HDFS

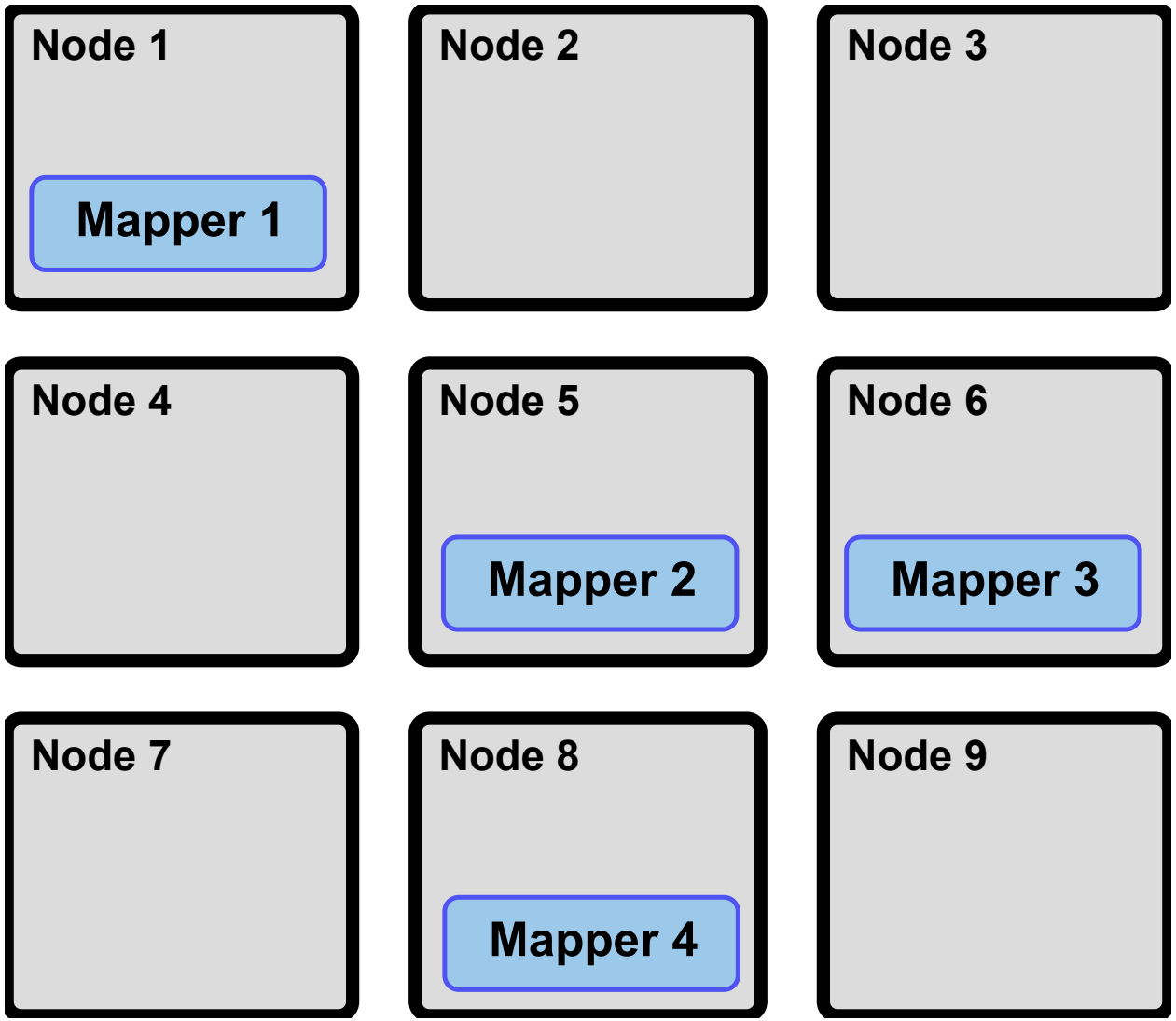
Hadoop Distributed File System

Labs

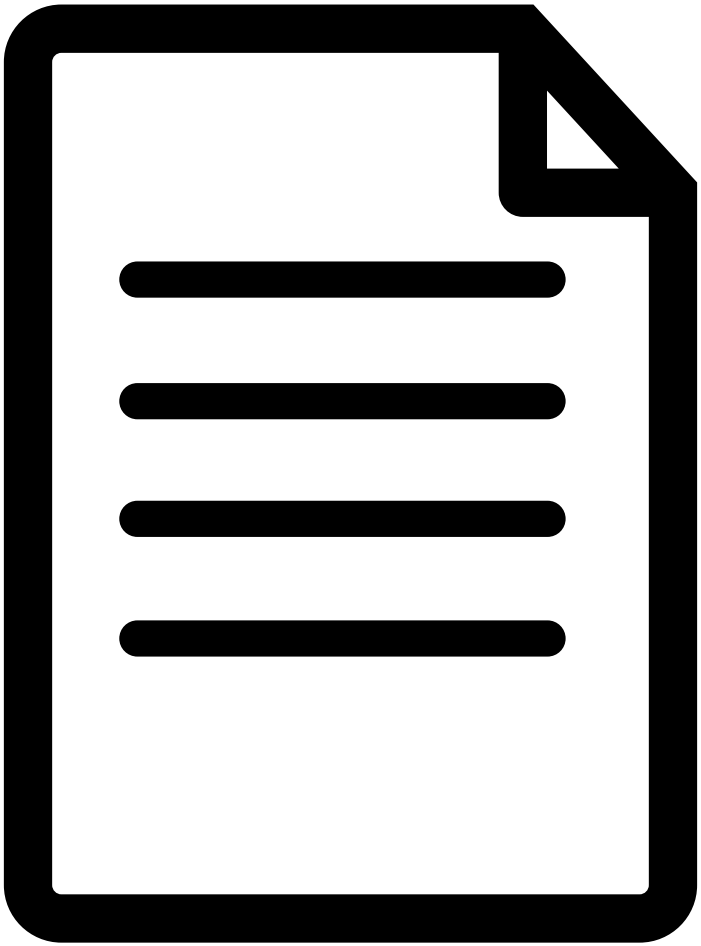
Run 2–27 October (four weeks) at these times:
Monday 9am
Monday 10am
Tuesday 2pm
Wednesday 10am
Wednesday 2pm
Thursday 9am
Thursday 11am
Friday 11am
Friday 2pm

Lab groups will be chosen online: student.inf.ed.ac.uk.

Distributed Map-Reduce

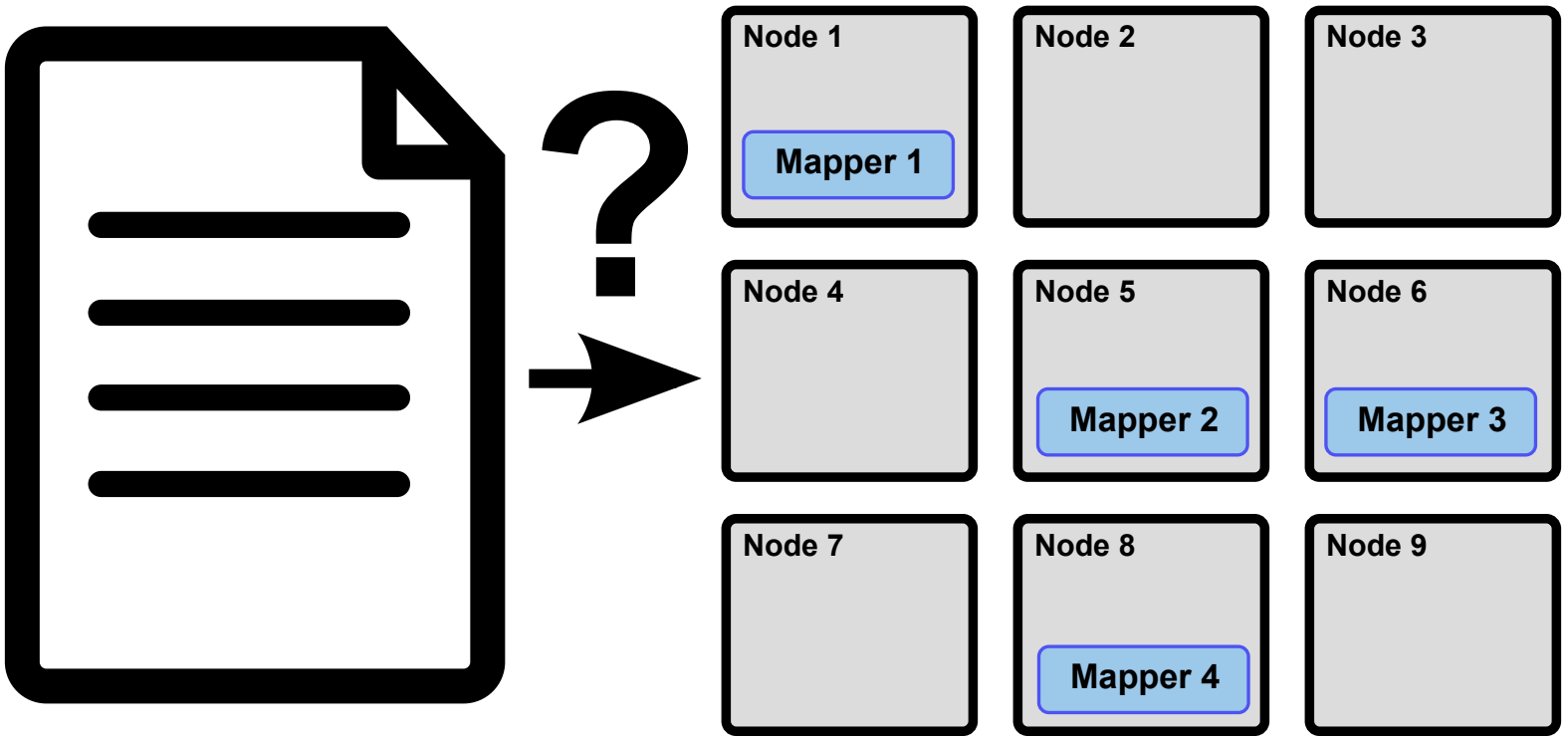


Large Data Sets

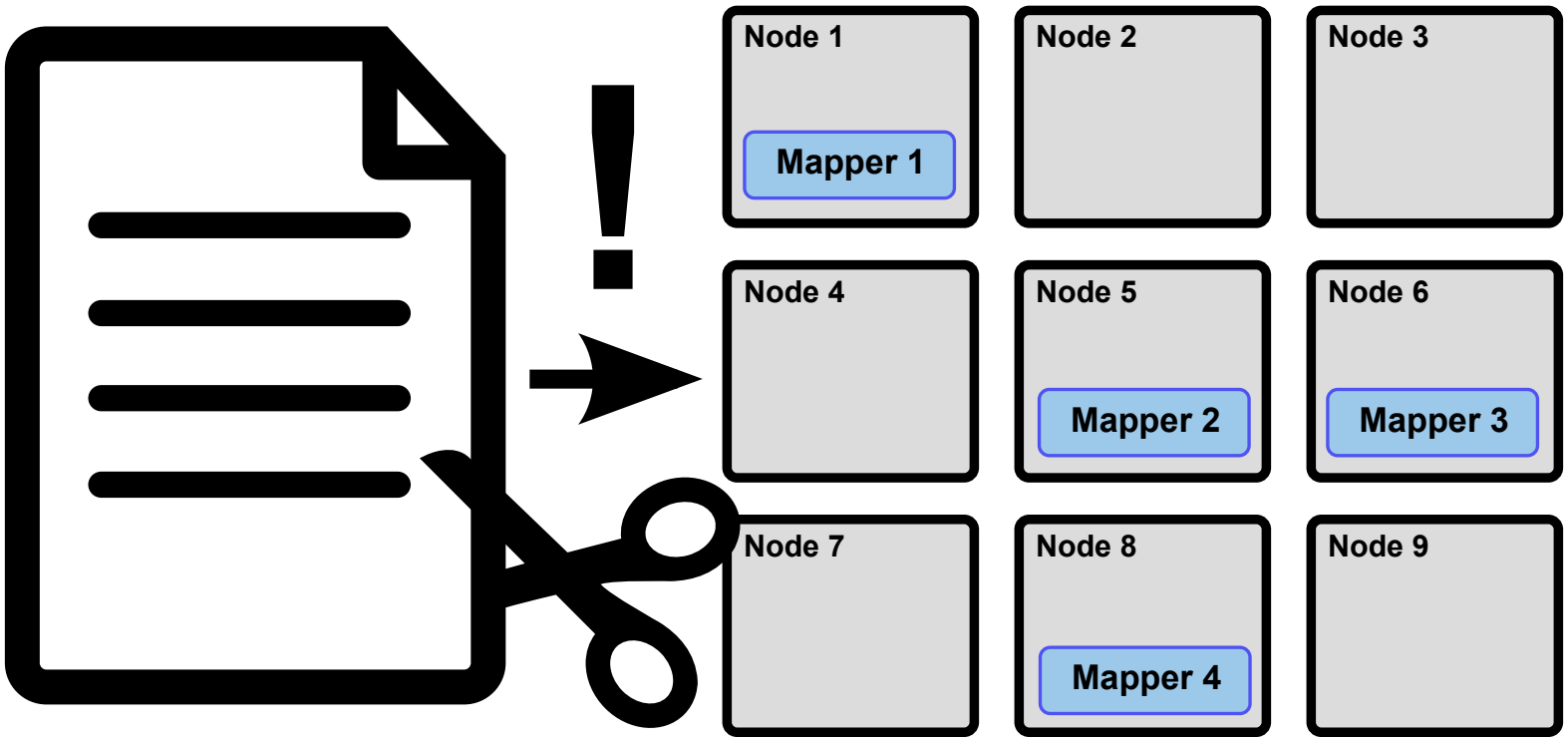


file sizes going up to petabytes

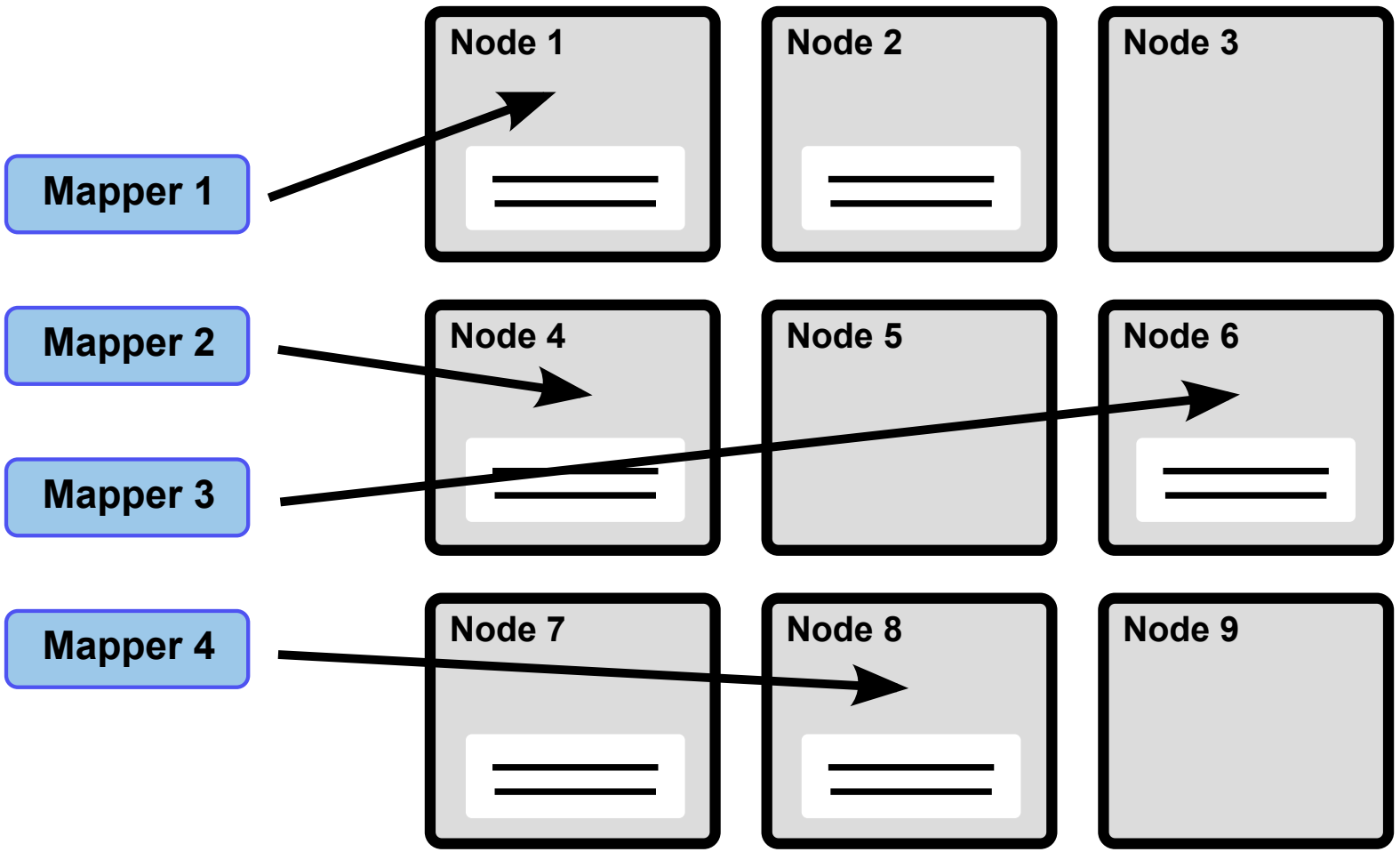
How to get Data to Mappers?



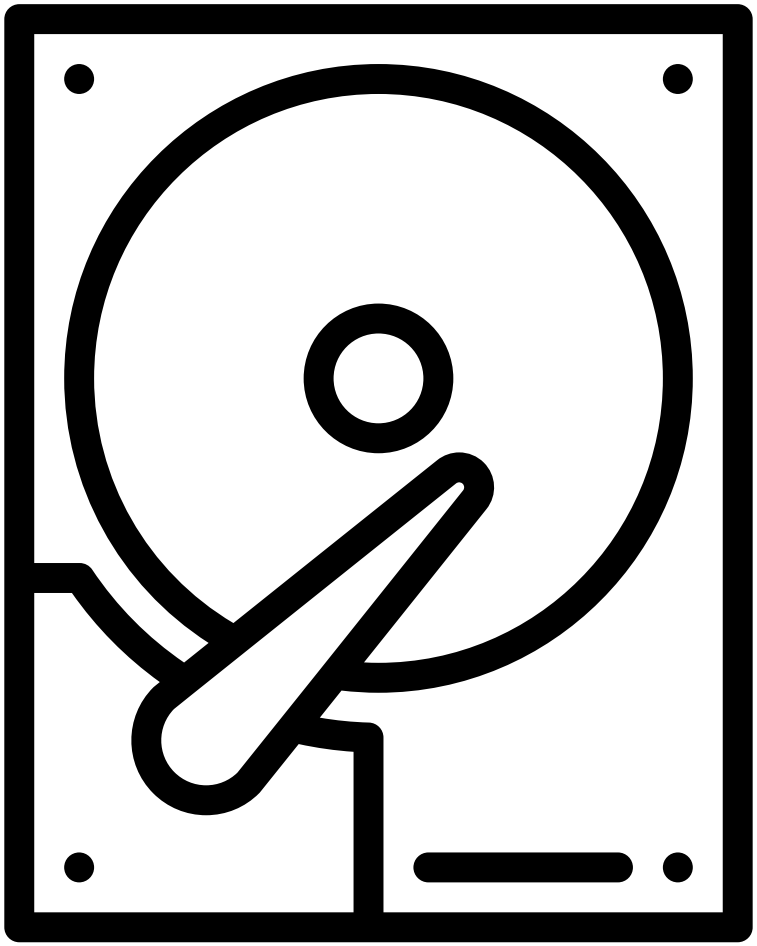
How to get Data to Mappers?



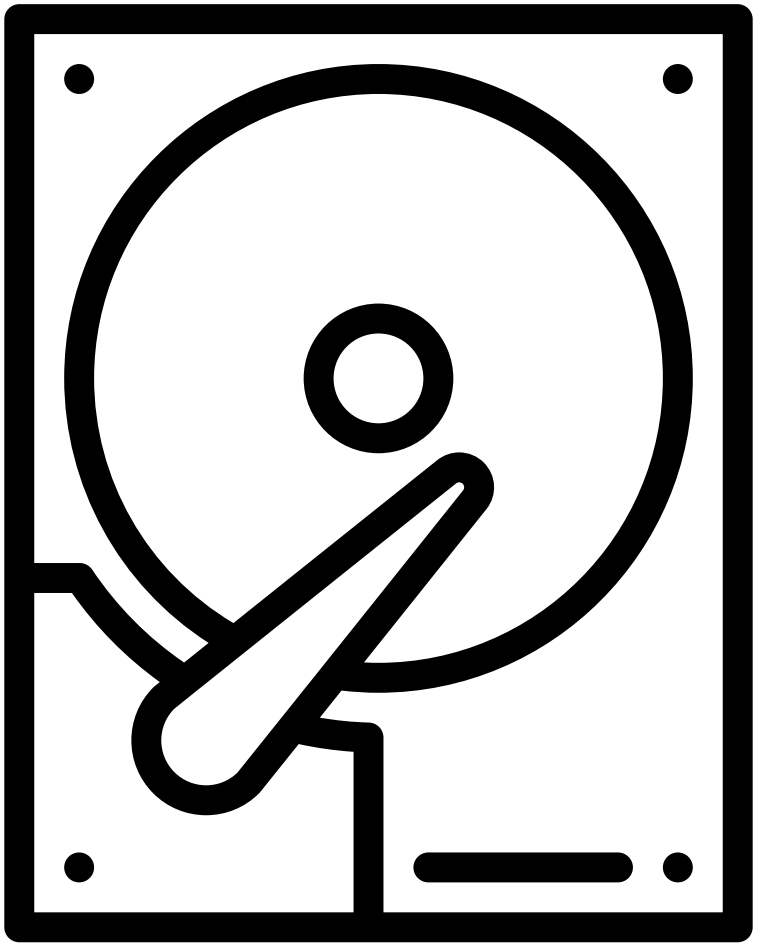
Bring Mappers to Data!



But disk access latency is so high!



But disk access latency is so high!



Yes, but throughput is acceptable.

Distributed File System

Distributed File System

HDFS is a GFS (Google File System) clone

HDFS Design Choices

1. Support handling of large files across multiple nodes

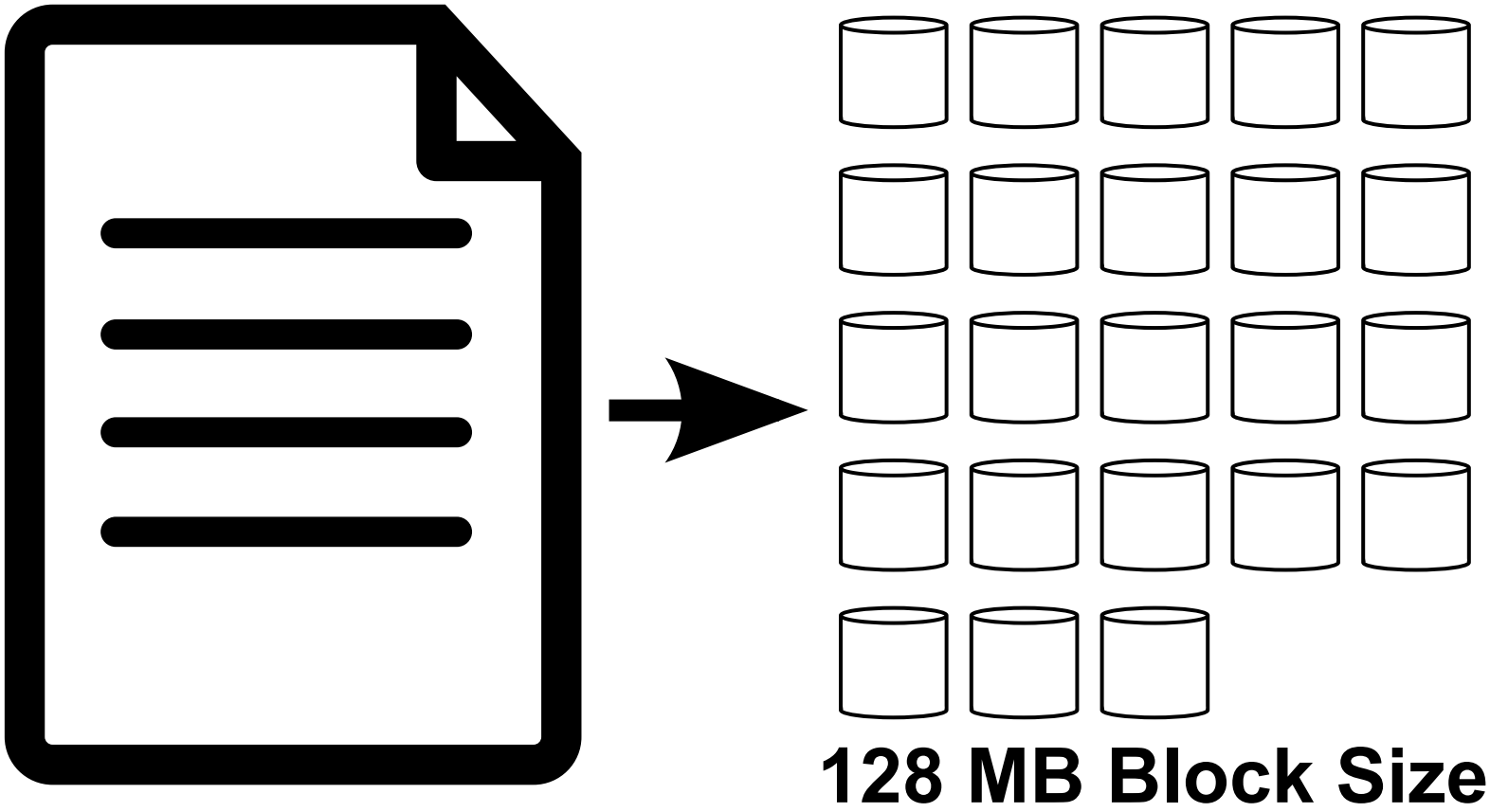
HDFS Design Choices

1. Support handling of large files across multiple nodes
2. Optimise for streaming access

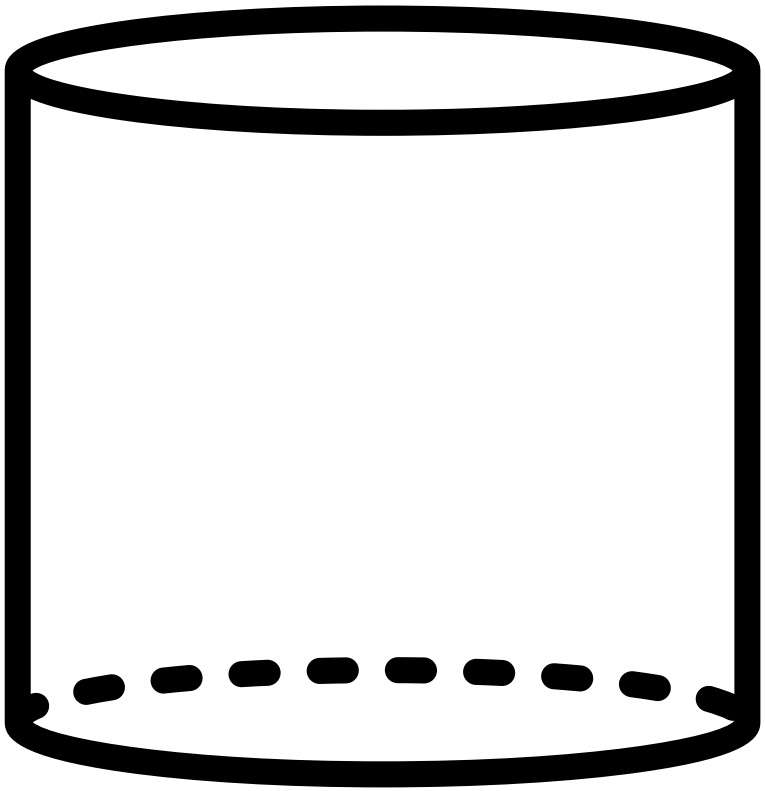
HDFS Design Choices

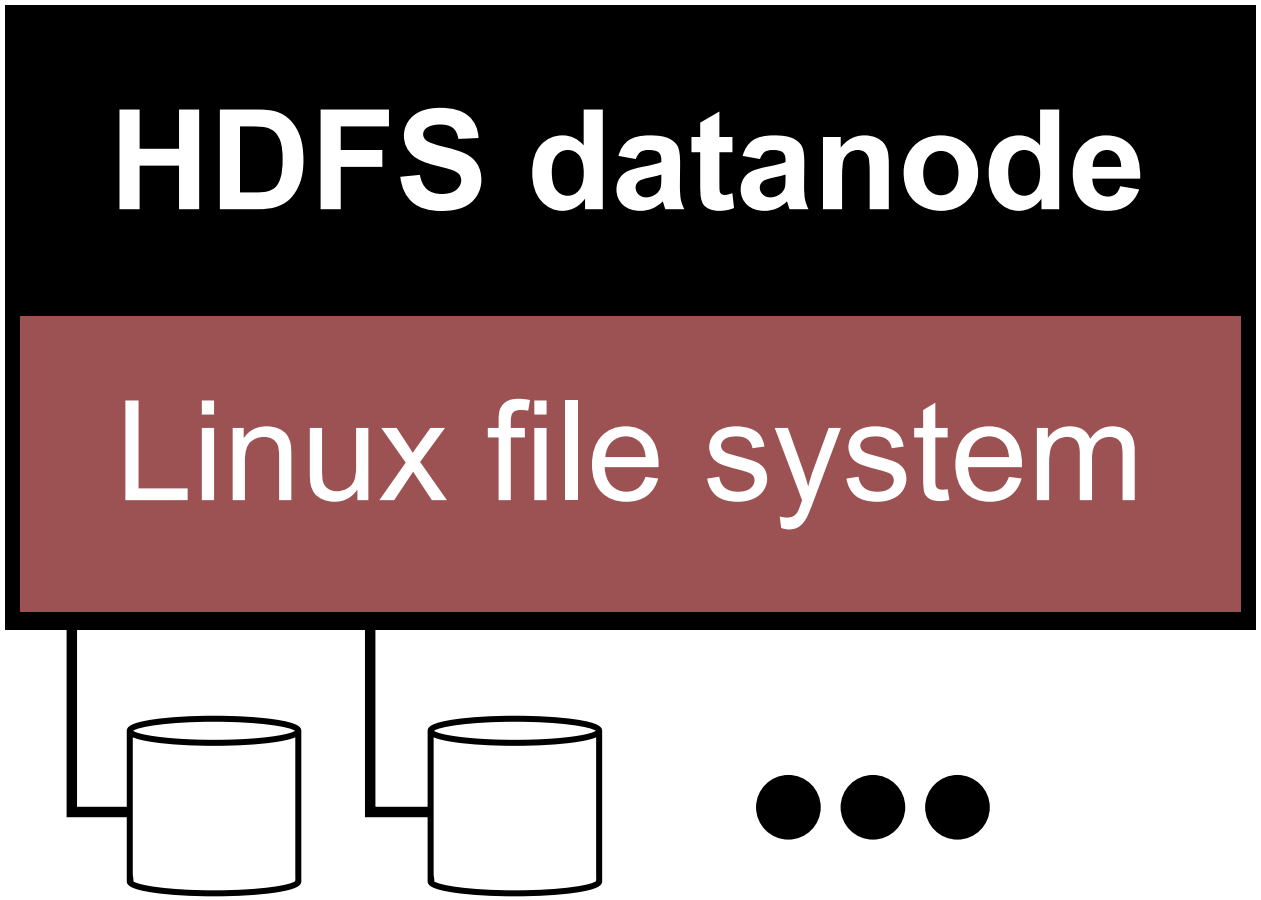
1. Support handling of large files across multiple nodes
2. Optimise for streaming access
3. Run on commodity hardware (e.g. high fault tolerance)

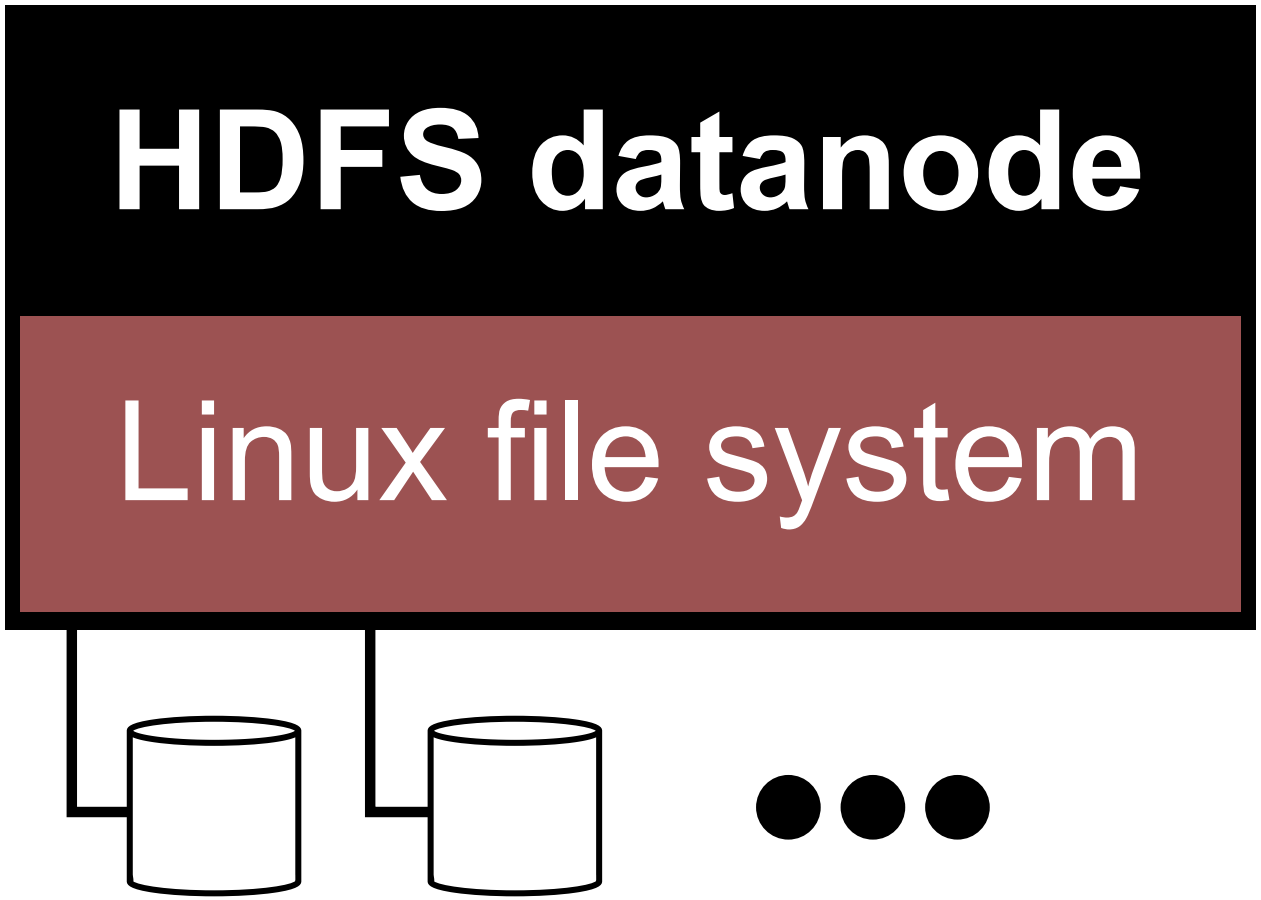
Large Files



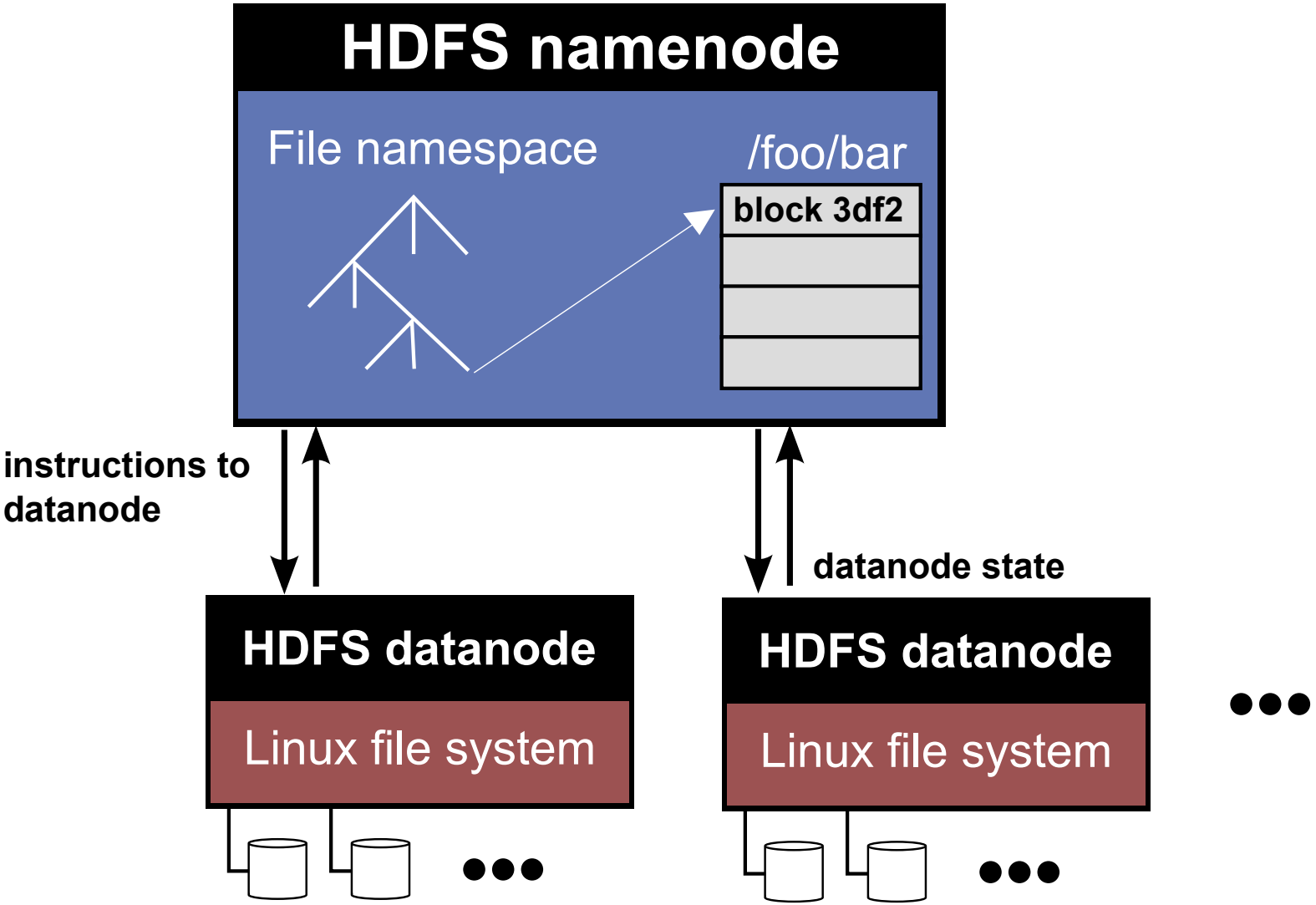
Why so large Blocks?



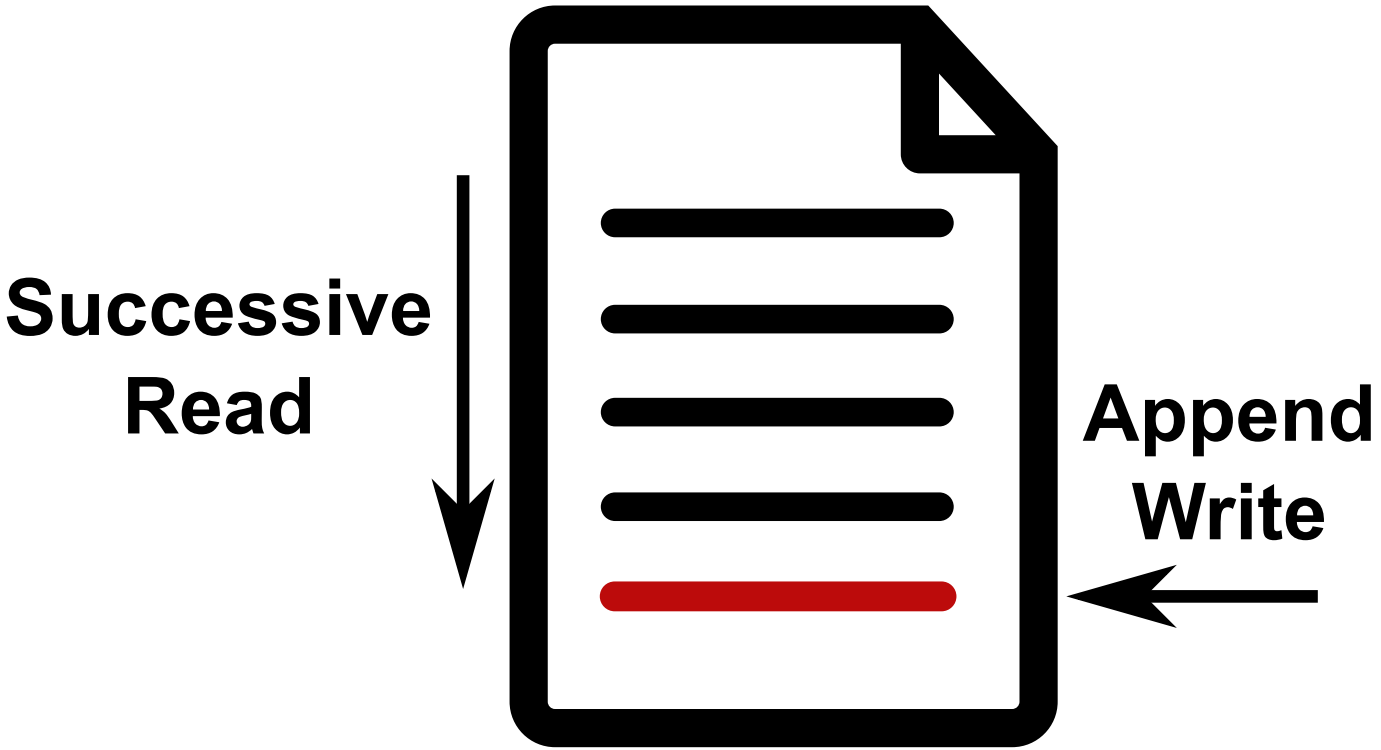

HDFS datanode
Linux file system



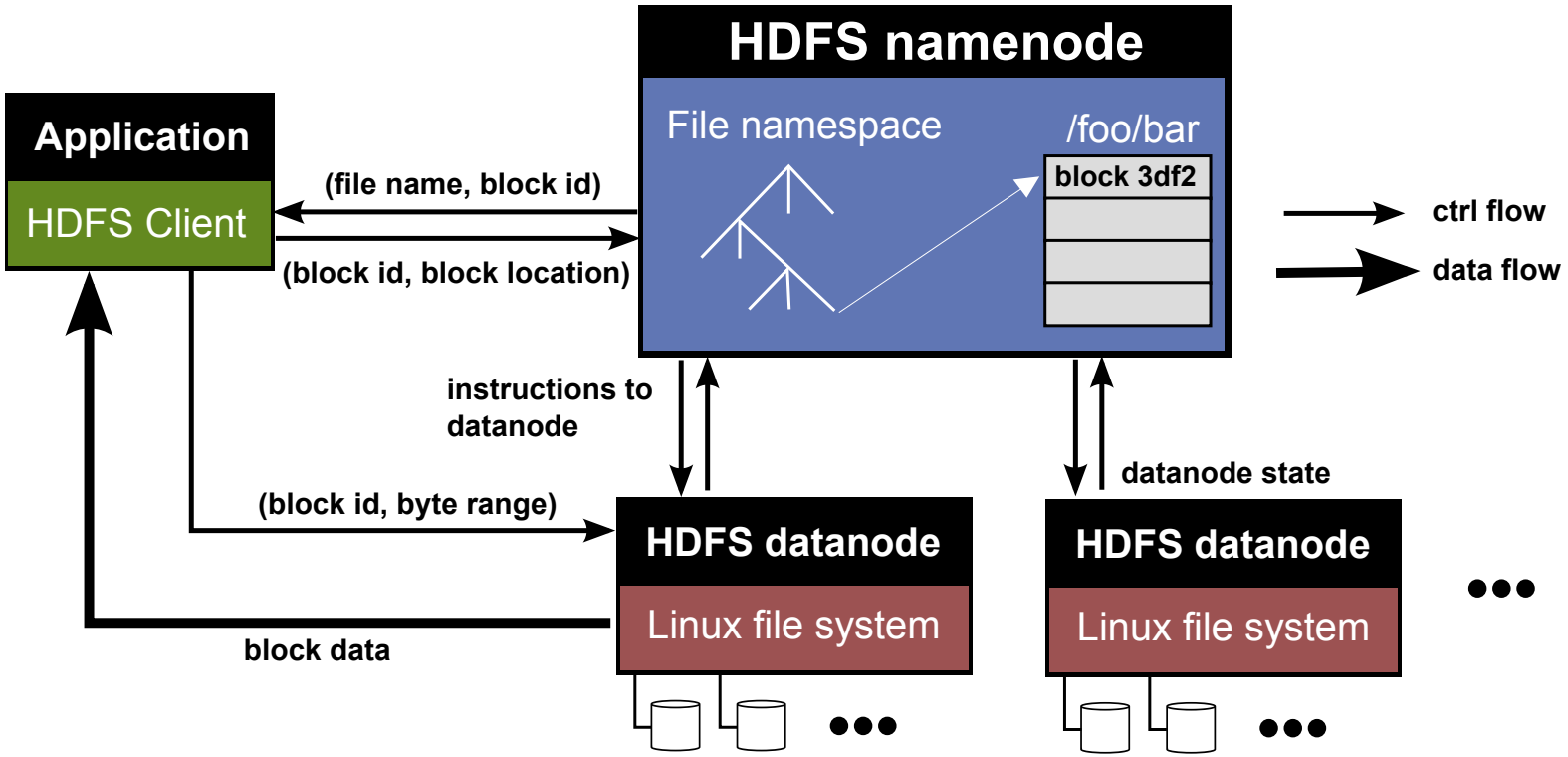
Demo

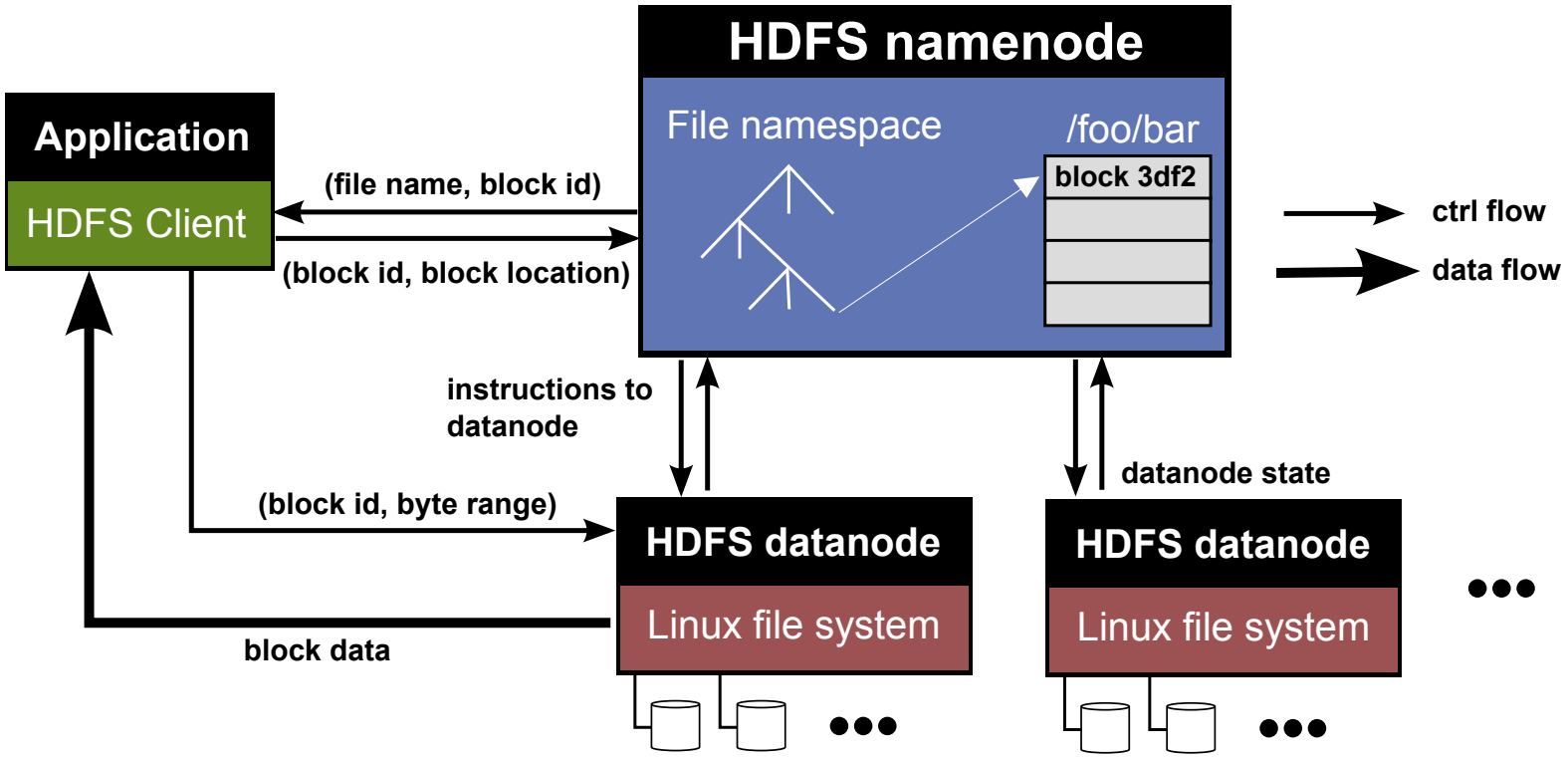

HDFS namenode
File namespace
/foo/bar
block 3df2
instructions to datanode
datanode state
HDFS datanode
Linux file system
HDFS datanode
Linux file system

Optimised for Streaming

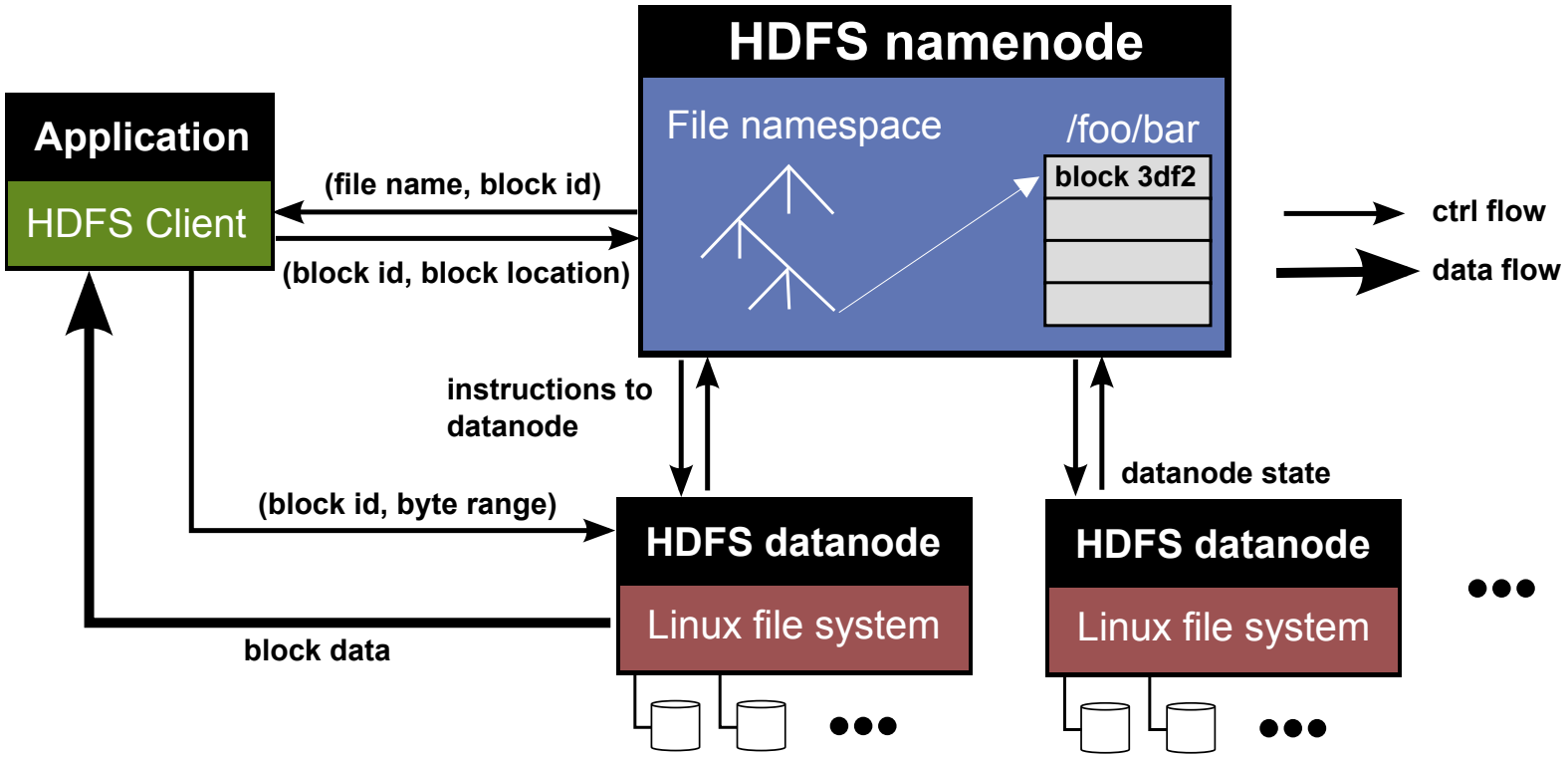


write once read many


HDFS namenode
File namespace
/foo/bar
block 3df2
Application
HDFS Client
(file name, block id)
(block id, block location)
ctrl flow
data flow
instructions to datanode
datanode state
(block id, byte range)
HDFS datanode
Linux file system
HDFS datanode
Linux file system
block data

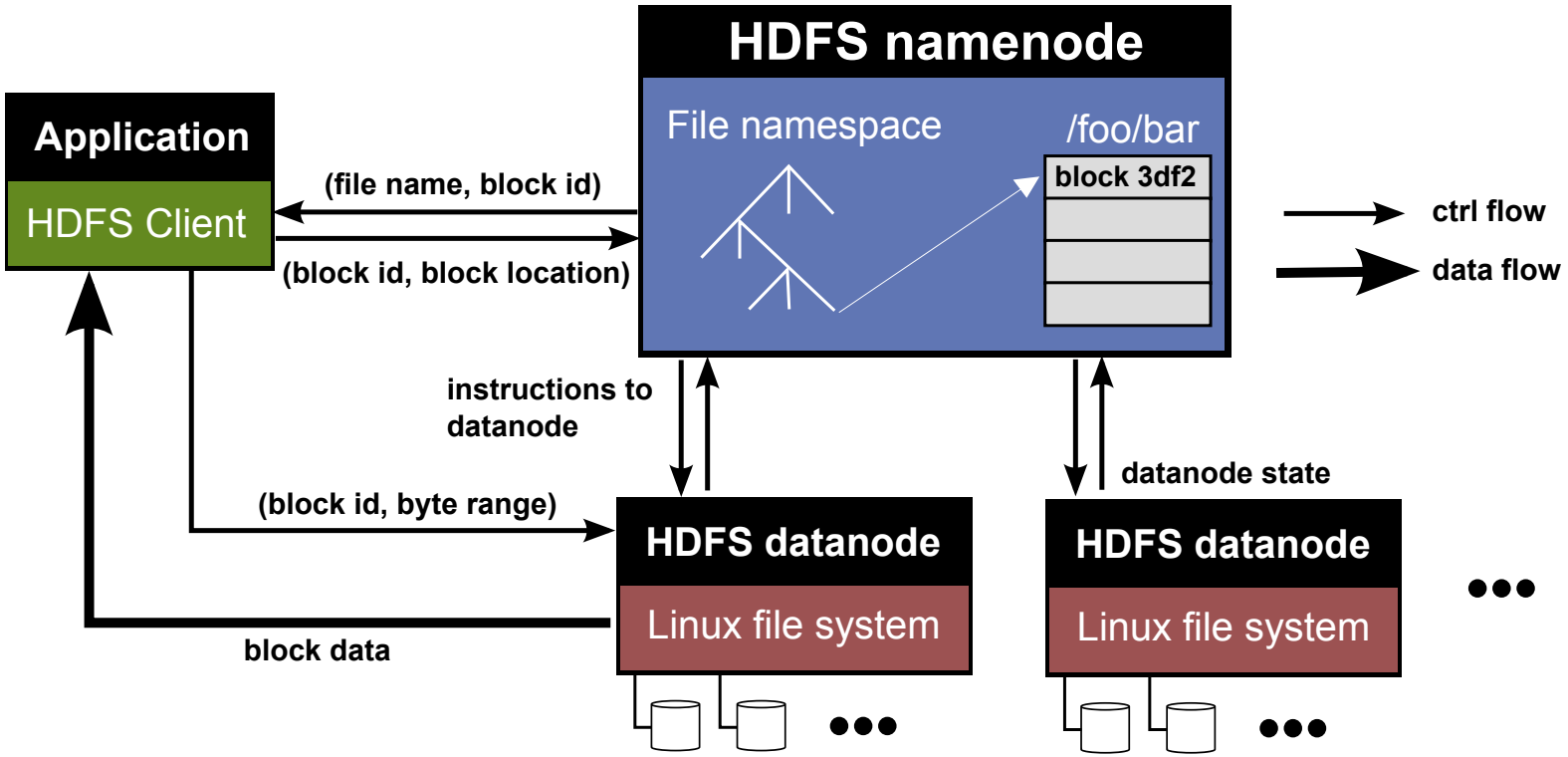


Why so large blocks?



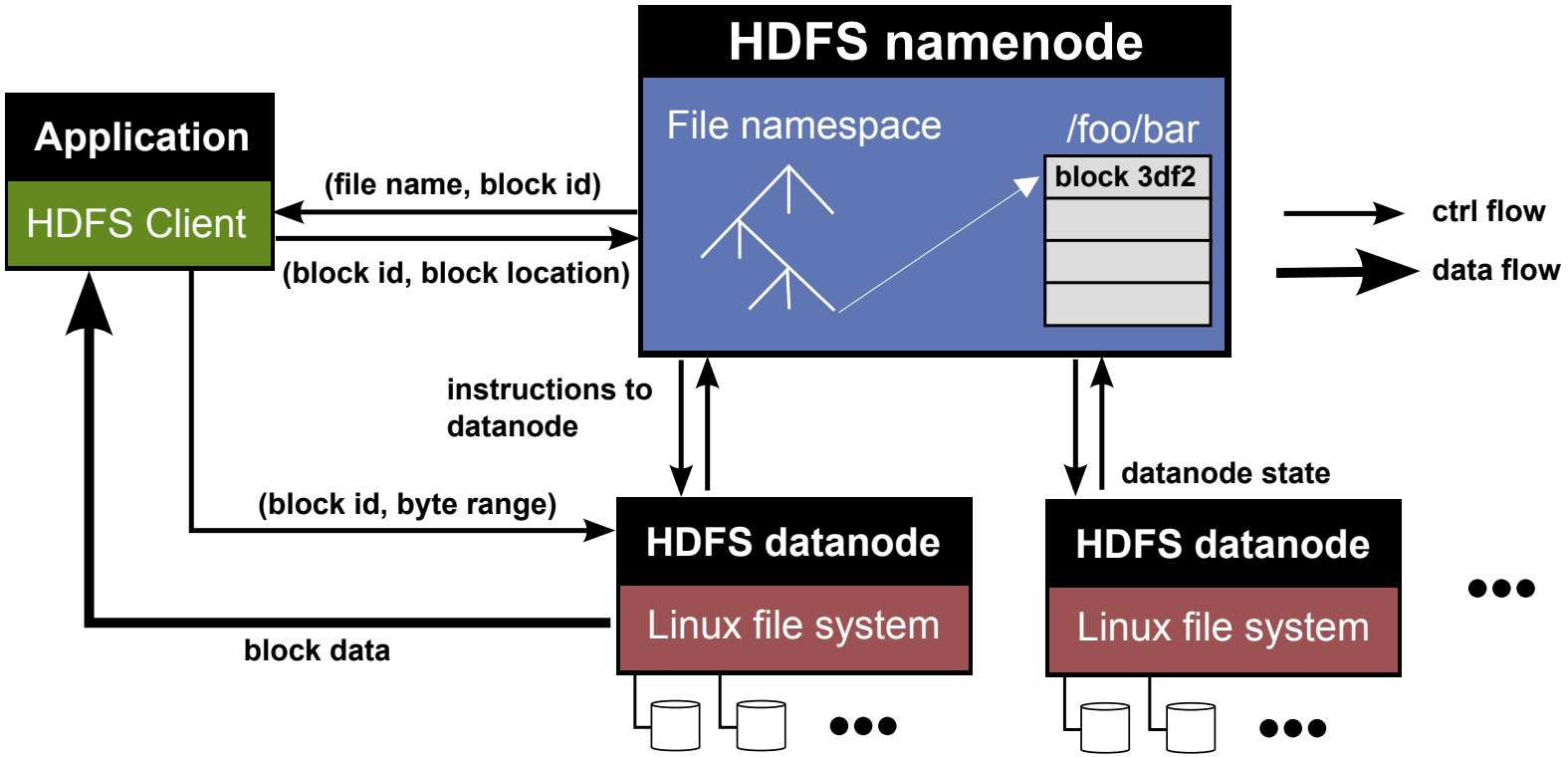
Why so large blocks?

1. less communication between master and workers



Why so large blocks?

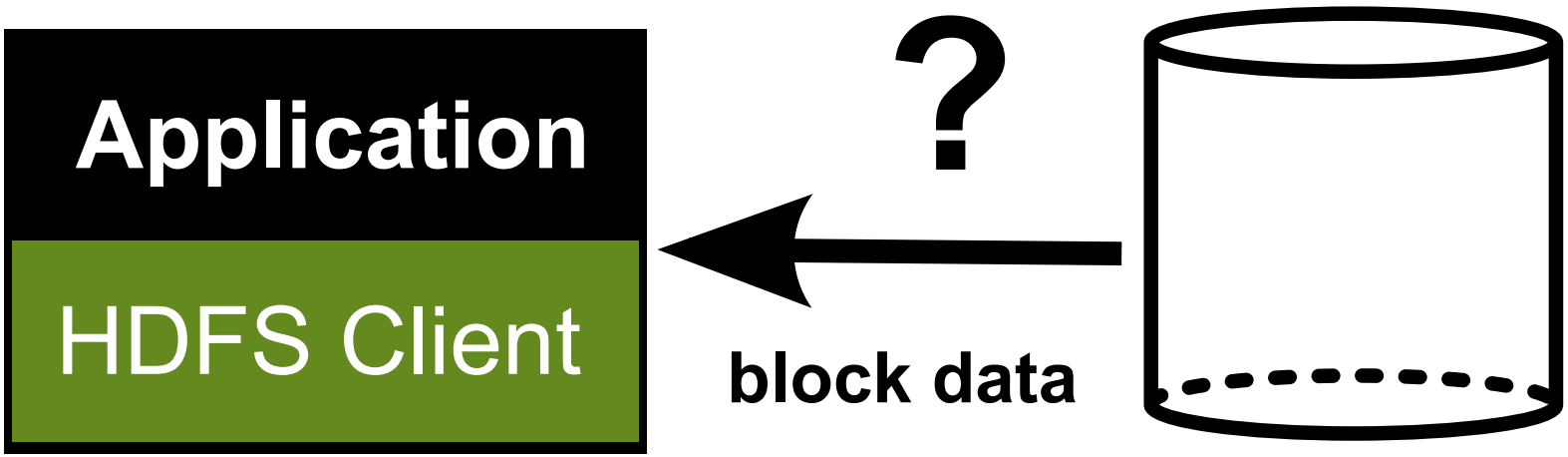
1. less communication between master and workers
2. reduced communication between client and datanodes



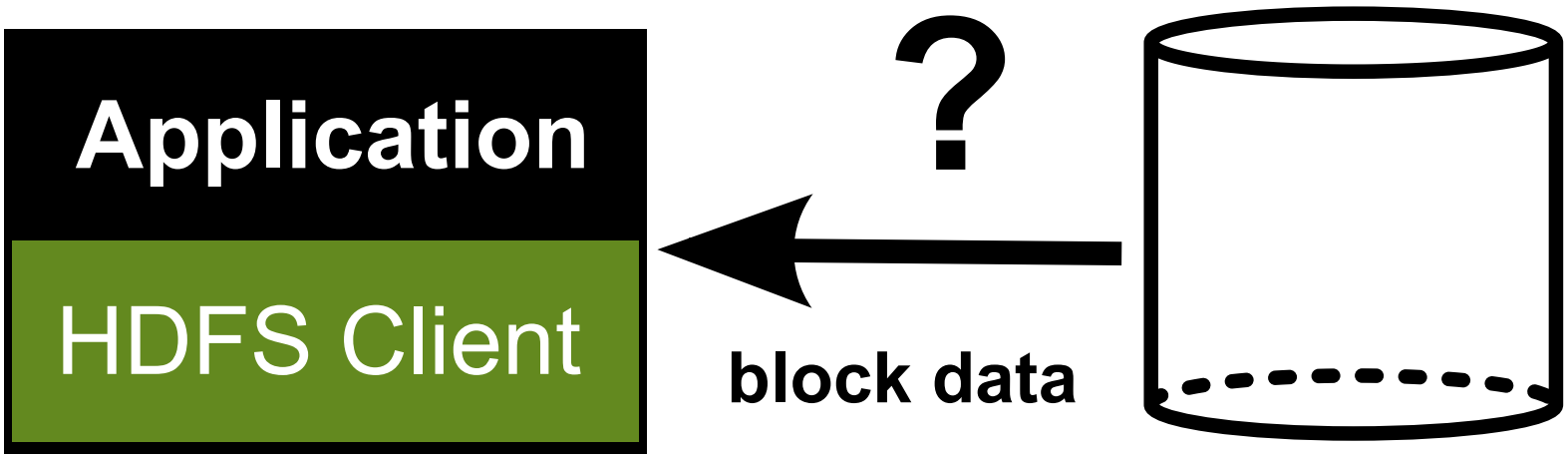
Why so large blocks?

1. less communication between master and workers
2. reduced communication between client and datanodes
3. less meta data to be saved in namenode

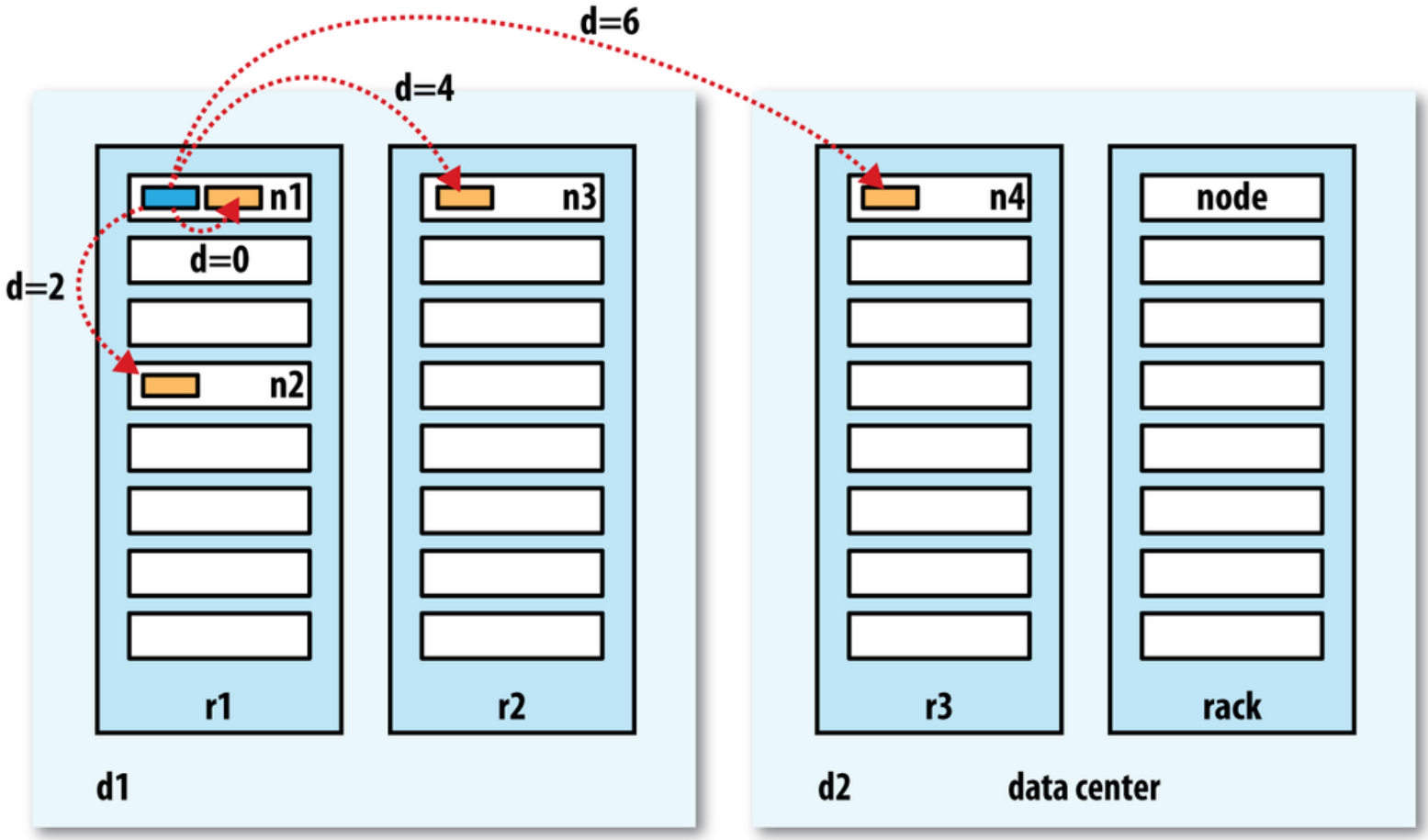
Which block location is best for the client?



Which block location is best for the client?



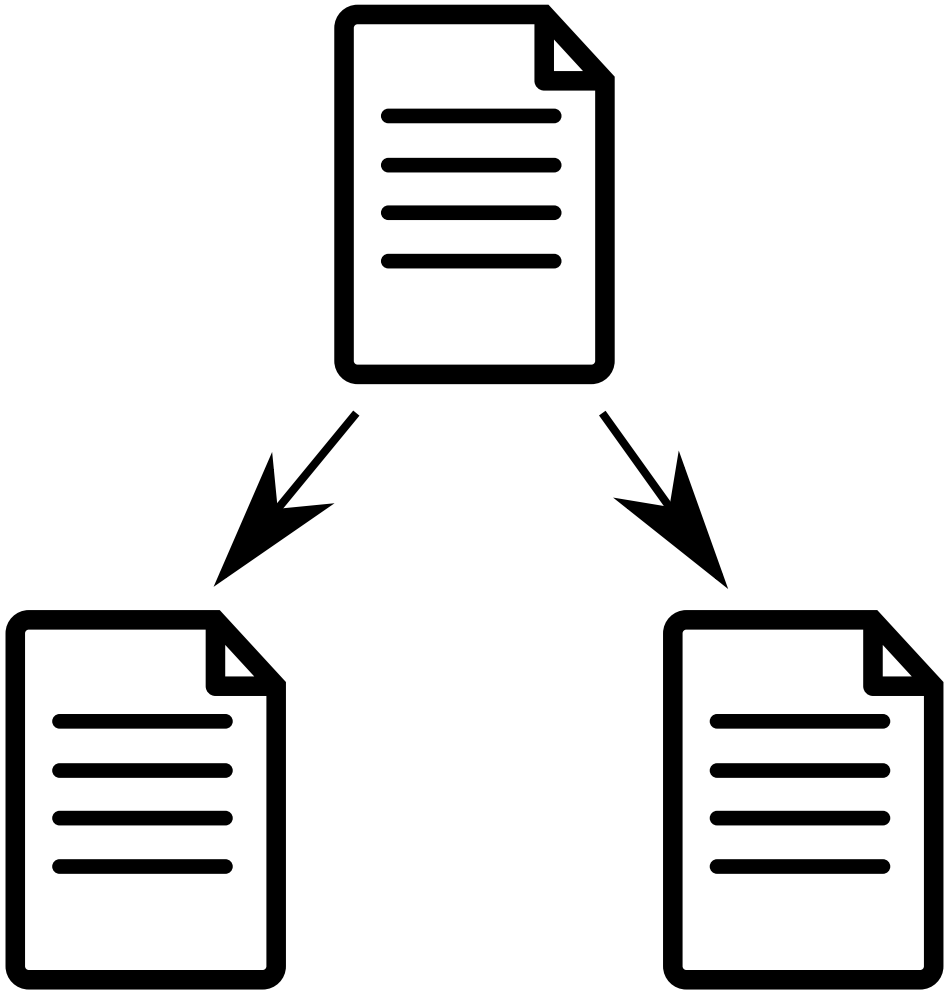
The closest one!



Network is represented as a tree.
Distance between two nodes is the sum of their distance to their closest common ancestor.

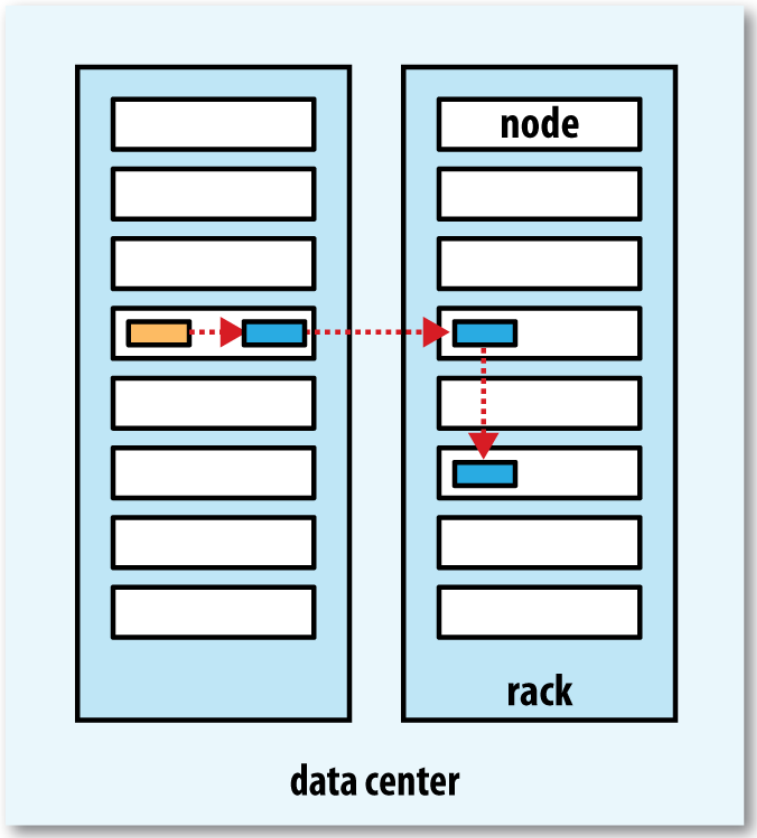
Fault Tolerance

Faults are the norm, not the exception.

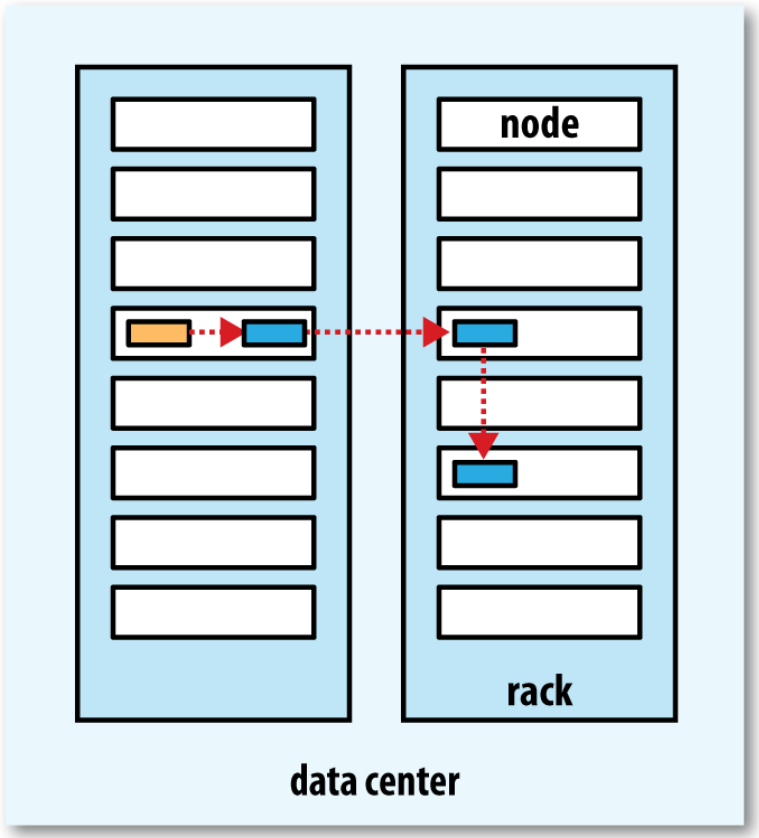


Hadoop keeps three versions by default.

How to spread over across the cluster?

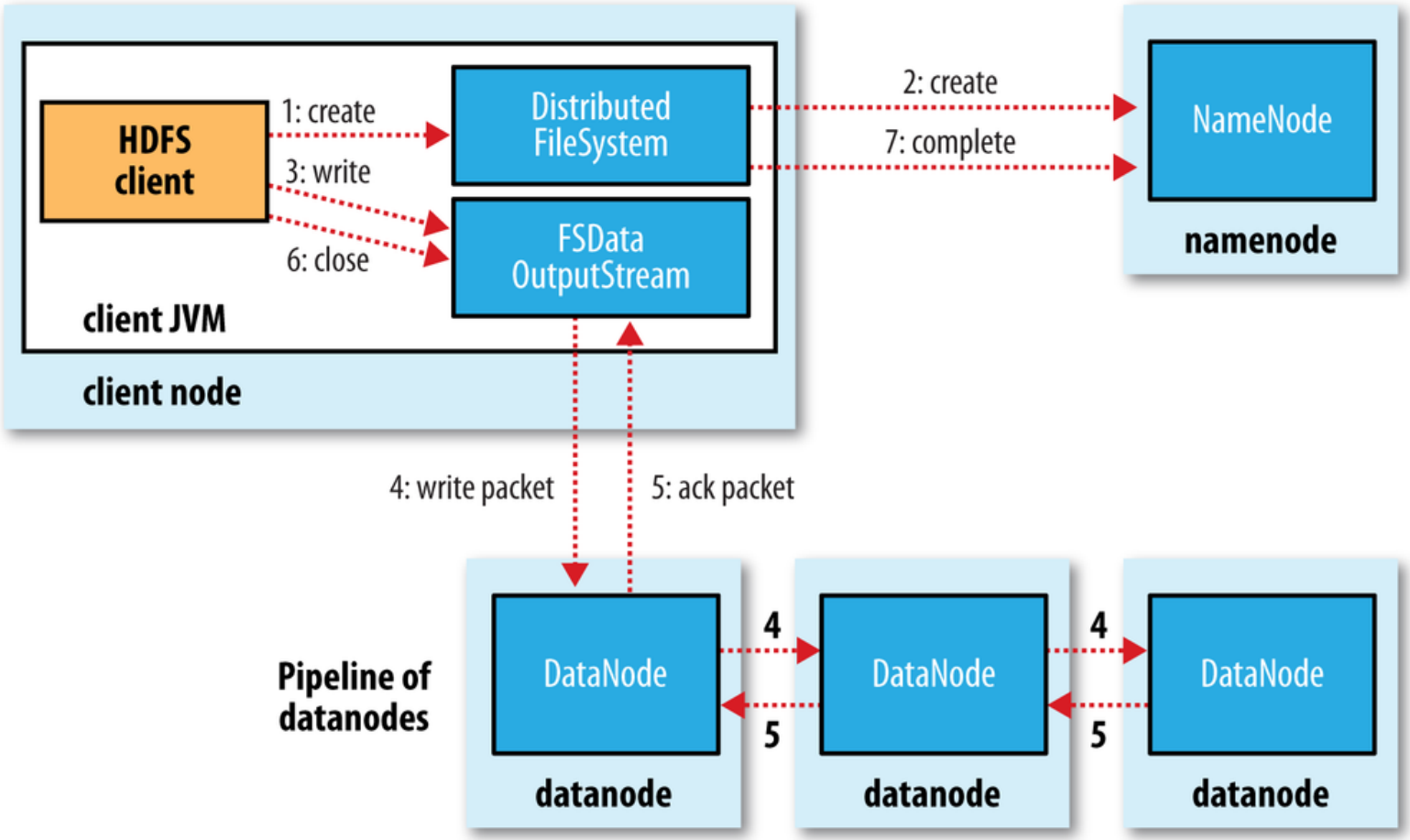


How to spread over across the cluster?



Demo

Anatomy of a Write



Summary

1. HDFS handles large files across the cluster
2. HDFS is optimised for streaming access to files
3. HDFS runs on commodity hardware and needs to be fault tolerant