Optimistic Concurrency Control in a Distributed NameNode Architecture for Hadoop Distributed File System

Qi Qi

Instituto Superior Técnico - IST (Portugal)
Royal Institute of Technology - KTH (Sweden)
Swedish Institute of Computer Science - SICS (Sweden)

qiq@kth.se

19 September 2014

Overview

Motivation

Industrial Standard in Big Data Era

Apache Hadoop Ecosystem

Limits to growth in HDFS

Number of Files	Memory Requirement	Physical Storage
1 million	0.6 GB	0.6 PB
100 million	**60 GB**	**60 PB**
1 billion	600 GB	600 PB

Hop-HDFS and Its Limitation

Distributed NameNode Architecture

Restricted Concurrency

Problem Statement

HDFS

System-level Lock

Hop-HDFS v1

System-level Lock

Hop-HDFS v2

Row-level Lock

MySQL Cluster

Read Committed / Anomalies

Contribution

Architectures and Namespace Concurrency Control

GFS, HDFS, Hop-HDFS and MySQL Cluster

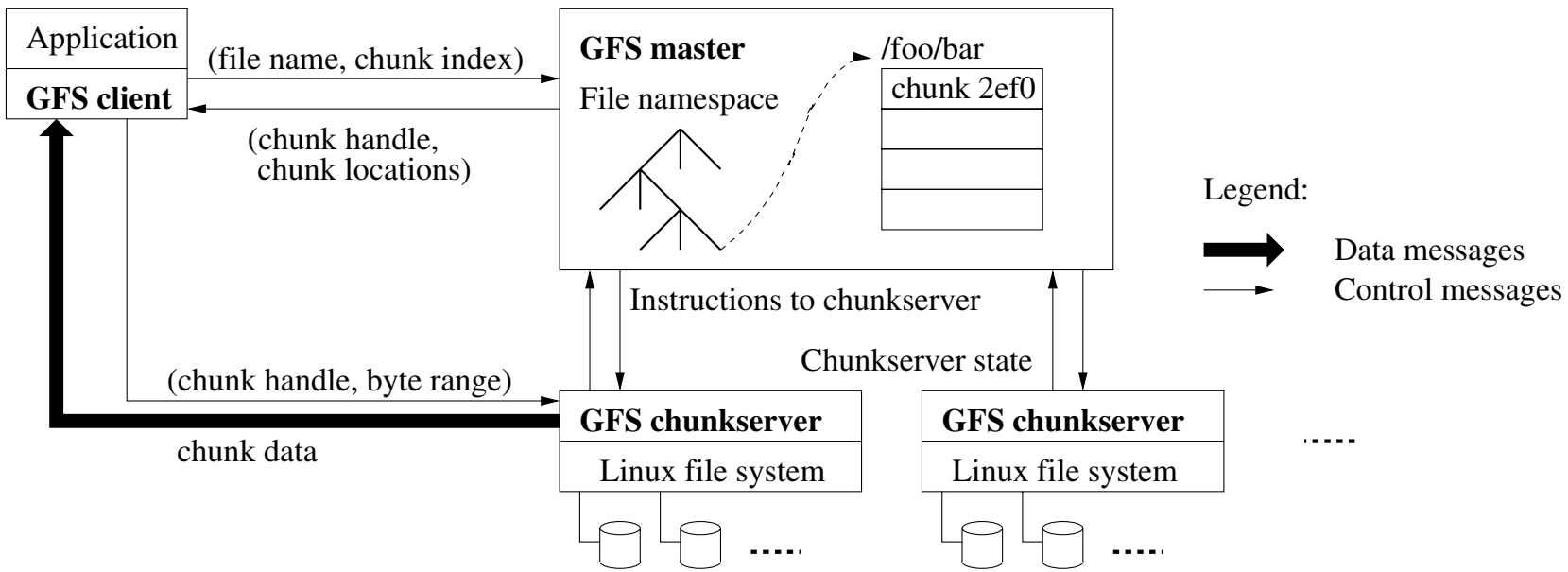
Performance Accessment and Limitation Analysis

HDFS v.s. Hop-HDFS v2 (PCC version)

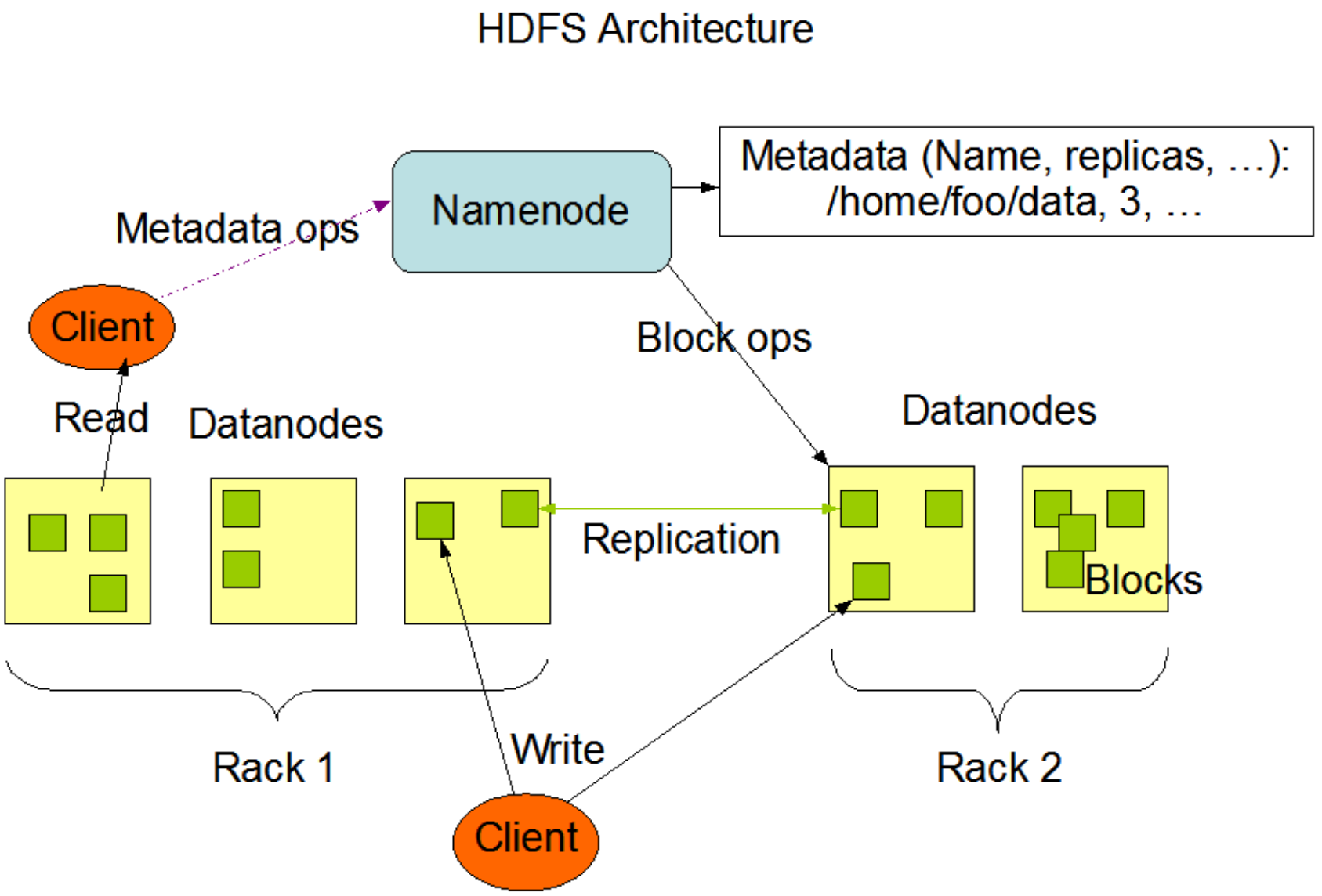
Solution for Hop-HDFS

Optimistic Concurrency Control with Snapshot Isolation on Semantic Related Group

GFS Architecture



HDFS Architecture



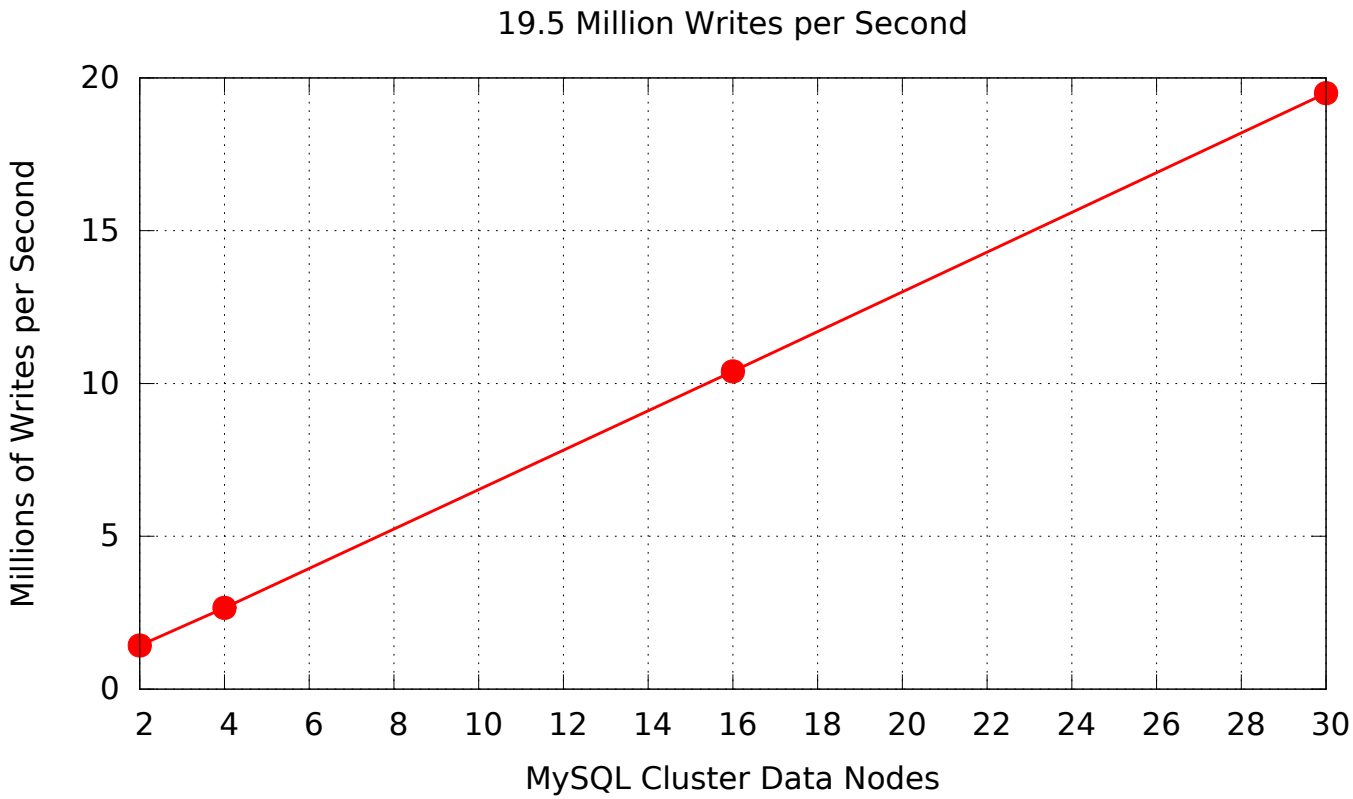
Isolation Level

Berenson, Hal, et al. "A Critique of ANSI SQL Isolation Levels." ACM SIGMOD Record 24.2 (1995): 1-10.

Isolation Level	**Lost Update**	**Fuzzy Read**	**Phantom**	**Read Skew**	**Write Skew**
Read Uncommitted	✓	✓	✓	✓	✓
Read Committed	✓	✓	✓	✓	✓
Cursor Stability	sometimes	sometimes	✓	✓	sometimes
Repeatable Read	X	X	✓	X	X
Snapshot	X	X	sometimes	X	✓
Serializable	X	X	X	X	X

MySQL Cluster

- Distributed, in-memory, replicated database
- Supports only **Read Committed**
- High throughput:

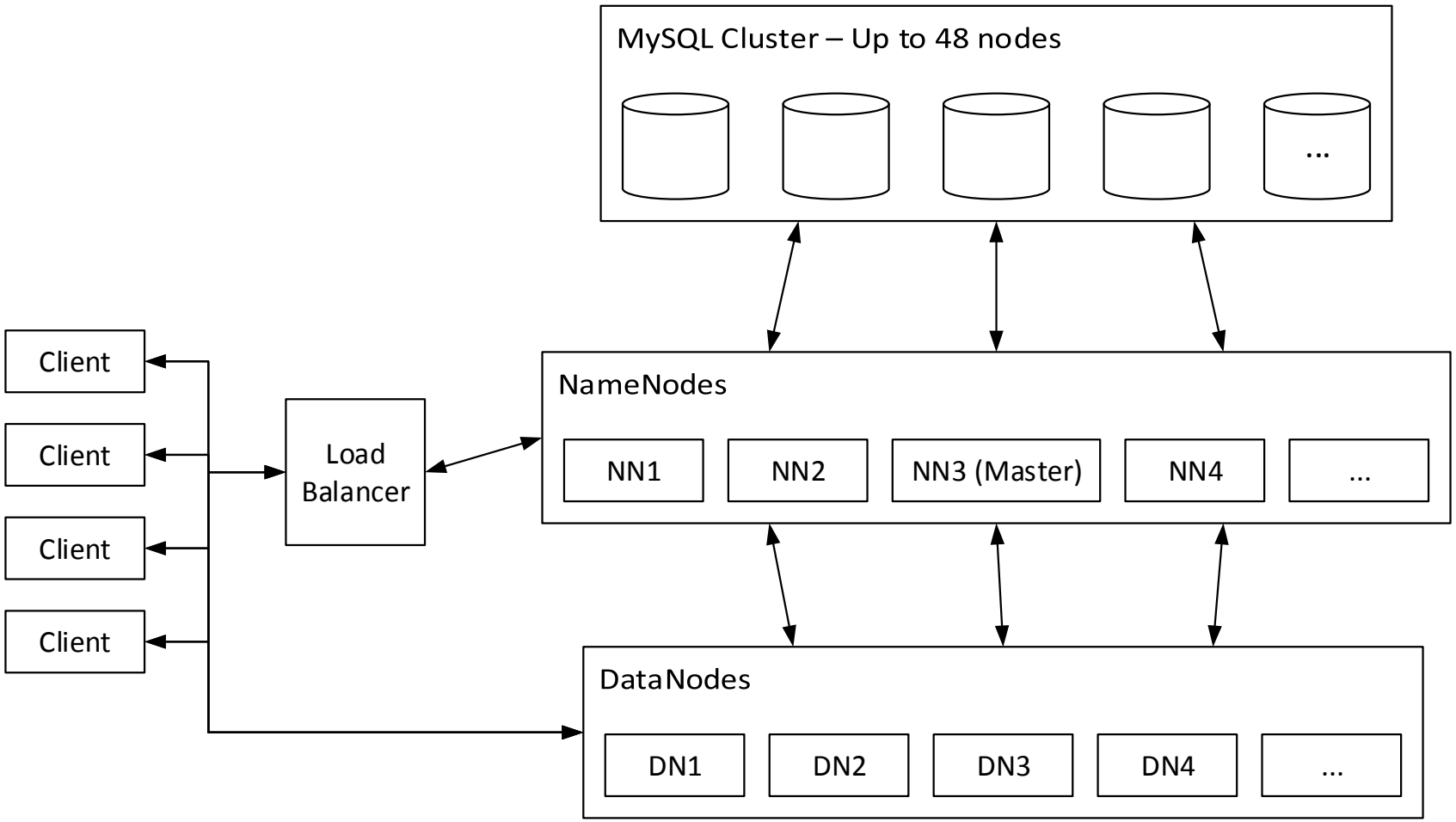


Hop-HDFS

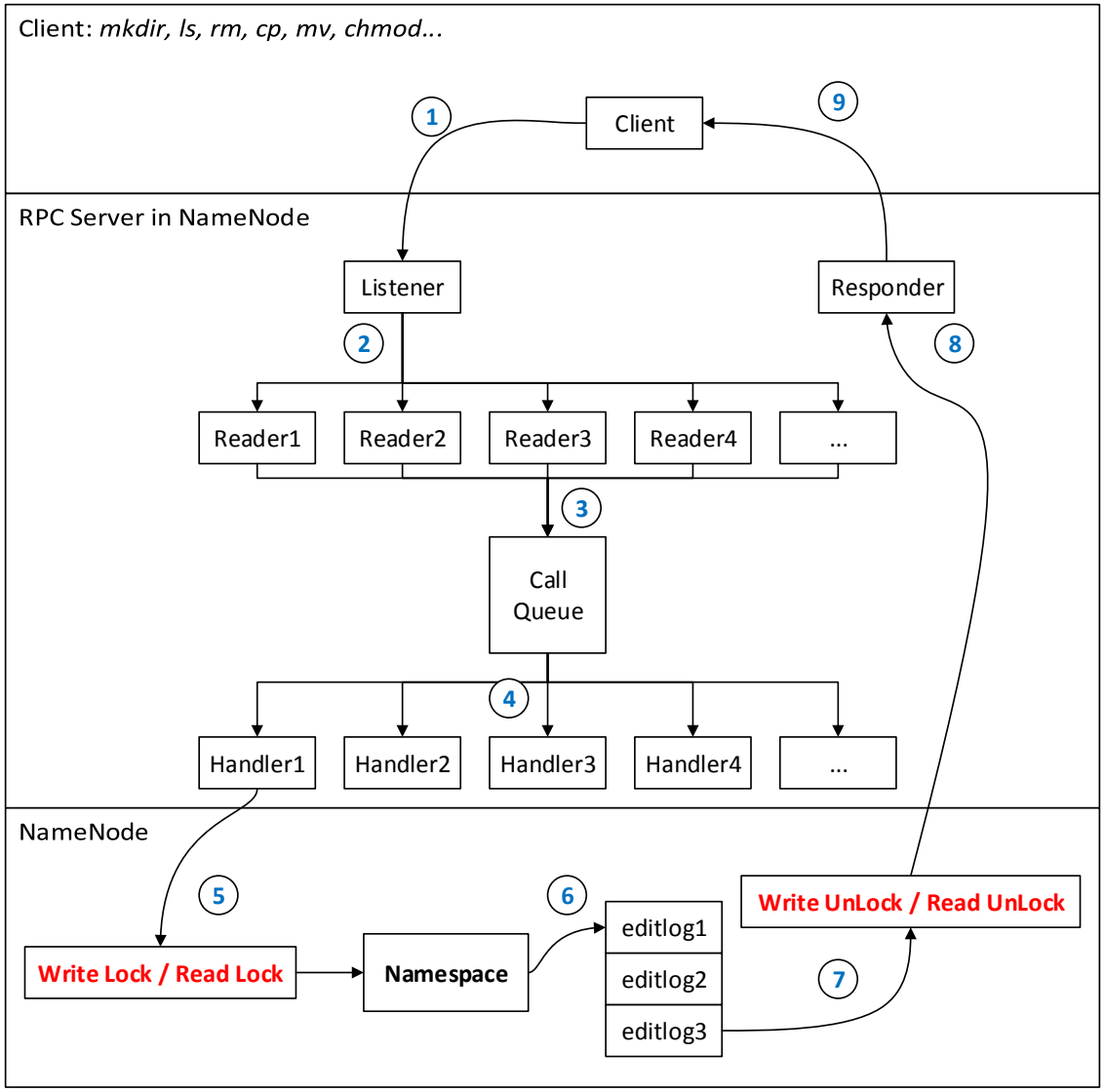
Overcome Limitations in HDFS NameNode

- Scalability of the Namespace
- Throughput Problem
- Failure Recovery

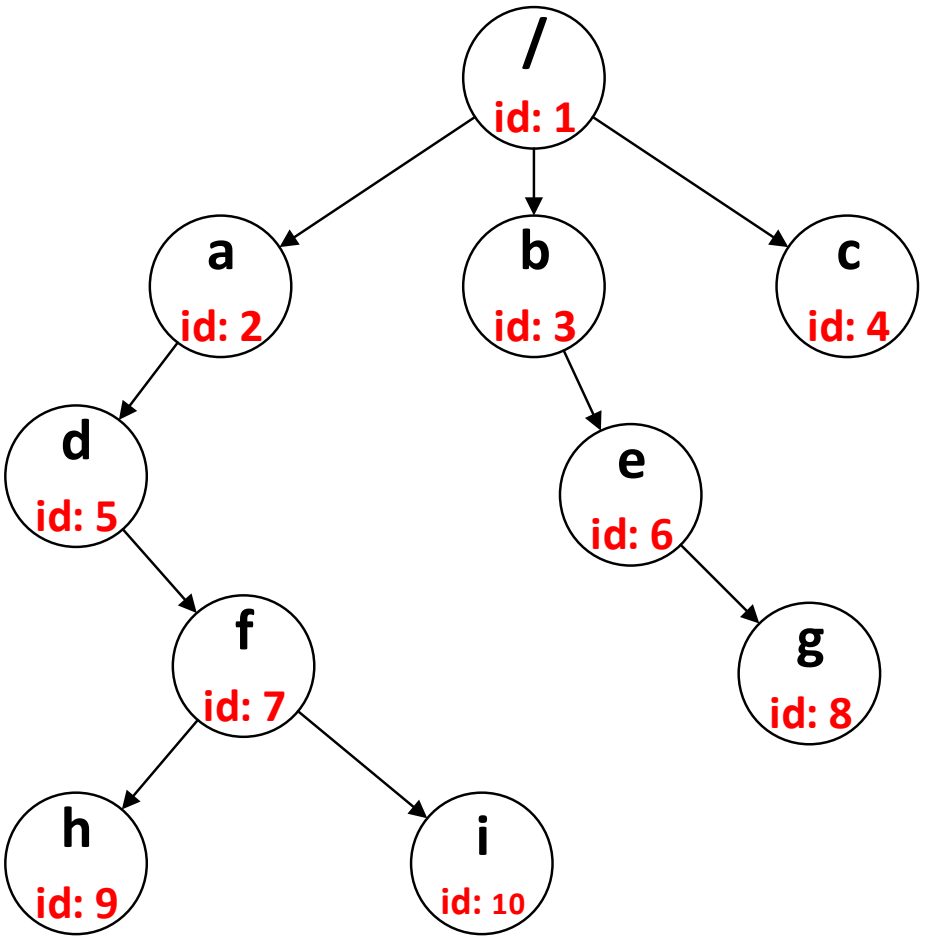
Hop-HDFS Architecture



HDFS Namespace Concurrency Control



HDFS Namespace Structure



id	parent_id	name
1	0	/
2	1	a
3	1	b
4	1	c
5	2	d
6	3	e
7	5	f
8	6	g
9	7	h
10	7	i

Table

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table : Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

References

Shvachko, K. V. (2010).
Hdfs scalability: The limits to growth.
login 35(2), 6–16.

Thank you.