クラウドの基盤技術
分散ファイルシステムの話

2010-02-20

増田和弘

kazuhiro.masuda at justsystems.com

自己紹介 & バックグラウンド

- 1960年生まれ
- 組込系、PC用DBMS、PC通信、全文検索など
- 業務で最初に使ったコンピュータ
  - SEIKO 9500
  - マルチCPU(8086+8087+8088+8088)
- オープンソースへの貢献はとくになし
  - ユーザ or ウォッチャー
    - ハックしていません
  - Linux, PostgreSQL, Firebird
    - 社内利用・布教

内容概要

- GoogleFileSystem
- HDFS(Hadoop)

Webスケール

- 「規模」についての最近の表現
  - ユーザ数にも使うが、
  - ここでは総データ量
- GoogleEarth(2006) で 70TB
- GoogleCache(2006)で800TB

GFS=Google File System

- 2003年公表のWhitePaper
  - “Google File System”
  - http://labs.google.com/papers/gfs.html
- 和文解説
  - 「Googleを支える技術」西田圭介,技術評論社

GFS=Google File System

- 分散ファイルシステム
  - HDD1台にもPC1台にも入りきらない巨大なデータを多数のPCで分割して持ち合い
- 冗長性・信頼性確保
  - 通常3箇所に同じデータを保存
  - 1つを故障で失えば、他のPCにコピーを作って冗長度3を保持する
- 順アクセスに性能焦点
  - 最初から最後までRead、最後に追加するWrite
  - ブロックサイズは64MBと大きい(Chunk:チャンク)
  - 非POSIX

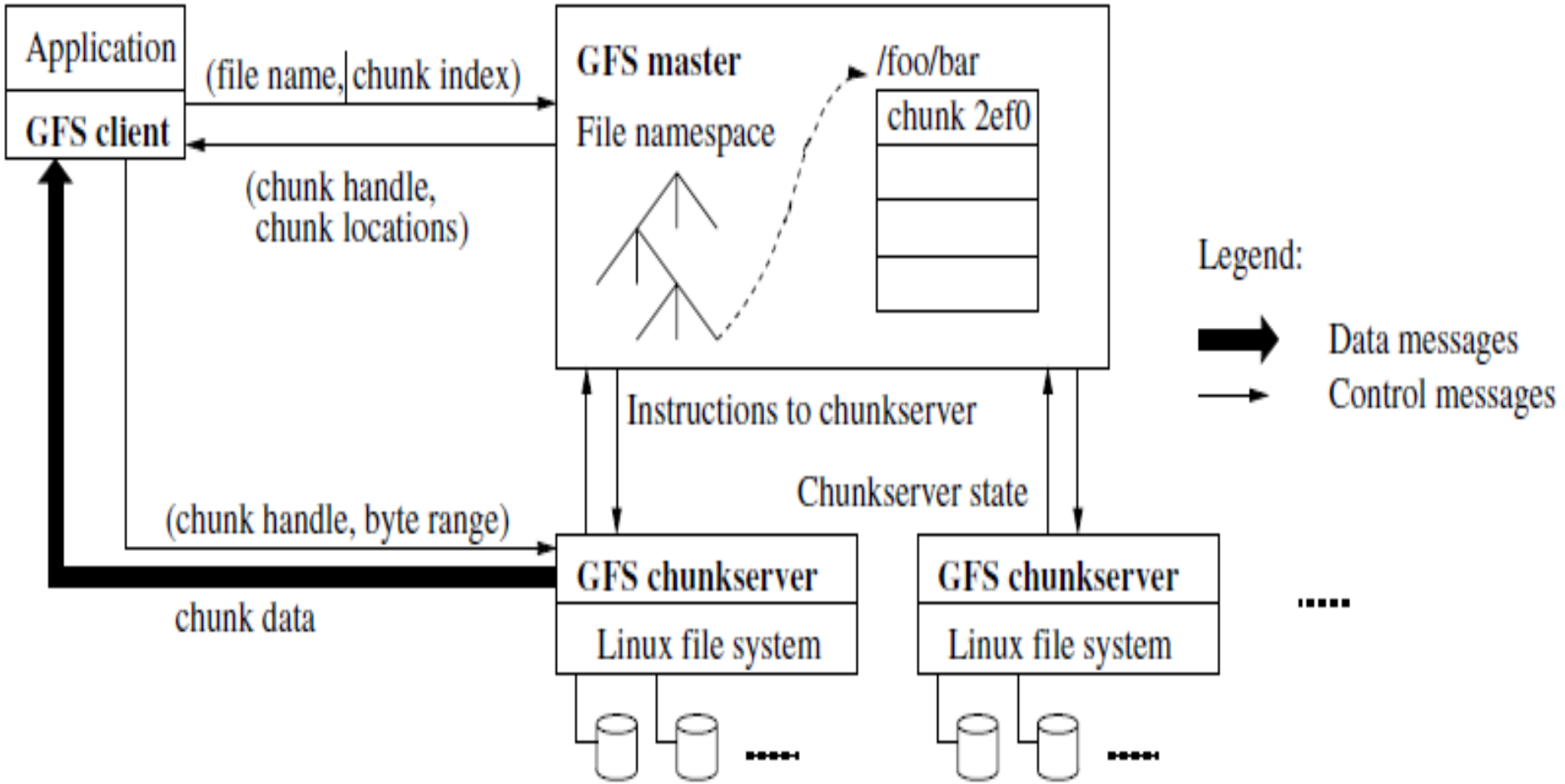

Figure 1: GFS Architecture

Chunk

- 64MBは、管理の単位
  - サイズ大
  - →全体Chunk数小
  - →マスターのメモリ消費抑制
- 転送は必要な分だけ(chunk handle, byte range)
  - Mapタスクの処理単位(Key, Value)の末尾がChunkの切れ目を跨いだとき
- Replica(複製)数3は、耐故障性確保の最小値
  - 読込性能スケールアウト目的ならもっともっと増やしていい

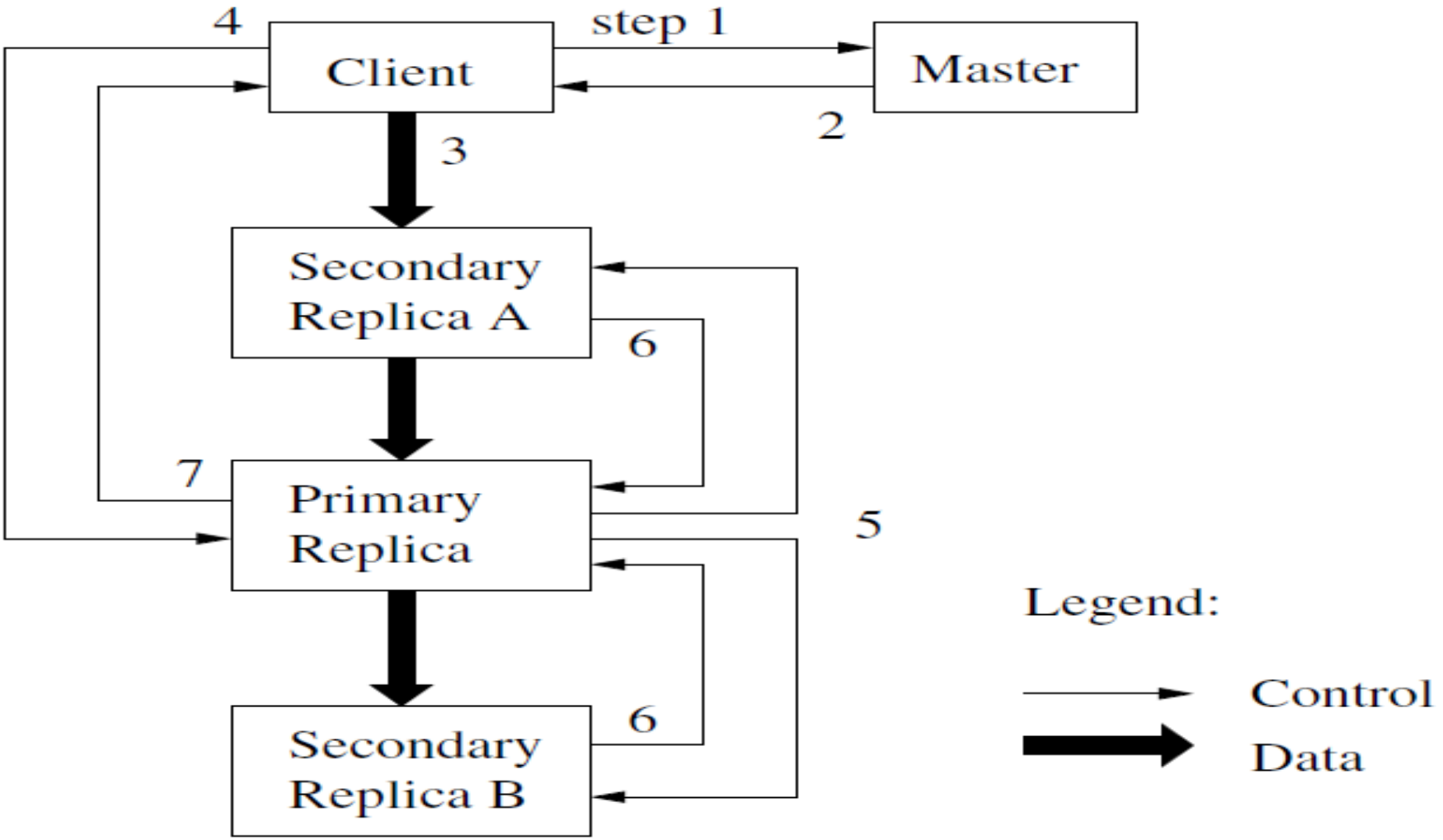

Figure 2: Write Control and Data Flow

GFSの特異なところ

- 基本　シングルマスター
  - 単純
  - 小さいNameSpace（少ないファイル数）の想定
- 一度書いたら、もう更新しない
  - コピーはメタデータをコピーするだけ
  - 更新の排他制御が要らない
  - 更新の伝播や、伝播遅延の問題がおきない
- Co-Design
  - 追記時のエラーで、重複レコードが挟まっているかもしれない
  - それをアプリケーションが許容して、重複を読み飛ばす

HDFS:Hadoop DFS

- http://hadoop.apache.org/
- 分散ファイルシステム
  - 手本：GoogleFileSystem
  - DFS(DistributedFileSystem)
- Java実装のオープンソースプロダクト

Hadoop Subprojects

- Hadoop Common: The common utilities that support the other Hadoop subprojects.
- Avro: A data serialization system that provides dynamic integration with scripting languages.
- Chukwa: A data collection system for managing large distributed systems.
- HBase: A scalable, distributed database that supports structured data storage for large tables.
- HDFS: A distributed file system that provides high throughput access to application data.
- Hive: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- MapReduce: A software framework for distributed processing of large data sets on compute clusters.
- Pig: A high-level data-flow language and execution framework for parallel computation.
- ZooKeeper: A high-performance coordination service for distributed applications.

GFS-HDFSの名前の対応

GFS	HDFS
Master	NameNode
ChunkServer	DataNode
Chunk	Block

NameNodeの情報

- HDFS NameSystem
  - フォルダ/ファイル名
  - パーミッション、サイズ、タイムスタンプ
- HDFile→(Chunkリスト)
- 各Chunk→(有効Replica数, DataNode1, DataNode2,DataNode3,...)
- 各DataNode→(保有Chunk数,空き容量,生死,...)

Hadoop のHDFSはOne of them

- Hadoopはいろんなファイルシステムで使える
  - Local
  - HDFS
  - KFS (DFS by C++),AmazonS3,etc
- 普通にDownloadすると、実はHDFSのJavadocが入ってない
  - org.apache.hadoop.hdfs
  - hadoop.fs を介して環境に応じて使い分け
- NameNode,DataNodeのJavadocも入ってない
  - org.apache.hadoop.server.*
  - ant javadoc-dev

もっと知りたい

- O'ReileyのHadoop本
- ソース, javadoc-dev
- ちょっと古いけど
  - オープンソース分散システム「Hadoop」解析資料
  - PFIとNTTレゾナントが共同調査
  - http://preferred.jp/2008/08/hadoop.html
- 注目集まっているので日本語Web記事増殖中

終わりです