Running the largest HDFS cluster

Hairong Kuang, Tom Nykiel
hairong@fb.com tomasz@fb.com

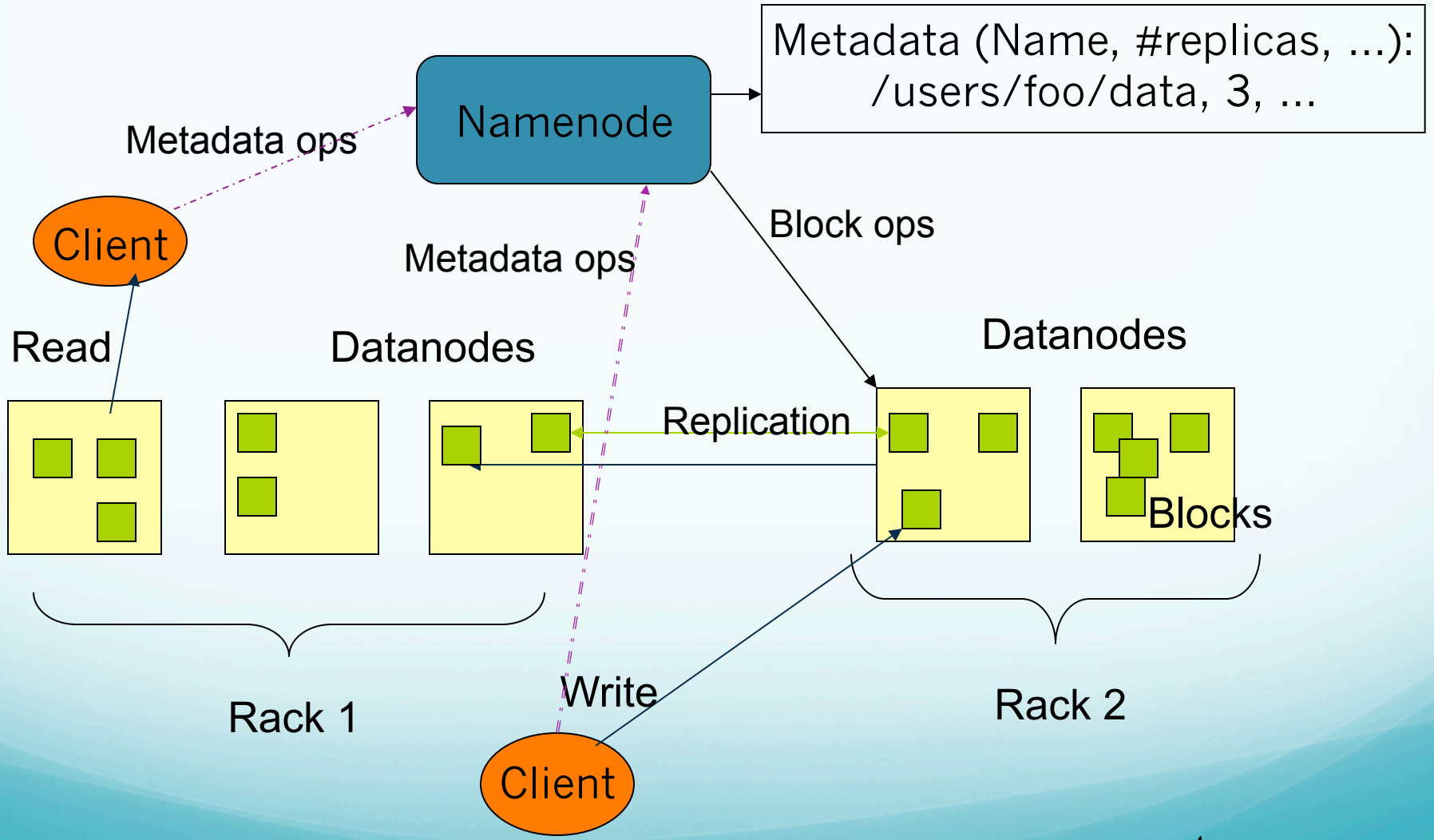
Agenda

- HDFS at Facebook
- Improving HDFS scalability
- HDFS federation

What is HDFS

- HDFS:
  - Storage layer for Hadoop Open Source Apache project
  - Scale: petabytes of data on thousands of nodes
- Characteristics:
  - Uses clusters of commodity computers
  - Use replication across servers to deal with unreliable storage/servers
  - Metadata-data separation - simple design
  - Slightly Restricted file semantics
    - Focus is mostly sequential access
    - Single writers
    - No file locking features
  - Supports moving computation close to data
    - Single ‘storage + compute’ cluster vs. Separate clusters

HDFS Architecture



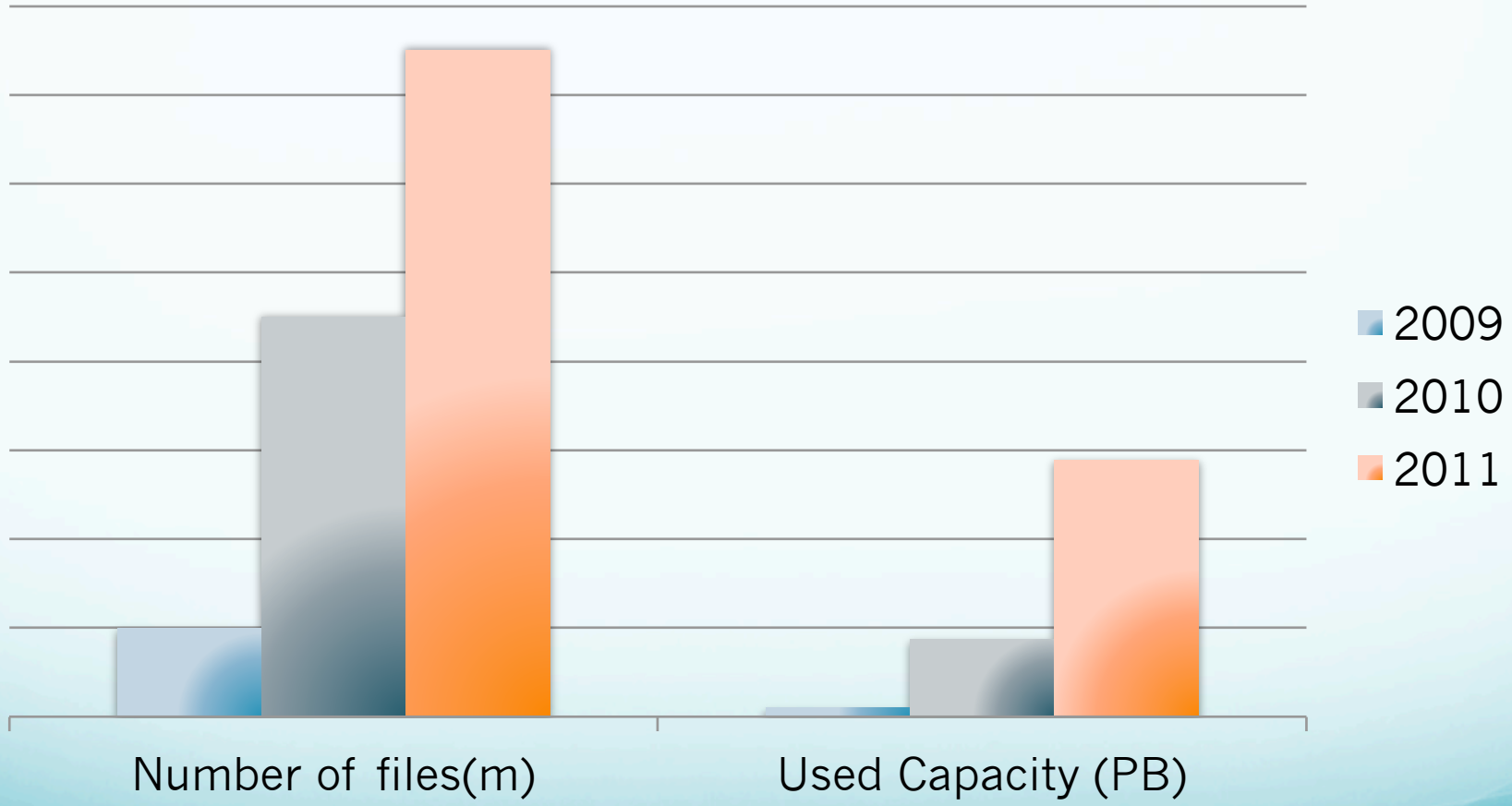
Facebook Use of HDFS

- Quiz: What is the total number of HDFS clusters used in Facebook ???
- The biggest one: warehouse cluster storing Hive tables

The largest HDFS cluster

- Thousands of nodes
- Close to 100 PB of configured capacity
- 100+ million files
- Thousands of concurrent clients access the cluster
- At peak hour, thousands of audit requests per second
- It is growing each day

Growth of the cluster



Agenda

- HDFS at Facebook
- Improving HDFS scalability
  - Scale of the system
  - System monitoring
  - Communication
  - Synchronization
  - Data structures and algorithms
  - Network awareness
  - Handling persistency
  - Memory management
  - Tiny bugs – huge losses
- HDFS federation

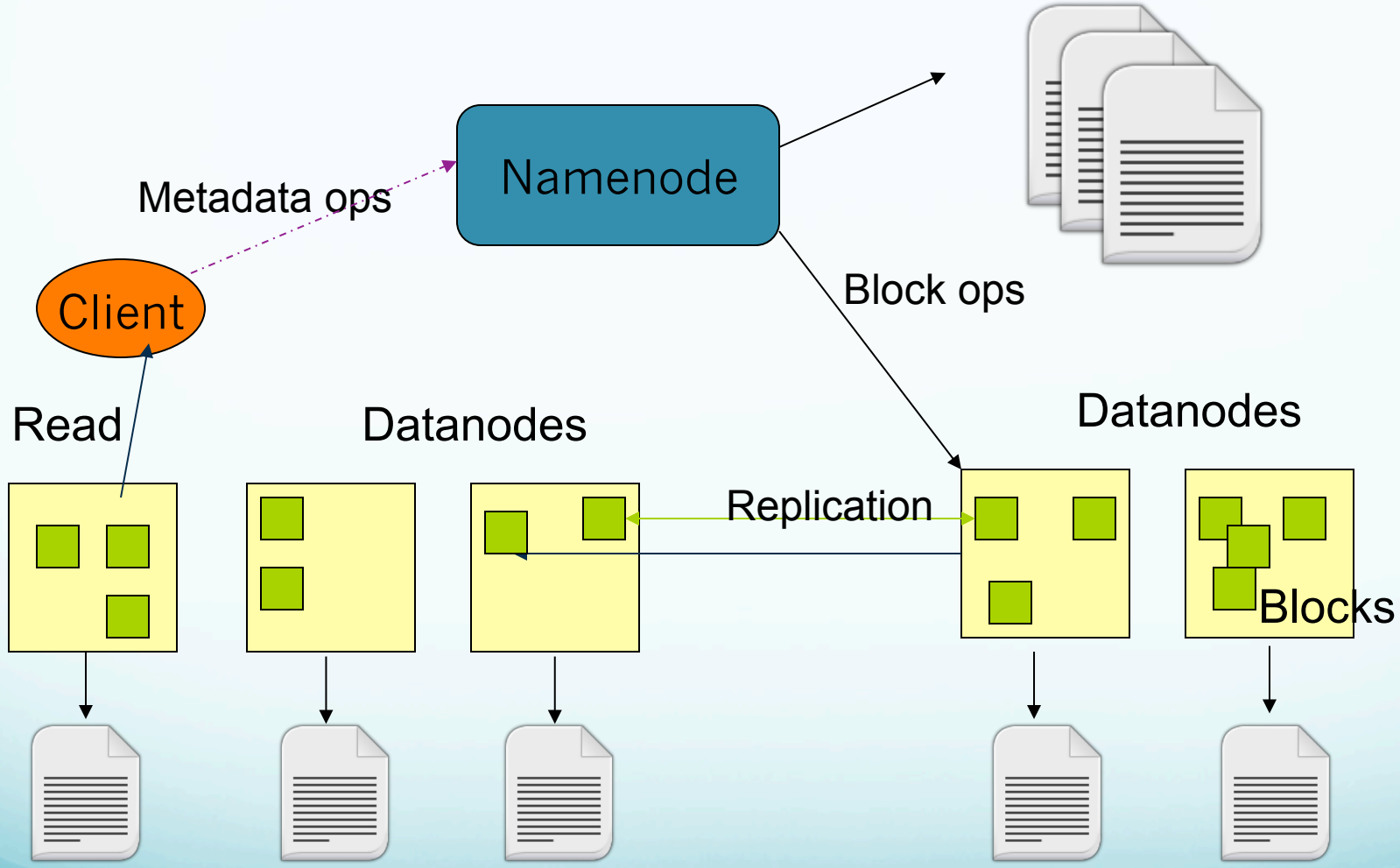
Scale of the system

- FSDirectory
  - Information about all files/directories in the namespace
- BlocksMap
  - Information about all the blocks in the filesystem
- Other associated structures - examples:
  - Queues for storing replication status
  - Table of datanodes

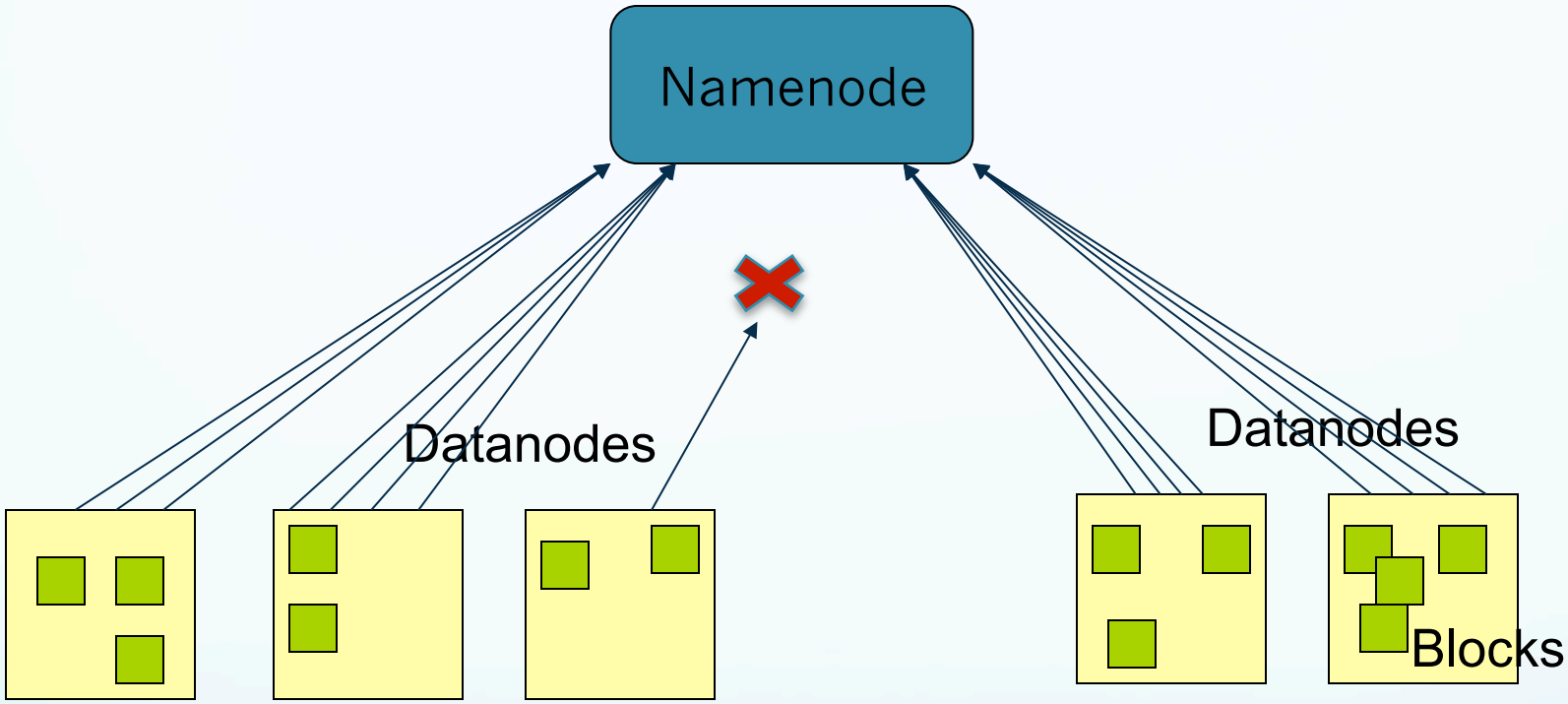
Memory utilization:

- In memory state
- FSImage + FSEditsLog
  - ensuring persistency
- System logs
  - debugging and monitoring

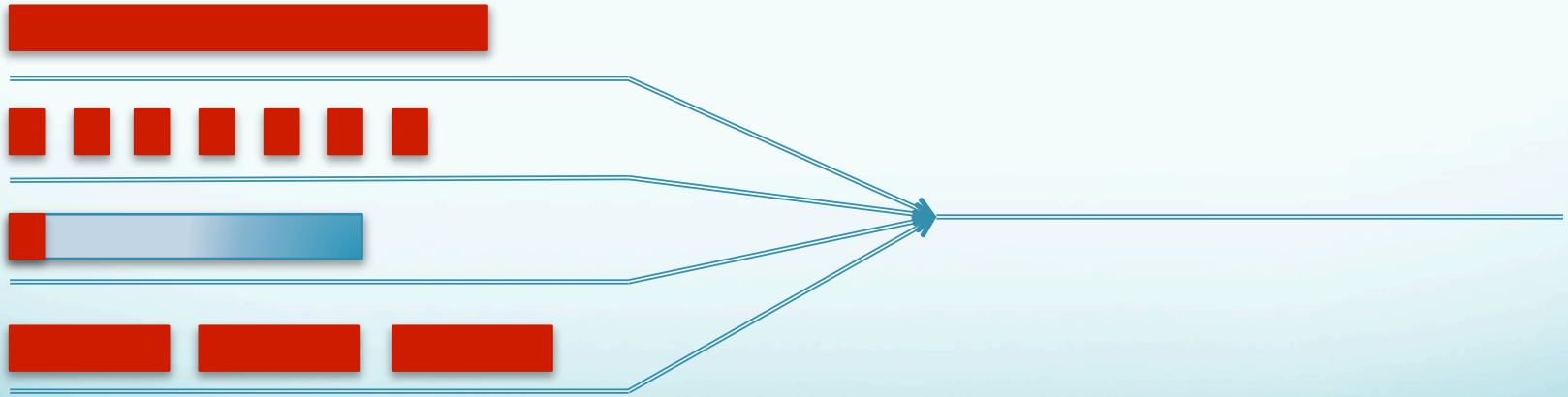
System monitoring - logging



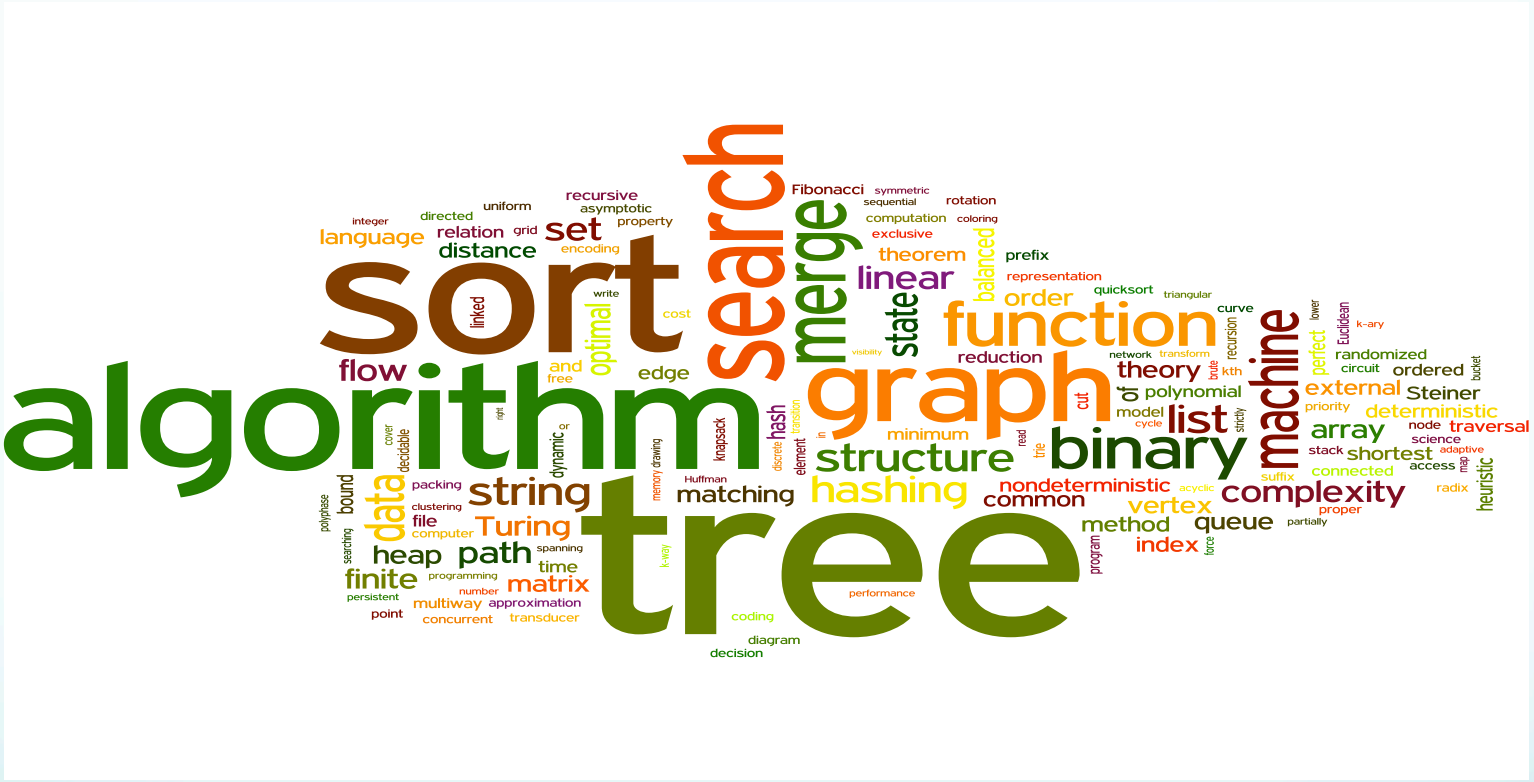
Communication



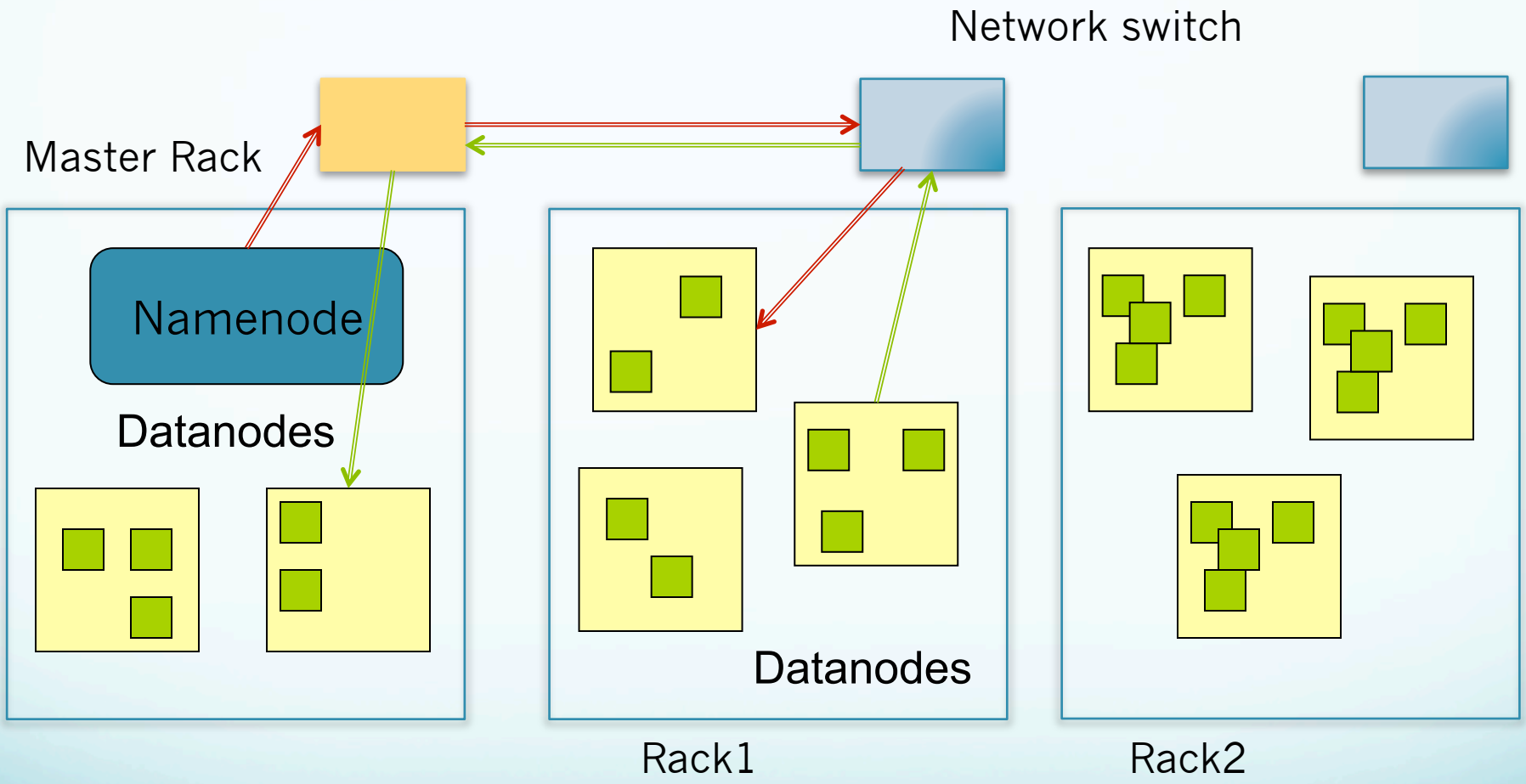
Synchronization



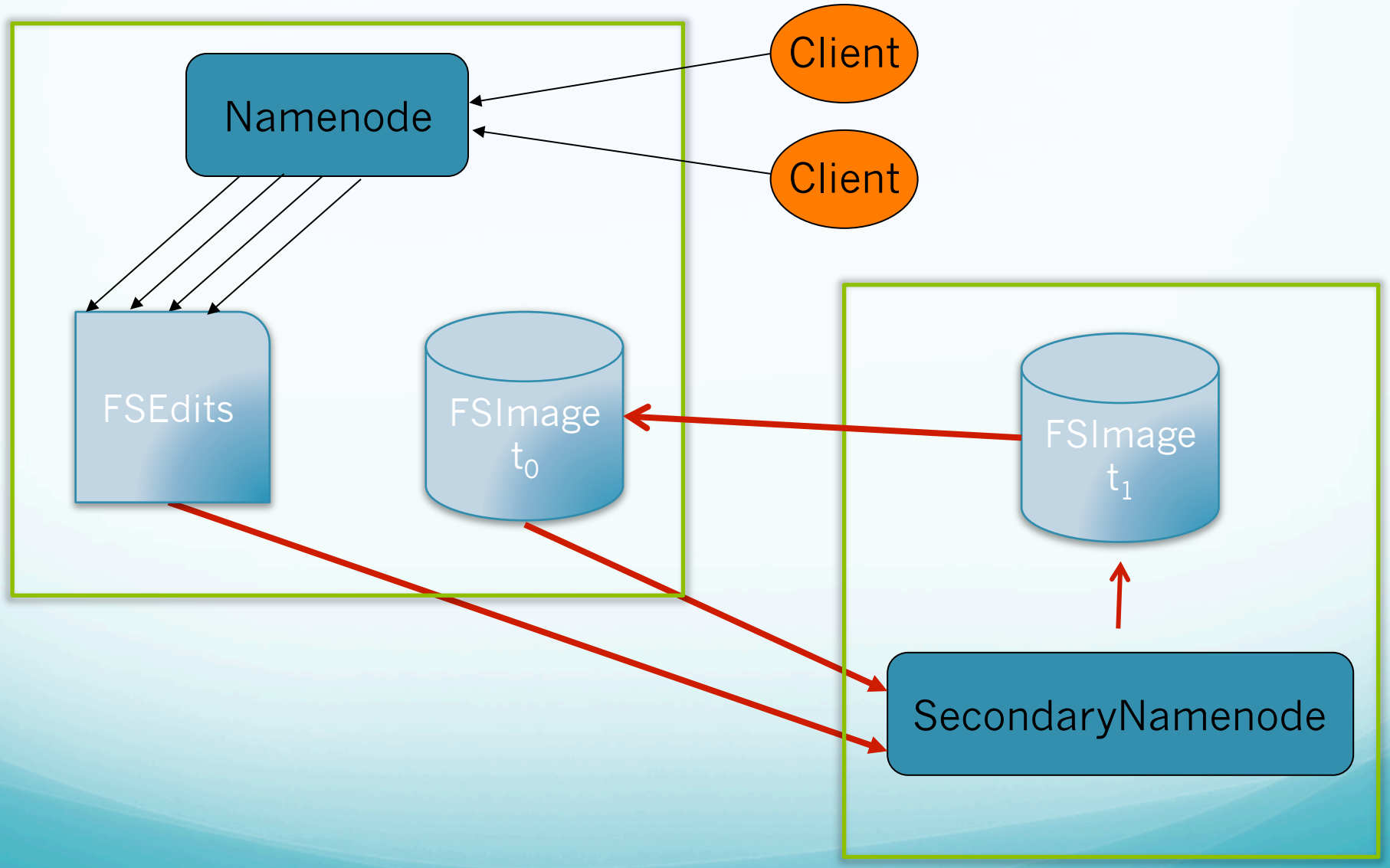
Data structures and algorithms



Network awareness



Handling persistency



Memory management

- Namenode running with enormous heap space
- Problem: A full GC takes at least 10 minutes
  - The NameNode is non-responsive !
- Improvements:
  - Configuration changes
  - Avoid unnecessary creations of temporary data

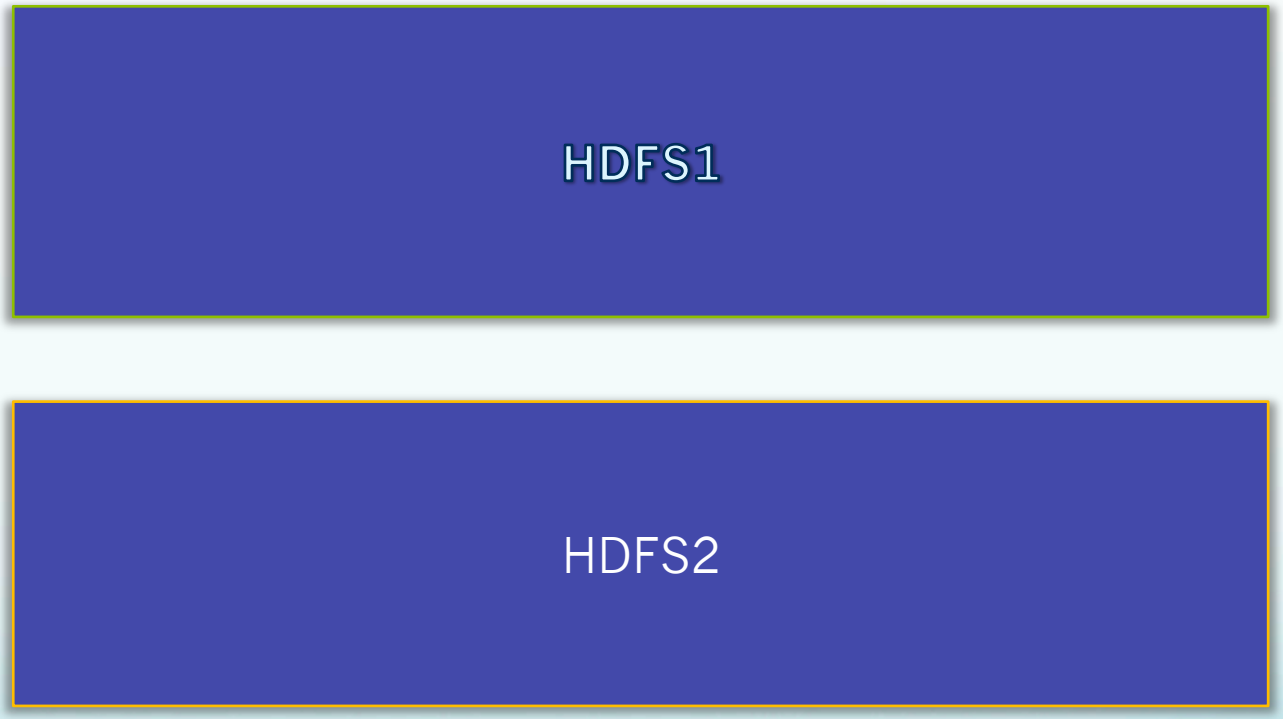
Tiny bugs - huge losses

- A bug in the MR application layer caused the scan of the whole /tmp subtree for each job submitted:
  - Huge number of VALID requests to the NameNode
- Another bug in the application layer exploded the number of metadata read requests by 12 times.

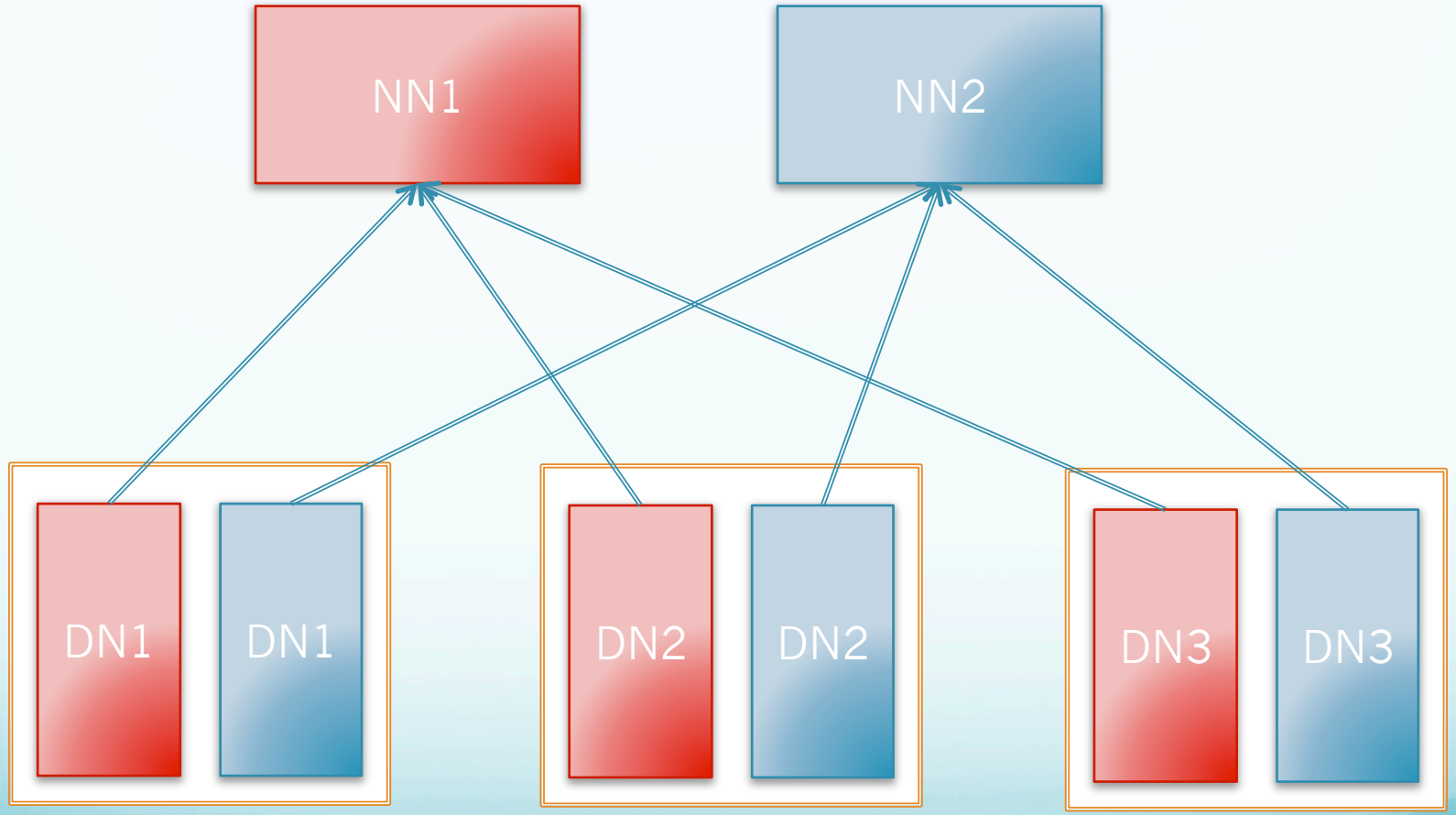
Agenda

- HDFS at Facebook
- Improving HDFS scalability
- HDFS federation

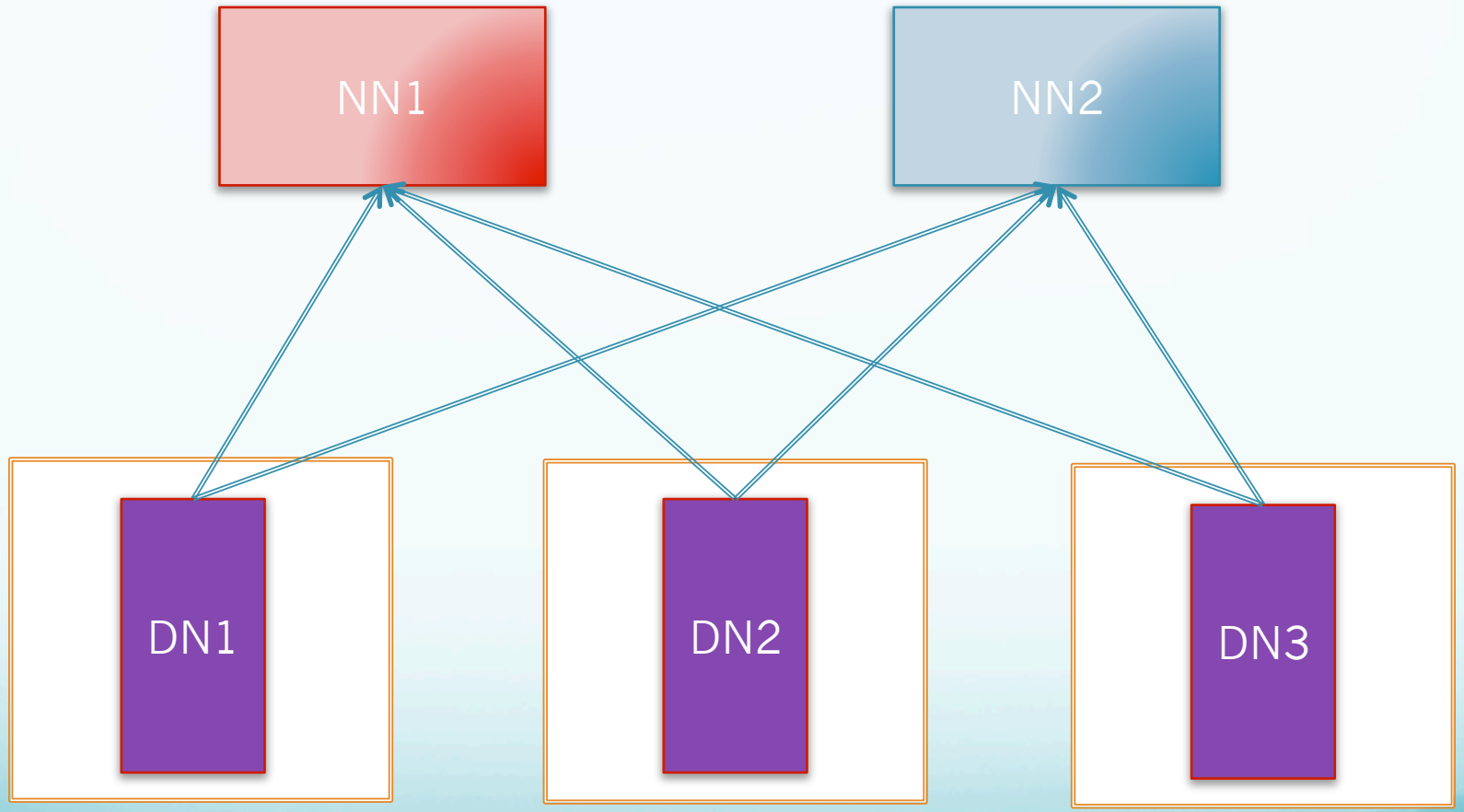
Static Partitioned Clusters



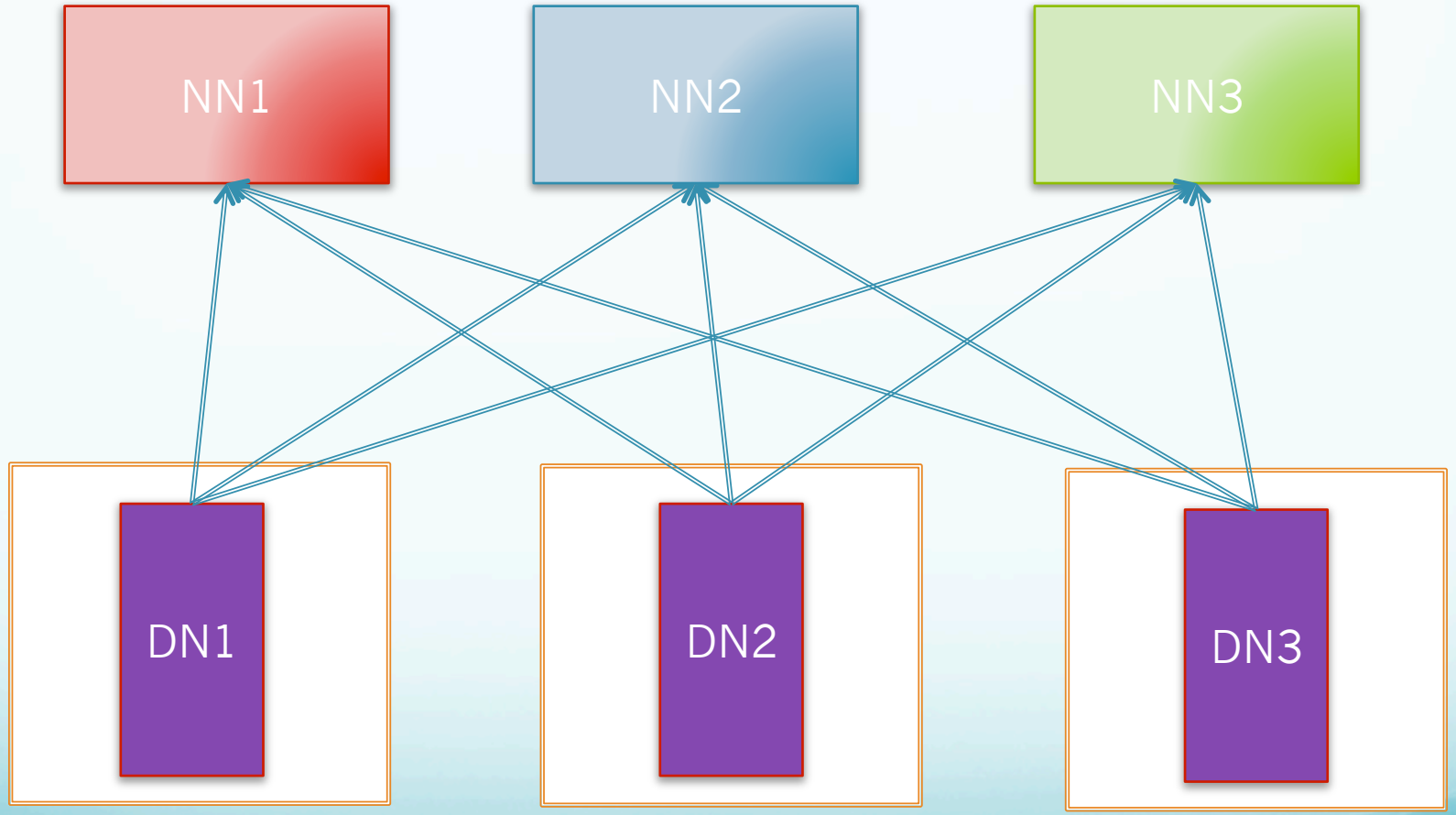
Cluster overlay



Federation



Federation



Conclusion

- We have tons of data stored in HDFS in many clusters, including one of the largest clusters in the world.
- We need to deal with problems never faced before
- Our job is to keep it running efficiently, not lose data, and make it highly available !

Future

- Improve NameNode availability
  - Manual / Automatic failover
- Improve I/O efficiency
- Cross data-center support