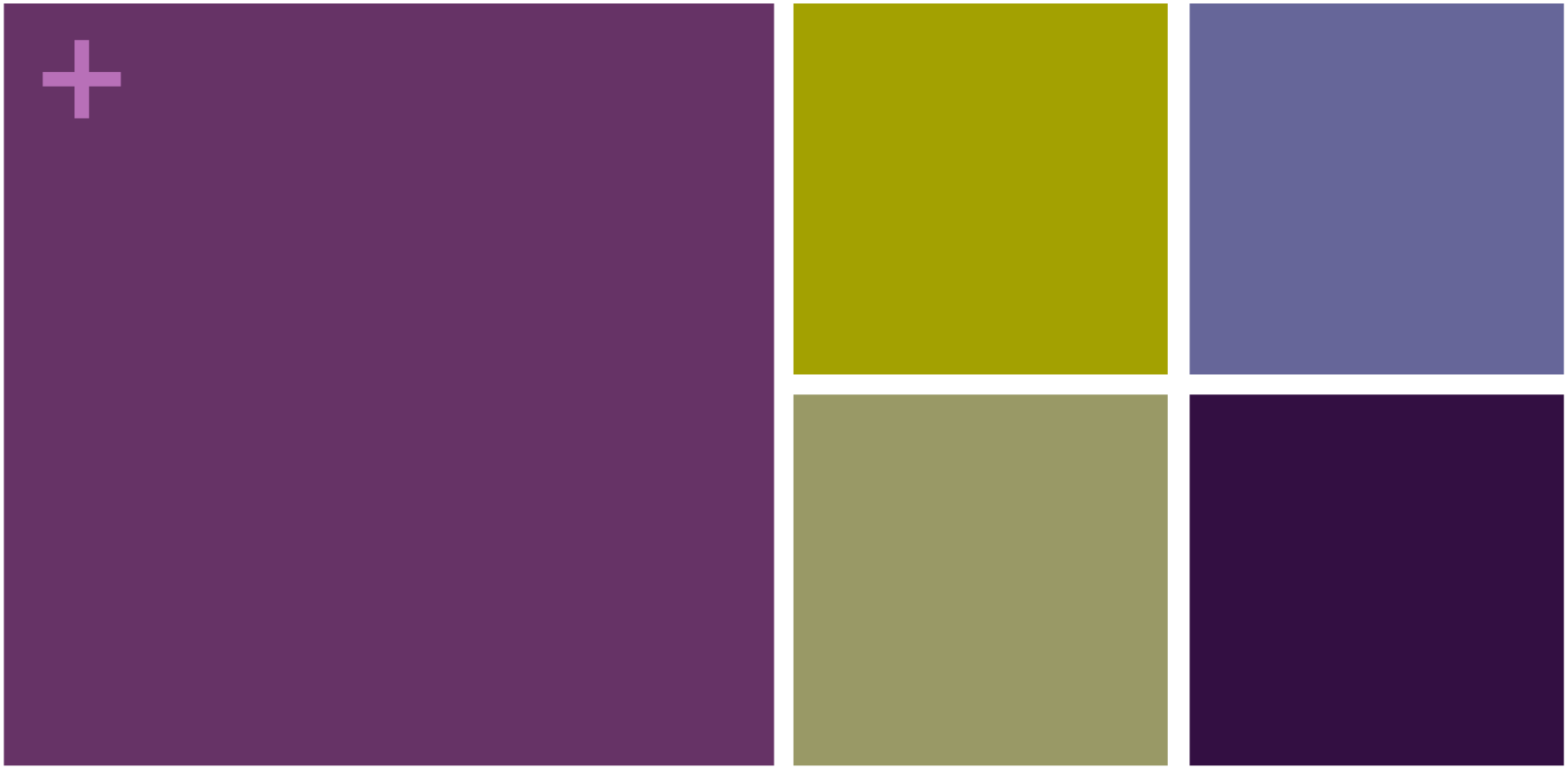


Sistemas Distribuidos

Diseño de Sistemas Distribuidos: Google

Rodrigo Santamaría

Diseño de Sistemas Distribuidos

- Introducción: Google como caso de estudio
- Servicios
- Plataforma
- Middleware

Introducción

- Crear un sistema distribuido no es sencillo
  - **Objetivo**: obtener un sistema consistente que cumpla con los requisitos identificados
    - Escala, seguridad, disponibilidad, etc.
  - Elección del **modelo**
    - Tipo de fallos que asumimos
    - Tipo de arquitectura
  - Elección de la **infraestructura**
    - Middleware (RMI, REST, Kademlia, etc.)
    - Protocolos existentes (LDAP, HTTP, etc.)

Introducción

Google como caso de estudio

- Google es un ejemplo de un diseño distribuido exitoso
  - Ofrece búsquedas en Internet y aplicaciones web
  - Obtiene beneficios principalmente de la publicidad asociada
- **Objetivo**: *"organizar la información mundial y hacerla universalmente accesible y útil"*
- Nació a partir de un proyecto de investigación en la Universidad de Stanford, convirtiéndose en compañía con sede en California en 1998.
- Una parte de su éxito radica en el algoritmo de posicionamiento utilizado por su motor de búsqueda
  - El resto, en un sistema distribuido eficiente y altamente escalable

Introducción

Desafíos en Google

- Google se centra en cuatro desafíos
  - **Escalabilidad**: un sistema distribuido con varios subsistemas, dando servicio a millones de usuarios
  - **Fiabilidad**: el sistema debe funcionar en todo momento
  - **Rendimiento**: cuanto más rápida sea la búsqueda, más búsquedas podrá hacer el usuario -> mayor exposición a la publicidad
  - **Transparencia**: en cuanto a la capacidad de reutilizar la infraestructura disponible, tanto internamente como externamente (plataforma como servicio)

Introducción

Desafíos: escalabilidad

- Google ve estos desafíos en términos de tres dimensiones
  - *Datos*: el tamaño de Internet sigue creciendo (p.ej. por la digitalización de bibliotecas o los contenidos en redes sociales)
  - *Peticiones*: el número de usuarios sigue creciendo
  - *Resultados*: la calidad y cantidad de resultados sigue mejorando
- Logros
  - 1998: sistema de producción inicial
  - 2010: $88 \cdot 10^9$ millones de búsquedas al mes
  - Sin perder eficiencia: tiempo de consulta < 0.2s
- Ejemplo [Dean 2006]
  - Asumimos que la red está compuesta de ~$20 \cdot 10^9$ páginas
    - Cada una de 20KB -> tamaño total de 400TB
  - Un ordenador que lea 30MB/s tardaría 4 meses en explorarla
    - 1000 ordenadores lo harían en menos de 3h

Introducción

Desafíos: fiabilidad

- Google ofrece un nivel de fiabilidad del 99.9% en sus contratos de pago
  - Generalmente lo ha cumplido, pero en 2009, sufrió una caída de 100 minutos en Gmail durante un mantenimiento rutinario
  - Algunas caídas en servicios gratuitos, sobre todo Gmail
    - https://en.wikipedia.org/wiki/Gmail#Outages
- Para mantener la fiabilidad, es necesario anticipar los fallos (HW y SW) con una frecuencia razonable
  - Técnicas de detección de fallos
  - Estrategias para enmascarar o tolerar fallos -> principalmente mediante la replicación de la arquitectura física

Introducción

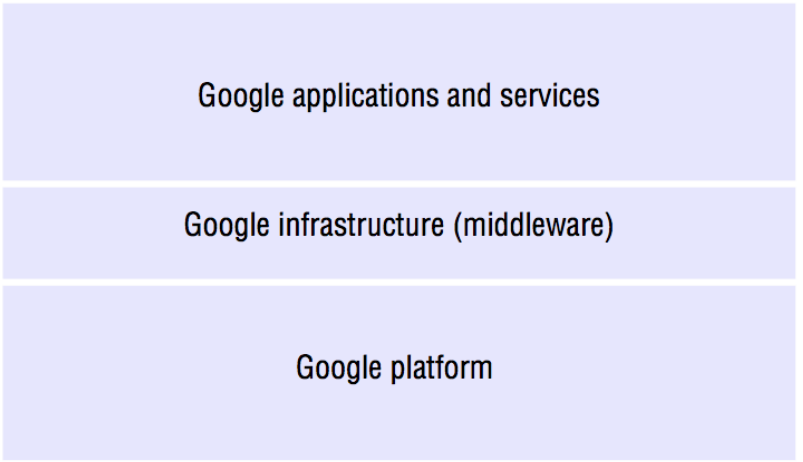
Desafíos: rendimiento

- Involucra a todas las fases de cada servicio
  - Por ejemplo, en cuanto al motor de búsqueda
    - Involucra a crawling, indexing y ranking
    - Objetivo: operaciones de búsqueda < 0.2s
- Involucra a todos los recursos subyacentes
  - Red
  - Almacenamiento
  - Procesamiento

Introducción

Desafíos: transparencia

- Requisito fundamental para el uso de la plataforma de Google como servicio, entendida como:
  - Extensibilidad
  - Soporte para el desarrollo de nuevas aplicaciones
- Para ello Google desarrolla su propia infraestructura (middleware) entre la arquitectura física y las aplicaciones y servicios



Google

Principios de diseño

- **Simplicidad**: ‘cada API tiene que ser tan sencilla como sea posible y no más’
  - Aplicación de la navaja de Occam
- **Rendimiento**: ‘cada milisegundo cuenta’
  - Medición de los costes de operaciones primitivas (acceso a disco y memoria, envío de paquetes, etc.)
- **Testeo**: ‘si no se ha roto, es que no lo has probado lo suficiente’
  - Régimen estricto de pruebas y monitorización

Diseño de Sistemas Distribuidos

- Introducción
- Servicios
  - Motor de búsqueda
  - Cloud computing
- Plataforma
- Middleware

Motor de búsqueda

- Objetivo:
  - Dada una consulta, obtener una lista ordenada de resultados relevantes, a partir de una búsqueda en la red.
- Subsistemas del motor:
  - *Crawling*: obtener información
  - *Indexing*: procesar información
  - *Ranking*: clasificar información

Motor de búsqueda

Crawling

- **Tarea**: localizar y recuperar contenidos de la Web y pasarlos al sistema de indizado.
- **Ejecución**: servicio software *Googlebot*
  1. Lee recursivamente una página web
  2. Recolecta todos sus enlaces
  3. Planifica posteriores operaciones de crawling en dichos enlaces
     - Esta técnica (*deep searching*) permite penetrar en prácticamente todas las páginas indexadas de la Web*
- La ejecución se realiza en batería. En el pasado, el Googlebot se ejecutaba una vez cada pocas semanas
  - No es suficiente para páginas de noticias o con otro tipo de contenido cambiante (blogs, redes sociales, etc.)

*Todas las páginas indexadas de la web ***no*** son toda la web
https://en.wikipedia.org/wiki/Deep_web

Motor de búsqueda

Crawling

- Con la arquitectura de búsqueda *Caffeine*, se introduce en 2010 una nueva filosofía de crawling/indexing
  - El crawler está en funcionamiento *continuo*, y tan pronto como se explora una página, se realiza su indexado.
- Este modo reduce la antigüedad de los índices en un 50%
  - Se pasa de un procesamiento por lotes global a un procesamiento continuo **incremental**.
    - Es una estrategia clásica de mejora (compresión de archivos, copias de seguridad, sistemas de versiones ...)

Motor de búsqueda

Indexing

- **Tarea**: producir un índice de contenidos de la red similar al de un libro (título, autor, tema, etc.)
- Ejecución: *indizado* inverso de palabras
  - Dada una página web (varios formatos: html, pdf, doc), se identifican características textuales clave: posición, tamaño de letra, capitalización.
  - También se realiza un índice de enlaces encontrados en la página
- Utilizando este índice, se reduce el número de páginas candidatas de miles de millones a unas decenas de miles, según el poder discriminativo de las palabras buscadas.

Motor de búsqueda

Ranking

- **Tarea**: clasificar los índices en función de su relevancia
- **Ejecución**: algoritmo PageRank
  - Basado en los sistemas de ranking de publicaciones científicas: un artículo científico es importante si ha sido citado por otros colegas del área.
    - Análogamente, una página es importante si ha sido enlazada por un gran número de páginas.
  - También tiene en cuenta factores relacionados con la proximidad de la búsqueda a las palabras claves de la página obtenidas en el indizado inverso

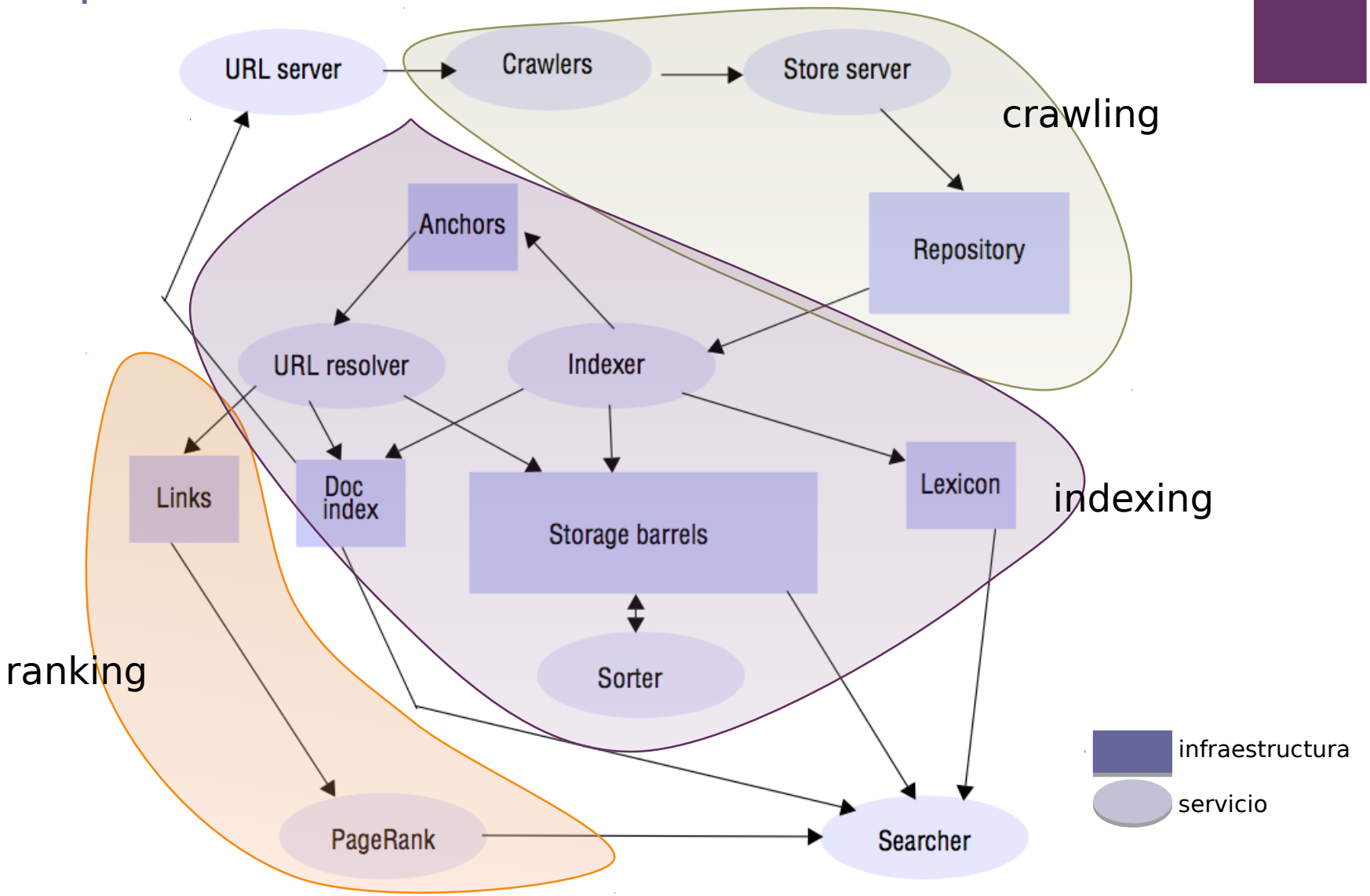
Motor de búsqueda

Ejemplo de búsqueda

- Imaginemos que buscamos “distributed systems book”
  - El buscador seleccionará páginas que contengan las tres palabras en sus índices
    - Preferentemente en el título o alguna sección importante (p. ej. listas de libros de sistemas distribuidos)
  - El buscador priorizará aquellas con un alto número de enlaces, poniendo primero las más enlazadas, y desde páginas más ‘importantes’:
    - *amazon*, *wikipedia*, *bookdepository* son priorizadas por su relevancia en la red.

Motor de búsqueda

Arquitectura



Modificado a partir de [Brin and Page, 1998]

Motor de búsqueda

Arquitectura

- *Repository*: datos web recogidos por el crawler, comprimidos
- *Barrel(s)*: almacén de *hits*, es decir, entradas que contienen el identificador del documento, la palabra encontrada en el documento, su tamaño de letra y capitalización
- *Sorter*: reordena el barrel por identificadores de palabra para generar el índice invertido
- *Anchors*: información sobre los enlaces en los documentos
- *URL resolver*
  - Preparación de los enlaces (*links*)
  - Obtención de nuevos enlaces para alimentar al Crawler (*Doc index*)

Motor de búsqueda

Arquitectura

- Los detalles específicos de esta arquitectura han variado, pero los servicios clave (crawling, indexing, ranking) siguen siendo los mismos.
- Modificaciones en
  - Almacenamiento optimizado
  - Bloques de comunicación más reutilizables
  - Procesamiento por lotes a continuo

Motor de búsqueda

Críticas

- Peligro de autocomplacencia [Maurer, 2007]
  - Google nos dice qué es importante -> no miramos más allá
- ¿importancia = nº enlaces?
  - Aquello más enlazado es más famoso, no necesariamente más relevante o interesante para el usuario
  - Discriminación de los sitios nuevos
- Peligro de manipulación
  - Neutralidad: Google prioriza sus servicios en las búsquedas [NexTag, 2011]
  - Google bombing: enlazar con un determinado texto a una página
    - Enlaces ‘miserable failure’ a páginas relacionadas con George Bush
  - Search Engine Optimization (SEO): modificar los contenidos de una página en base a los buscadores y los usuarios (~marketing)

Google

Otros servicios: Cloud computing

- Modificación de un cliente ligero que permite usar archivos y aplicaciones remotas sin perder autonomía.
- Software como servicio: Google Apps
  - Alternativa a las suites de escritorio:
    - Gmail, Google Drive, Google Sites, Google Calendar
    - Google Maps, Google Earth
- Plataforma como servicio: Google App Engine
  - Pone a disposición del usuario parte de la infraestructura distribuida de Google para sus Google Apps y su servidor de búsqueda
    - Para que pueda desarrollar sus propias aplicaciones web mediante su plataforma
    - Mediante el uso de APIs

Diseño de Sistemas Distribuidos

- Introducción
- Servicios
- Plataforma
  - Modelo físico
  - Componentes
- Middleware

Arquitectura

Modelo físico

- Utilización de gran número de PCs como base para producir un entorno para computación y almacenamiento distribuido
  - Cada unidad de PC utilizada cuesta entorno a 1000$
  - Típicamente tiene 2TB de disco y 16GB de DRAM
  - Corre con una versión reducida del kernel de Linux
- Modelo de fallos
  - Al utilizar PCs ‘de coste’, hay un riesgo de fallos
  - Según [Hennessy and Patterson, 2006]
    - 20 máquinas tienen fallos software cada día (se reinician manualmente!)
    - 2-3% de los PCs tienen un fallo hardware al año (el 95% son fallos en los discos o en la DRAM) -> Un fallo HW por cada 10 fallos SW.
  - Se diseñarán estrategias de tolerancia de fallos

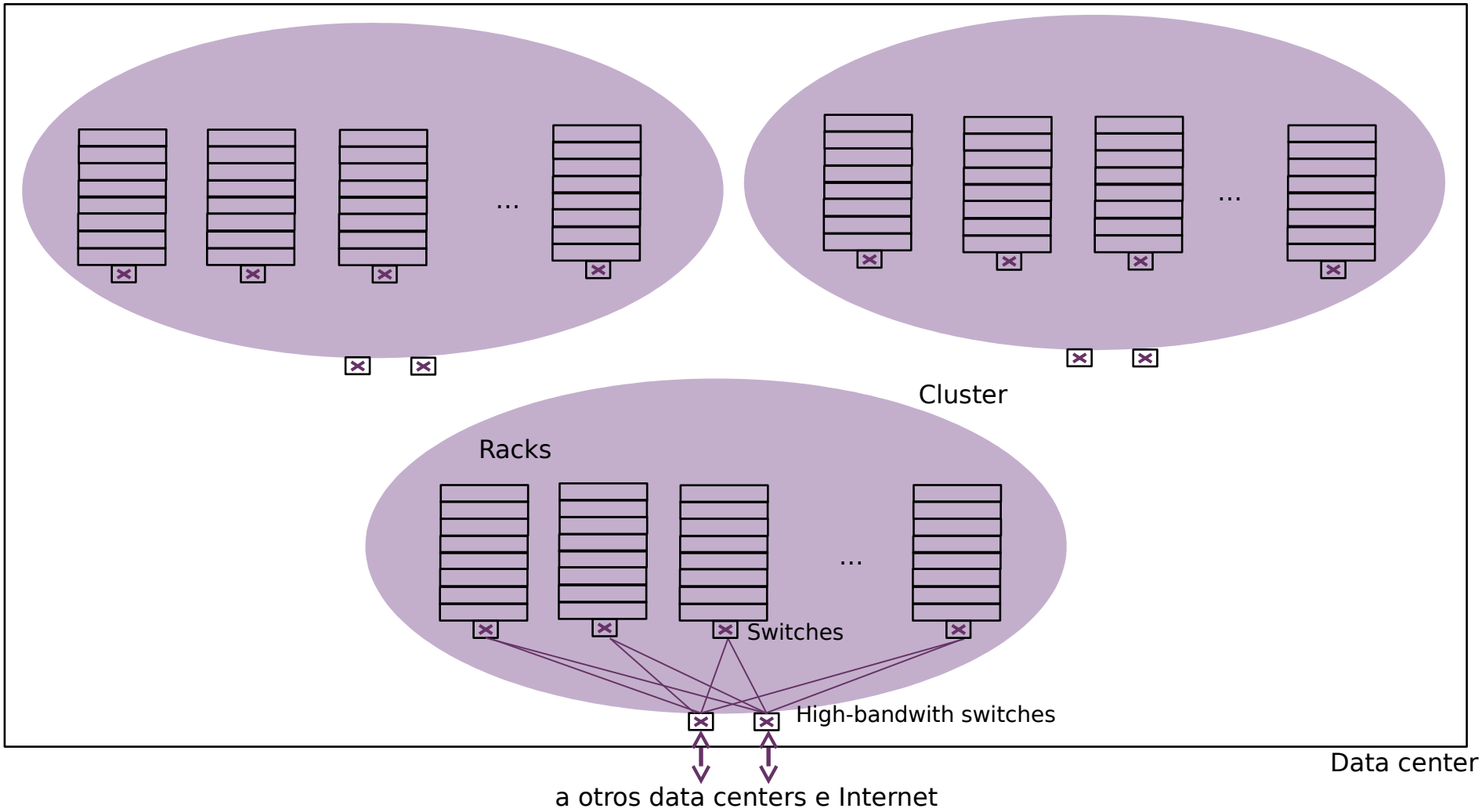
Arquitectura

Componentes [Hennessy and Patterson, 2006]

- Rack
  - Entre 40 y 80 PCs
  - Switch Ethernet que provee conexión en el rack y hacia fuera
- Cluster
  - 30+ racks
  - 2 switches de banda ancha (redundancia)
  - Unidad básica de gestión
- Data Center
  - Decenas repartidos por el mundo
  - Cada uno hospeda uno o más clusters.

Arquitectura

Componentes



Arquitectura

Data Centers

- 12 dedicados, 24 compartidos (2008)

http://royal.pingdom.com/2008/04/11/map-of-all-google-data-center-locations/
http://www.google.com/about/datacenters/locations/index.html

Arquitectura

Capacidad de almacenamiento

- Un PC -> 2 terabytes
- Un rack de 80PCs -> 160 terabytes
- Un cluster de 30 racks -> 4.8 petabytes
- Asumiendo unos 200 clusters entre todos los data centers -> 960 petabytes ~ 1 exabyte (10^{18} bytes)

Diseño de Sistemas Distribuidos

- Introducción
- Servicios
- Plataforma
- Middleware
  - Comunicación
  - Almacenamiento
  - Computación

Middleware

- Para implementar los servicios ofertados a partir de la plataforma distribuida disponible, hace falta toda una infraestructura (middleware) intermedia
- Por lo general, Google implementa sus propias soluciones middleware en vez de utilizar estándares
  - Son muy parecidas a las soluciones existentes
  - Pero están optimizadas para los requisitos de servicio y las características de la plataforma
- En esta sección veremos por encima cómo es este middleware en términos de comunicación, almacenamiento y computación distribuida

Middleware

Secciones

- **Paradigmas de comunicación**: invocación remota y multidifusión
  - *Protocol buffers*: ofrecen un formato de serialización y comunicación común
  - *Publish-subscribe*: servicio para la diseminación de eventos
- **Datos y coordinación**: almacenamiento y acceso coordinado a datos
  - *GFS*: sistema de archivos distribuido para grandes volúmenes de datos
  - *Chubby*: almacenamiento de volúmenes pequeños de datos y coordinación
  - *Bigtable*: base de datos distribuida construida sobre GFS/Chubby
- **Computación distribuida**: paralelización sobre la arquitectura física
  - *MapReduce*: computación distribuida sobre conjuntos grandes de datos
  - *Go/Sawzall*: lenguaje/biblioteca para computación distribuida

Middleware

Comunicación – protocol buffers

- Siguiendo el principio de simplicidad, Google adopta un servicio de invocación remota mínimo y eficiente
- Para ello utiliza los búferes de protocolo (*protocol buffers*), un mecanismo general para
  - Almacenamiento/marshalling, más sencillo que XML
  - Comunicación, intercambiándose mediante RPC
- Los búferes de protocolo son neutrales respecto a
  - Plataforma
  - Lenguaje
  - Protocolo RPC

Middleware

Comunicación – protocol buffers

- Utilizan un lenguaje sencillo de especificación de *mensajes*
- Cada *mensaje* se parece a un objeto:
  - Con campos enumerados
    - Secuencialmente
  - Y etiquetados
    - Política de existencia (repetido, requerido, opcional)
  - Con mensajes anidados
- Para cada campo, el compilador genera métodos de gestión
  - exists, clear, set, get

```
message Book {
    required string title = 1;
    repeated string author = 2;
    enum Status {
        IN_PRESS = 0;
        PUBLISHED = 1;
        OUT_OF_PRINT = 2;
    }
    message BookStats {
        required int32 sales = 1;
        optional int32 citations = 2;
        optional Status bookstatus = 3 [default = PUBLISHED];
    }
    optional BookStats statistics = 3;
    repeated string keyword = 4;
}
```

Middleware

Comunicación – protocol buffers

- Los búferes de protocolo son más ligeros que XML
  - No son tan abundantes en tags
    - 30-10% del tamaño en XML
  - Métodos de acceso y numeración secuencial
    - 10 a 100 veces más rápidos en operación y acceso
- La comparación no es del todo justa
  - XML tiene un propósito general
    - Tiene una semántica mucho más rica
    - Busca la interoperabilidad entre todo tipo de sistemas

Middleware

Comunicación – protocol buffers

- Los búferes de protocolo se intercambian mediante RPC con una sintaxis adicional para generar interfaces y métodos

```
service SearchService {
      rpc Search (RequestType) returns (ResponseType);
}
```

  - Interfaz de un servicio de búsqueda *SearchService*
    - Con una operación remota *Search* que toma un parámetro del tipo *RequestType* y retorna uno de tipo *ResponseType*

- Sólo se puede enviar un parámetro y retornar otro
  - Simplicidad
  - Filosofía tipo REST: operaciones simples, manipulación de recursos

- El compilador se encarga de generar los stubs para la comunicación RPC a partir del código

Middleware

Comunicación – publish-suscribe

- Ampliación de los protocol buffers para tareas de diseminación de eventos
  - RPC tiene un rendimiento bajo para este tipo de tarea
    - Necesidad de conocer la identidad de todos los destinatarios
- Google no ha hecho públicos sus detalles pero básicamente es un sistema de notificación de eventos
  - Un servicio publica los eventos que genera
  - Los suscriptores expresan su interés en determinados eventos
  - El sistema asegura la entrega de las notificaciones de eventos del servicio a sus suscriptores

Middleware

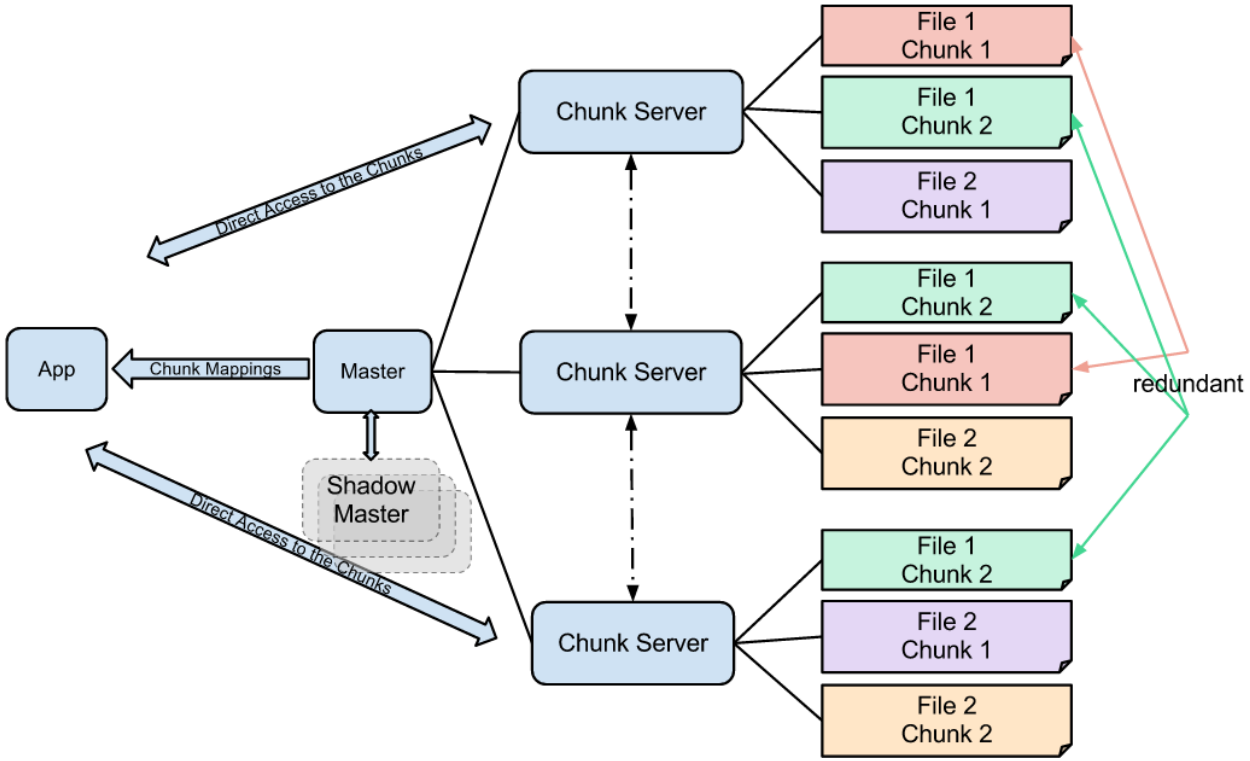
Almacenamiento - GFS

- Google File System es un sistema de archivos distribuido
  - Similar a otros de propósito general como NFS o AFS
  - Diseñado para los requisitos de Google:
    - Debe funcionar de manera fiable sobre la arquitectura física
      - Los componentes (HW y SW) pueden fallar
    - Optimizado para los archivos utilizados (de gran tamaño)
      - Millones de archivos de tamaño medio de 100MB
      - Algunos archivos superando 1GB
    - Optimizado para el tipo de consultas
      - Consultas y modificaciones secuenciales, no aleatorias
      - Nivel de acceso concurrente alto
      - Consistencia ‘relajada’ (filosofía optimista) [Gherawat 2003]

Middleware

Almacenamiento – GFS (II)

- Diseñado para replicación y para interacción entre sistemas, no entre usuarios y sistemas



https://en.wikipedia.org/wiki/Google_File_System

Middleware

Almacenamiento - Chubby

- Chubby es un servicio central de la infraestructura de Google para coordinación y almacenamiento, que provee:
  - **Acceso sincronizado** a actividades distribuidas
    - Bloqueos, semáforos, exclusión mutua, etc.
  - **Servicio de nombres** dentro de Google
  - **Servicio de archivos** para archivos de tamaño pequeño
    - Complementa a GFS
  - Servicio de **elección de réplicas** para lectura/escritura
- Puede parecer que no cumple el principio de simplicidad (*hacer una sola cosa bien hecha*) pero en el fondo, todas sus capacidades se derivan de un servicio central de ***consenso distribuido:* Paxos**

Middleware

Almacenamiento - Chubby

- Chubby comenzó como un sistema de bloqueo de archivos (concurrencia pesimista), evolucionando a una API para archivos pequeños
  - **Acceso sincronizado:** `Acquire, TryAcquire, Release`
  - **Servicio de archivos:**
    - Operaciones generales: `Open, Close, Delete`
    - Operaciones de archivo: `GetStat, GetContentsAndStat, ReadDir, setContents, setACL`
  - **Elección de réplicas primarias** en un sistema de réplicas primaria-respaldo.
    - La elección de la réplica primaria se hace en base a un consenso mediante el algoritmo Paxos*

* Distinto a, por ejemplo, el método del abusón o en anillo, donde la elección se hace en base a un criterio previamente definido (identificadores de proceso)

Middleware

Almacenamiento – Chubby/Paxos

- El algoritmo Paxos* permite consensuar de manera distribuida un determinado valor en un sistema ***asíncrono***
- Requiere un coordinador
  - Se asume que el coordinador puede fallar
  - Puede haber varios coordinadores coexistiendo
  - El coordinador decidirá el valor de consenso
- Google utiliza este algoritmo para asegurar la consistencia entre réplicas de archivos
  - El valor de consenso se refiere a la actualización
  - Los procesos son réplicas que contienen el archivo

*Ver el tema de coordinación y acuerdo

Middleware

Almacenamiento – Chubby/Paxos

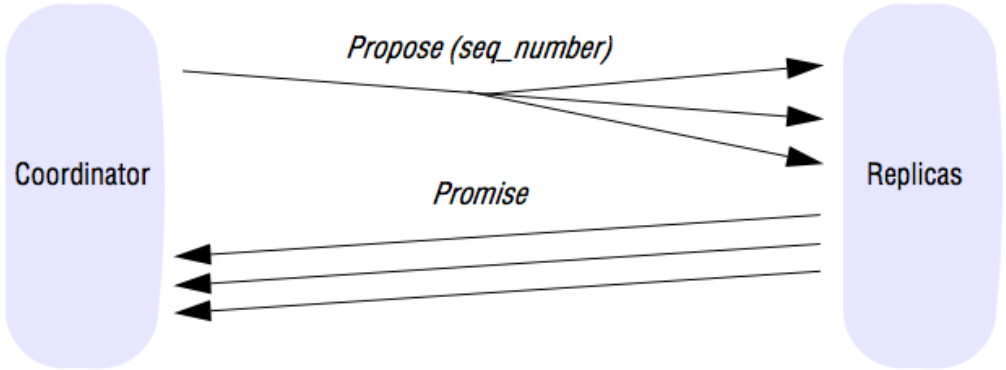
- 1) Elección de un valor
  - Sean *n* procesos (*réplicas*) con identificadores únicos i_r
  - Cada coordinador recibe un identificador único
  - En una elección, el coordinador p_r propone un nº de secuencia s'_r tal que:
    - $s'_r > s_r$ (mayor que el último valor acordado que tiene)
    - $s'_r \bmod n = i_r$ (único entre todas las réplicas)
  - p_r manda un mensaje *propose*(s'_r) a un quórum de réplicas
    - Un número mínimo suficiente de réplicas para obtener consenso
  - Las réplicas del quórum responden
    - *promise* si no han recibido un nº mayor
      - No se comprometerán con s_r menores que s'_r de ahora en adelante
    - *promise* + s''_r, si ya se ha comprometido con un s''_r menor
    - *nack* si ya se ha comprometido con un s_r mayor
  - Si el coordinador recibe suficientes mensajes *promise*
    - Determina como valor de acuerdo el mayor valor al que se haya comprometido el quórum, si existe; o el que él propuso, si no.

Middleware

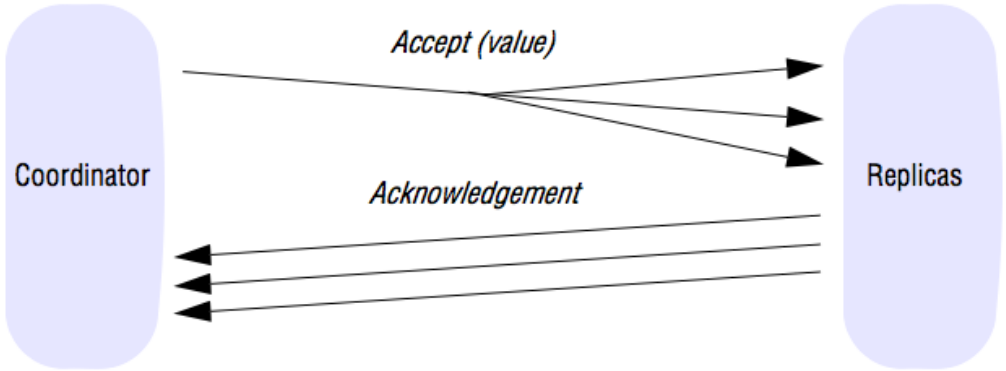
Almacenamiento – Chubby/Paxos

- 2) Diseminación del valor
  - El coordinador envía *accept* con el valor elegido al quórum
    - Todo miembro del quórum debe enviar un *ack*
  - El coordinador espera (indefinidamente) a recibir una mayoría de *acks*
    - Si una mayoría de réplicas ha hecho *ack*, se ha logrado el consenso. El coordinador hace un broadcast con el mensaje *commit* para notificar a todas las réplicas el acuerdo
    - Si no, el coordinador abandona la propuesta y el proceso reinicia

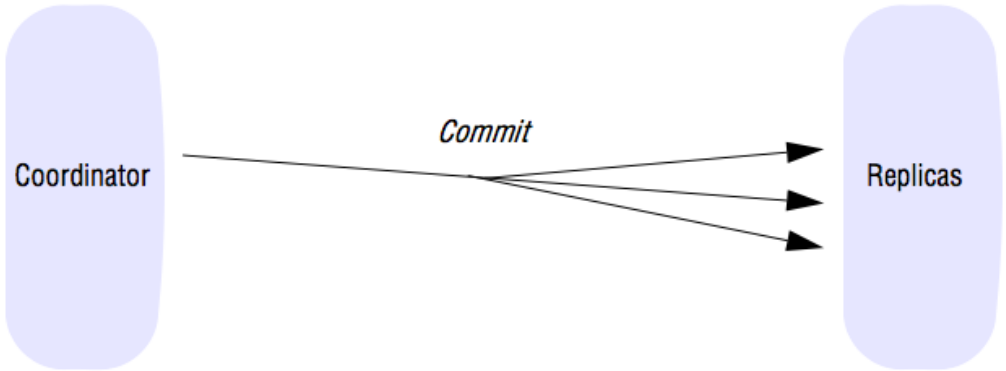
Step 1: electing a coordinator



Step 2: seeking consensus



Step 3: achieving consensus



Middleware

Almacenamiento – Chubby/Paxos

- Consenso
  - En ausencia de fallos, está asegurado
  - Con fallos (caída del coordinador o de una réplica, o mensajes perdidos/duplicados) también se asegura
    - Por la elección de números de secuencia al elegir coordinador
    - Por el mecanismo de quórum: si se acuerdan dos mayorías, debe haber al menos una réplica en común que votó por ambas
- Terminación
  - El algoritmo podría no terminar si dos candidatos compiten incrementando continuamente sus números de secuencia*

* Consistente con el teorema de imposibilidad de consenso distribuido en un sistema asíncrono con un proceso que falle (Fischer et al. 1985)

Middleware

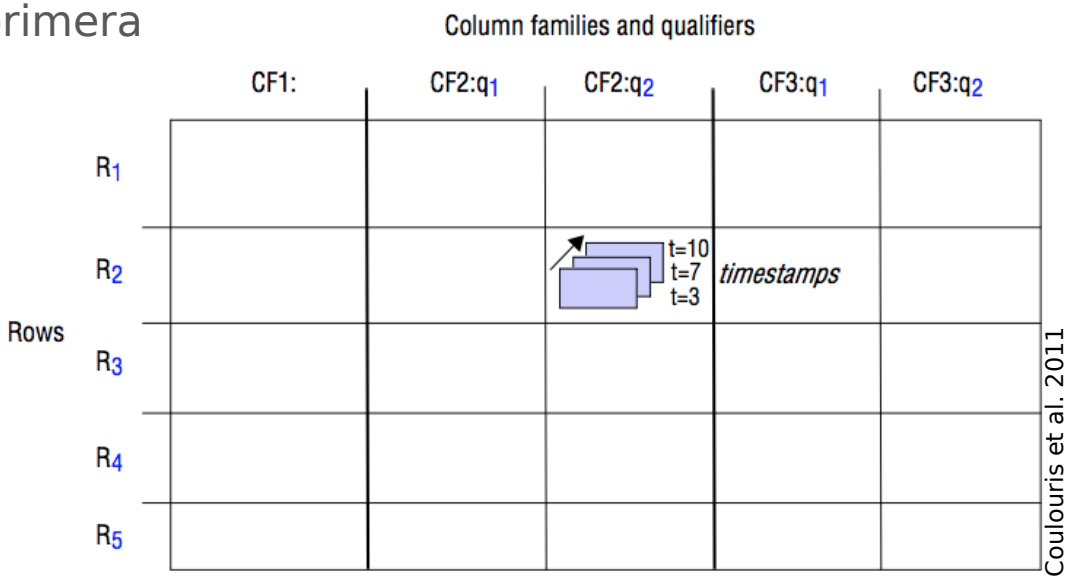
Almacenamiento – Bigtable

- GFS es un sistema de almacenamiento de archivos grandes
  - Pero ‘planos’, no indexados, accedidos secuencialmente
- Necesidad de acceso estructurado a los datos
  - Pero una base de datos distribuida es difícil de escalar y tiene bajo rendimiento, y no se necesita toda su funcionalidad
- Bigtable mantiene el modelo de tablas de una BD, pero simplifica la interfaz, centrándose en un almacenamiento y recuperación eficientes.
  - Cada tabla suele estar en el rango de los terabytes
  - Las tablas estarán almacenadas mediante GFS/Chubby

Middleware

Almacenamiento - Bigtable: interfaz

- Cada tabla se accede a través de tres dimensiones
  - Fila: indica el objeto
    - P. ej. la página web *www.bbc.co.uk/sports*
  - Columna: indica el atributo
    - P. ej. los *contenidos*, los *enlaces* o el *lenguaje*
  - Timestamp: indica la versión
    - Siendo la más reciente la primera



Coulouris et al. 2011

¿Os recuerda a alguna aproximación vista anteriormente?

Middleware

Almacenamiento - Bigtable: infraestructura

- Una bigtable ocupa varios TB
  - Para su almacenamiento, se divide en *tabletas* de 100-200MB
  - Cada tableta se almacena como un conjunto de archivos en un determinado formato
- La infraestructura de Bigtable se encarga de la división/reconstrucción en tabletas
- La infraestructura de GFS se encarga del almacenamiento de las tabletas
- La infraestructura de Chubby se encarga de la monitorización de sus estados y del balanceo de carga

Middleware

Computación - MapReduce

- La infraestructura de almacenamiento y comunicación maneja grandes conjuntos de datos distribuidos
  - Necesitamos operar con esos datos de manera distribuida
- MapReduce es un modelo de programación para aplicaciones que oculta al programador los aspectos distribuidos
  - Paralelización
  - Recuperación de fallos
  - Administración de los datos
  - Balanceo de carga

Middleware

Computación - MapReduce

- MapReduce sigue el patrón clásico en la computación en paralelo:
  - Dividir los datos (complejos) en trozos (simples)
  - Realizar el cómputo de los trozos simples (obteniendo resultados intermedios separados)
  - Unir los resultados intermedios en el resultado final
- Ejemplo: sumar todos los elementos de una matriz $10^3 x 10^3$
  - Dividimos la matriz en 100 matrices 100 x 100
  - 100 procesadores calculan las sumas parciales
  - Se hace la suma total a partir de las 100 sumas parciales

Middleware

Computación - MapReduce

- MapReduce utiliza para la paralelización en cuatro pasos
  - Divide los datos en trozos de igual tamaño
  - ***Map***: para cada trozo realiza la operación paralelizada, produciendo uno o más pares <clave, valor>
  - Se ordenan los pares intermedios según sus claves
  - ***Reduce***: se obtiene el resultado final fusionando los pares intermedios
- Ejemplo: contar el nº de veces que aparecen 1+ palabras
  - Se divide el conjunto de textos en el que buscar
  - *Map*: para cada subconjunto, cada vez que se encuentre una de las palabras buscadas, se emite un par <palabra, 1>
  - Se ordenan todos los pares según la clave (palabra)
  - *Reduce*: Se cuenta el nº de pares para cada palabra

Middleware

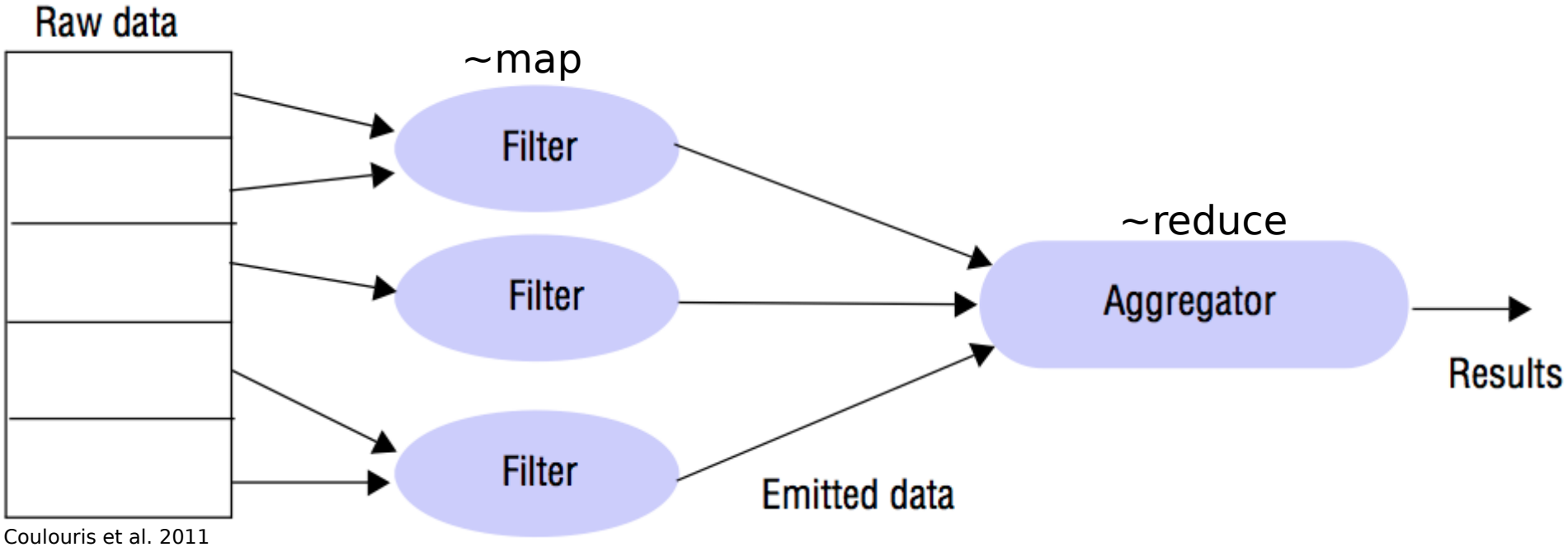
Computación - MapReduce

- MapReduce se implementa en una biblioteca que oculta los detalles asociados con la paralelización y la distribución
  - Construida sobre RPC y GFS principalmente
  - Es habitual usar como entrada y salida tablas de Bigtable
  - El programador sólo tiene que centrarse en especificar las funciones *map* y *reduce*
- Este modelo de programación ofrece varias ventajas
  - Simplifica el código de los programas
    - La fase de indizado pasó de 3800 a 700 líneas de código
  - Facilita la actualización y comprensión de los algoritmos
  - Separa la lógica de la aplicación de las tareas de distribución

Middleware

Computación - Go/Sawzall

- Sawzall es un lenguaje procedural de filtros y agregadores que utiliza la filosofía de MapReduce para ofrecer operaciones más complejas
  - Sumatorios, cuantiles, rankings, etc.
  - Programas de 10 a 20 veces más pequeños
  - Utiliza un patrón de paralelización similar



Coulouris et al. 2011

Middleware

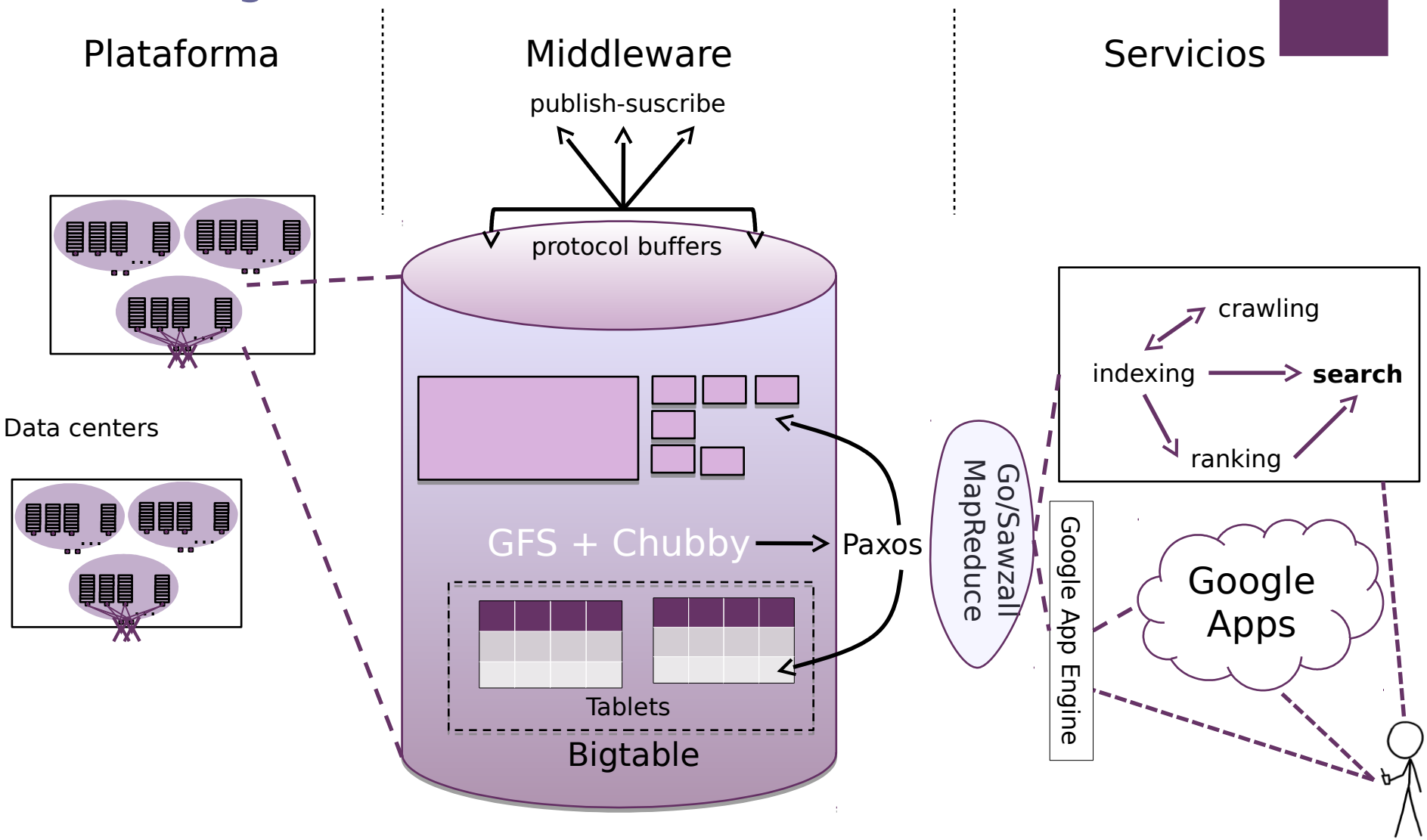
Computación - Go/Sawzall

- Sawzall muere de éxito en 2015, pues se le estaba dando uso como lenguaje de programación de propósito general.
- Se convierte en parte del sistema de logs del lenguaje de programación de propósito general Go*
  - En particular, se convierte en un conjunto de bibliotecas llamado Lingo que permite seguir usando los agregadores complejos en operaciones MapReduce

* https://www.unofficialgoogledatascience.com/2015/12/replacing-sawzall-case-study-in-domain.html

Google

Visión global

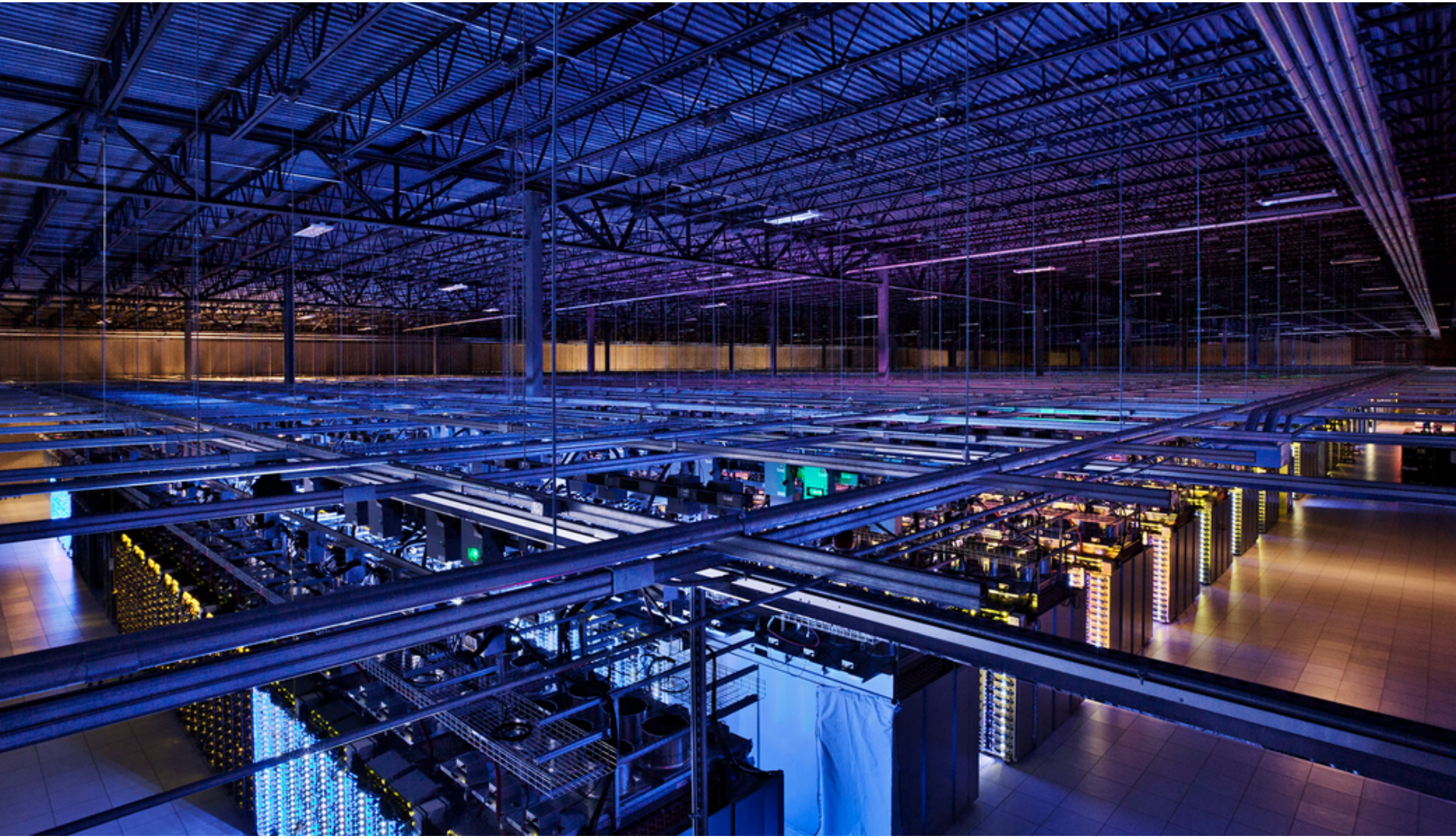


Google

- La **infraestructura** de **Google** es un buen **ejemplo** de casi todos los **problemas distribuidos**, y de su resolución mediante distintas aproximaciones
- Al diseñar un sistema distribuido es esencial
  - **Identificar** los **requisitos** primordiales para nuestro sistema
  - **Diseñar** una **plataforma** acorde a dichos requisitos
  - **Aprovechar** las **soluciones** distribuidas existentes o bien la **teoría** detrás de ellas
- Google **diseña sus propias soluciones** distribuidas **basándose en soluciones conocidas**, poniendo máxima atención en la escalabilidad y la fiabilidad.
- Para su **motor** de **búsqueda** usa tres componentes que recorren la web (**crawler**), la analizan (**indexer**) y priorizan (**ranking**)
- Para la **comunicación**, tienen una solución ad hoc basada en **RPCs** y una serialización específica más ligera que XML (**protocol buffers**)
- Para el **almacenamiento** distribuido, utilizan un sistema de archivos (**GFS**) similar a NFS pero especializado en grandes archivos. La consistencia se asegura mediante **Chubby**, una versión del algoritmo Paxos de Lamport
- Para la **computación** distribuida la solución es un **map-reduce**, que se optimiza con operaciones predefinidas de reducción (**Sawzall**)

Referencias

- G. Colouris, J. Dollimore, T. Kindberg and G. Blair. *Distributed Systems: Concepts and Design (5th Ed)*. Addison-Wesley, **2011**
  - Ch. 21
- J.L. Hennessy, D.A. Patterson. *Computer Architecture, A Quantitative Approach*. 5th ed. **2012**
- Google: http://googleblog.blogspot.com
- S. Gherawat et al. *The Google File System*. **2003**
- Algoritmo Paxos (versión original)
  - http://en.wikipedia.org/wiki/Paxos_algorithm
- Teorema de la imposibilidad de consenso
  - Fischer, M. J., Lynch, N. A., & Paterson, M. S. (1985). Impossibility of distributed consensus with one faulty process. *Journal of the ACM (JACM)*, *32*(2), 374-382.



Google Data Center en Council Bluffs, Iowa

http://www.google.com/about/datacenters/gallery/#/tech/2195/