distributed file system

Arun C. Murthy
CCDI, Yahoo!
acm@yahoo-inc.com

Challenge

- How do you scale up storage and applications?
  - Store tens of petabytes of data
  - 100s of terabytes per applications
- Need lots of cheap computers
  - Fixes speed problem (15 minutes on 1000 computers), but…
  - Reliability problems
    - In large clusters, computers fail every day
    - Cluster size is not fixed
- Need common infrastructure
  - Must be efficient and reliable

Hadoop Overview

- Open Source Apache Project
  - http://hadoop.apache.org/core
- Hadoop Core includes:
  - Hadoop Distributed File System - distributes data
  - Map/Reduce – parallel processing framework
- Written in Java
- Runs on
  - Linux, Mac OS/X, Windows, and Solaris

Hadoop Distributed File System

- Single multi-petabyte file system for entire cluster
  - Managed by a single *namenode*.
  - Files are written, read, renamed, deleted, but append-only.
  - Optimized for streaming reads of large files.
- Files are broken in to large blocks.
  - Transparent to the client
  - Blocks are typically 128 MB
  - Replicated to several *datanodes*, for reliability
- Intelligent client library talks to both namenode and datanodes
  - Data is not sent through the namenode.
  - Throughput of file system scales nearly linearly.
- Access from Java, C, or command line.

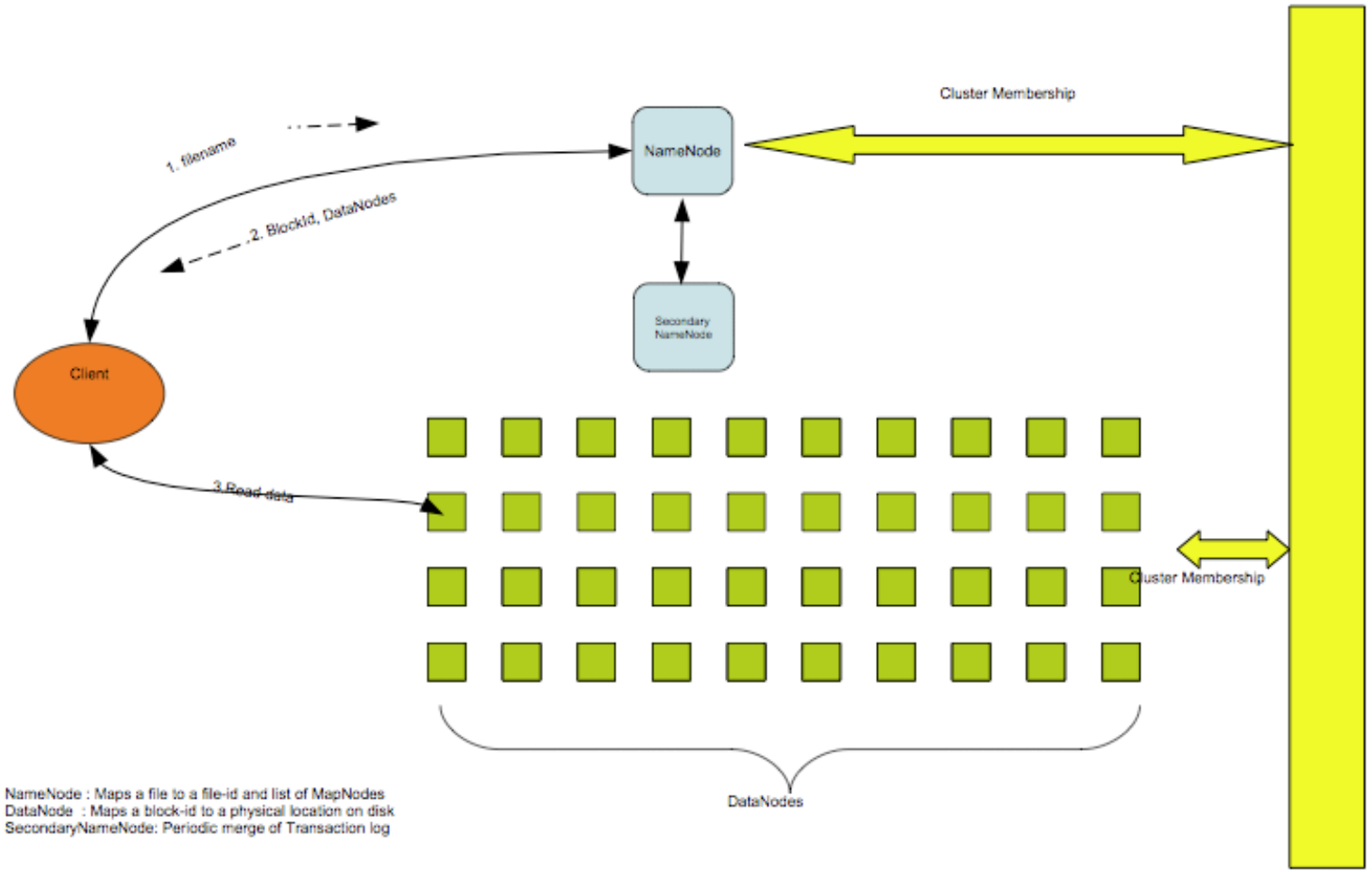
Block Placement

- Default is 3 replicas, but settable per file
- Blocks are placed (writes are pipelined):
  - On same node
  - On different rack
  - On the other rack
- Clients read from closest replica
- If the replication for a block drops below target, it is automatically re-replicated.
- Balancer

Data Correctness

- Data is checked with CRC32
- File Creation
  - Client computes checksum per 512 byte
  - DataNode stores the checksum
- File access
  - Client retrieves the data and checksum from DataNode
  - If Validation fails, Client tries other replicas
- Periodic validation by DataNode

Client Operations



Q&A

- For more information:
  - Website: http://hadoop.apache.org/core
  - Mailing lists:
    - core-dev@hadoop.apache
    - core-user@hadoop.apache
  - IRC: #hadoop on irc.freenode.org