Google File System / HDFS

Шугаепов Ильнур
VK.com
ilnur.shug@gmail.com

январь 2020 г.

Google File System (GFS)[1]

[1] Sanjay Ghemawat, Howard Gobioff и Shun-Tak Leung. “The Google file system”. B: (2003).

Цели

- Производительность
- Масштабируемость
- Надежность
- Доступность

Наблюдения

- Отказ компонент норма, а не исключение
- Поддерживать большое количество маленьких файлов сложно
- Большинство мутаций файлов — дописываение в конец

Предположения

- Система состоит из большого числа компонент, которые могут часто отказывать
- Система хранит преимущественно большие файлы (> 100 MB)
- Основные операции: потоковое чтение, запись в конец
- Много клиентов могут одновременно делать запись в конец файла

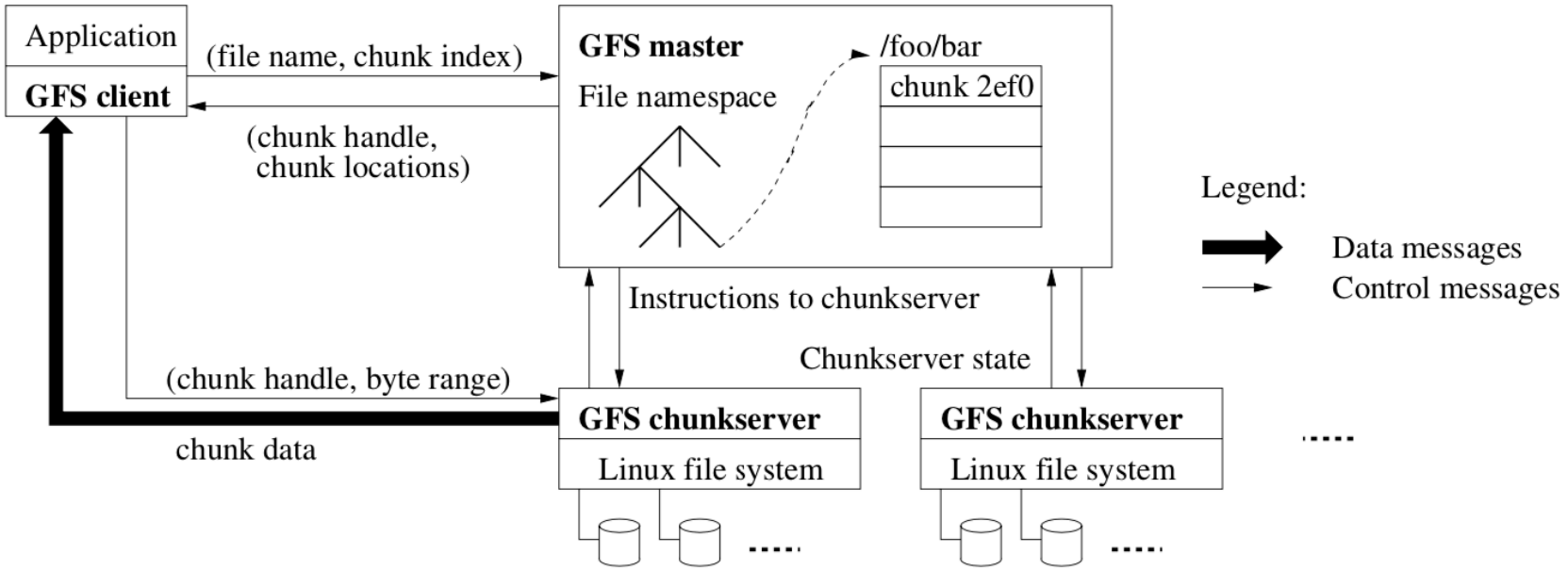

Рис.: Архитектура системы

Хранение файлов

1. Файл рабивается на куски (chunks) фиксированного размера (64 MB)
2. Каждый кусок идентифицируется глобально-уникальным `chunk_handle`, который выдается мастером
3. Куски файла раскидываются и реплицируются по разным chunkserver'ам, которые выбирает мастер

Master

- Хранит все метаданные
- Создает/реплицирует чанки
- Сборка мусора
- Общается с chunkserver’ами с помощью HeartBeat сообщений
- Обрабатывает запросы связанные с метаданными
- Ведет лог всех операций

Chunkserver

- Хранит данные на локальном диске
- Обменивается данными с клиентом напрямую
- Обменивается данными с другими chunlserver’ами

Чтение файла

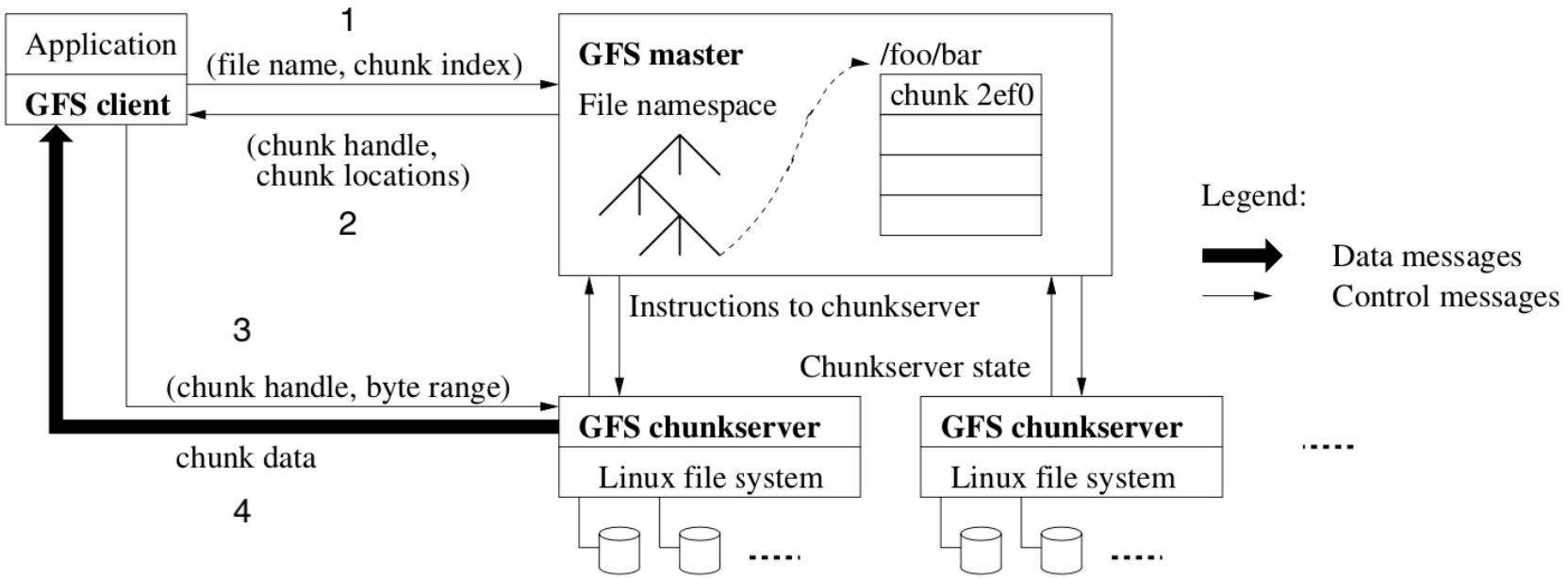

Рис.: Этапы чтения файла

Цели

- Надежность и доступность
- Для обеспечения и того и другого нужно раскидывать реплики еще и между стойками

Создание куска

Факторы влияющие на то, на каком chunkserver’е будет создан кусок

1. Утилизация диска
2. Утилизация сети
3. Как давно был создан последний кусок
4. Где находится сервер

Ре-репликация

1. Кусок ре-реплецируется, как только число реплик становится ниже заданного уровня
2. Ре-репликация происходит по приоритету

Балансировка

Мастер отдает команды на перераспределение кусков для лучшей утилизации дисков и сети

Механиз удаления файлов

1. Файл помечается удаленным (память не освобождается)
2. Освобождение ресурсов происходит во время очередного цикла сборки мусора (файл должен числится удаленным больше определенного периода)

Стратегии

- Быстрое восстановления мастера и chunkserver’ов
- Репликация данных
- Репликация состояния мастера

Логгирование

Изменения становятся доступны клиентам только после того как все действия будут внесены в лог

Hadoop Distributed File System (HDFS)[2][3]

[2] Konstantin Shvachko и др. “The hadoop distributed file system.”. В: *MSST*. Т. 10. 2010, с. 1—10.

[3] Tom White. *Hadoop: The definitive guide*. ” O’Reilly Media, Inc.”, 2012.

Отличия от GFS

- Open-source реализация GFS
- Master — NameNode
- Chunkserver — DataNode

 Ghemawat, Sanjay, Howard Gobioff и Shun-Tak Leung. “The Google file system”. B: (2003).

 Shvachko, Konstantin и др. “The hadoop distributed file system.”. B: *MSST*. T. 10. 2010, с. 1—10.

 White, Tom. *Hadoop: The definitive guide*. ” O’Reilly Media, Inc.”, 2012.

First Extra slide