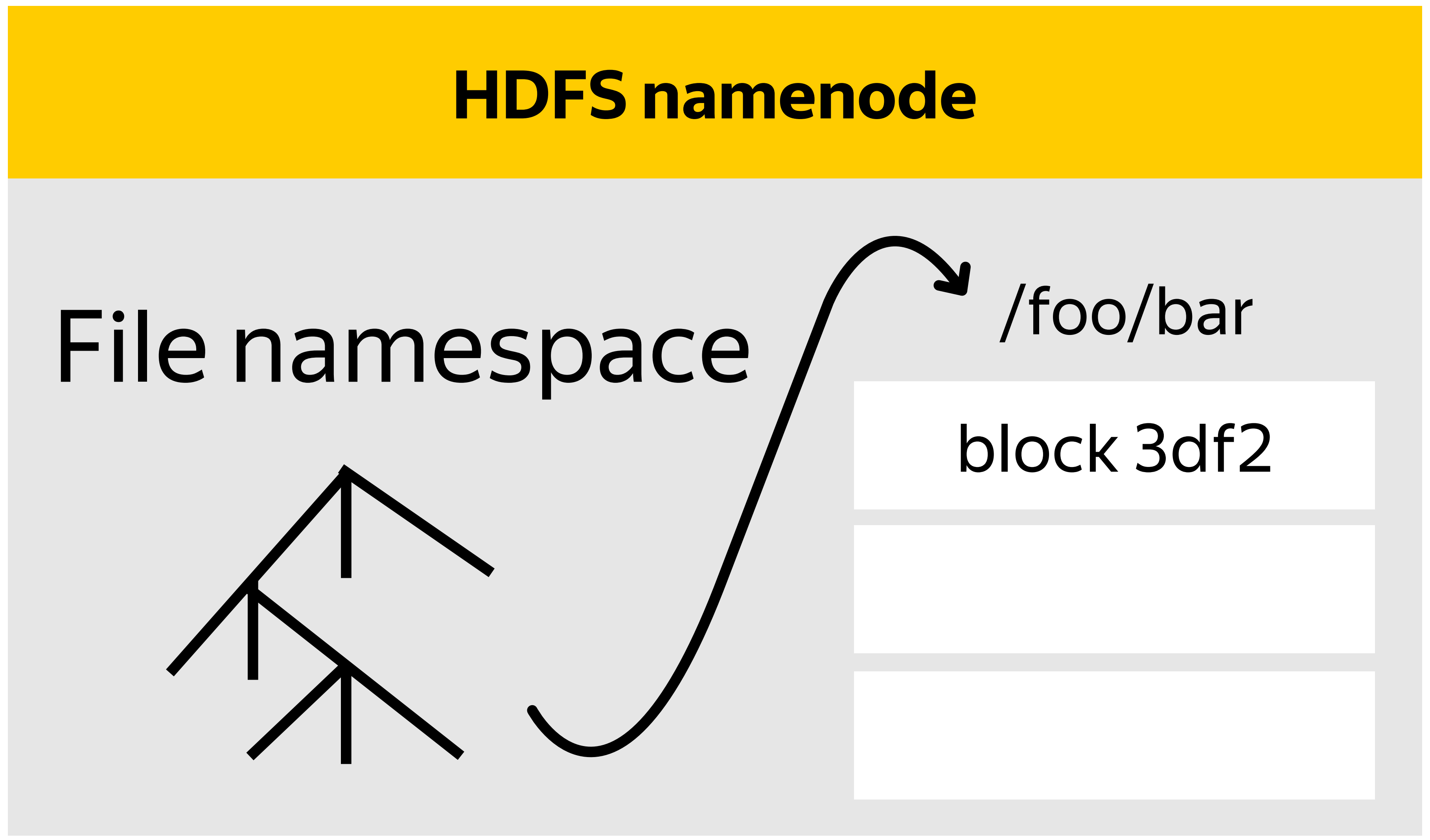
Yandex

# HDFS

Namenode Architecture

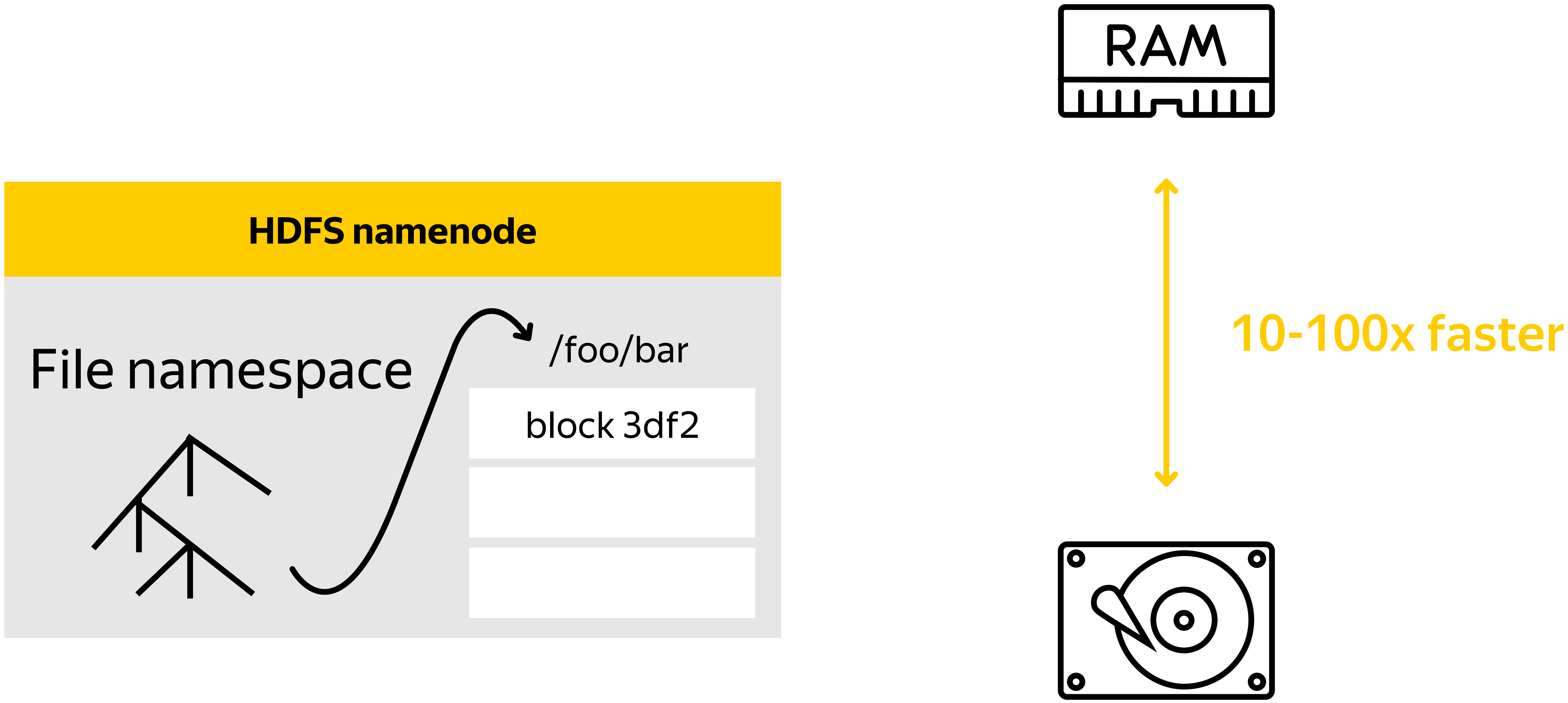


HDFS namenode

File namespace

/foo/bar

block 3df2



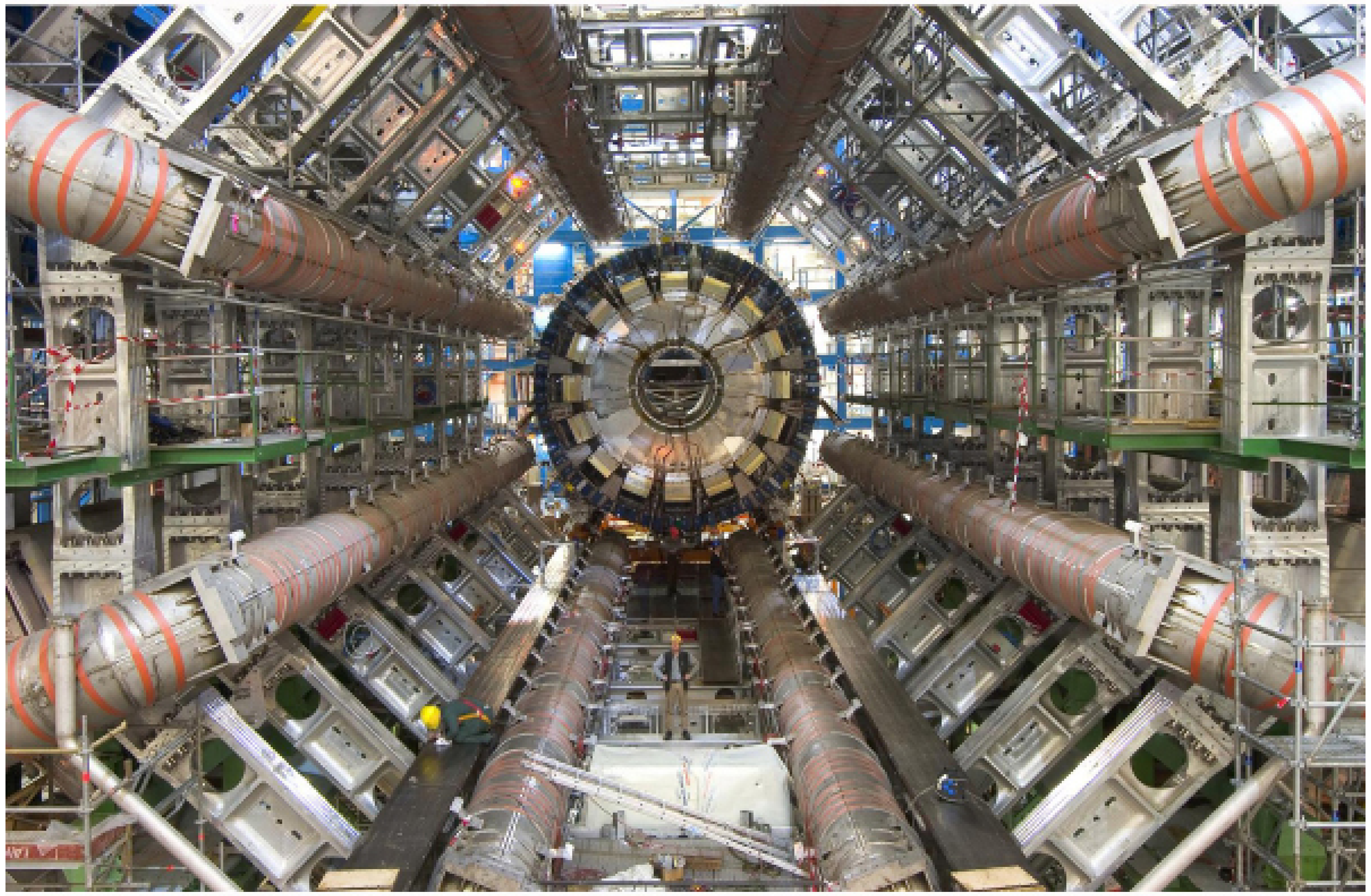
HDFS namenode

File namespace

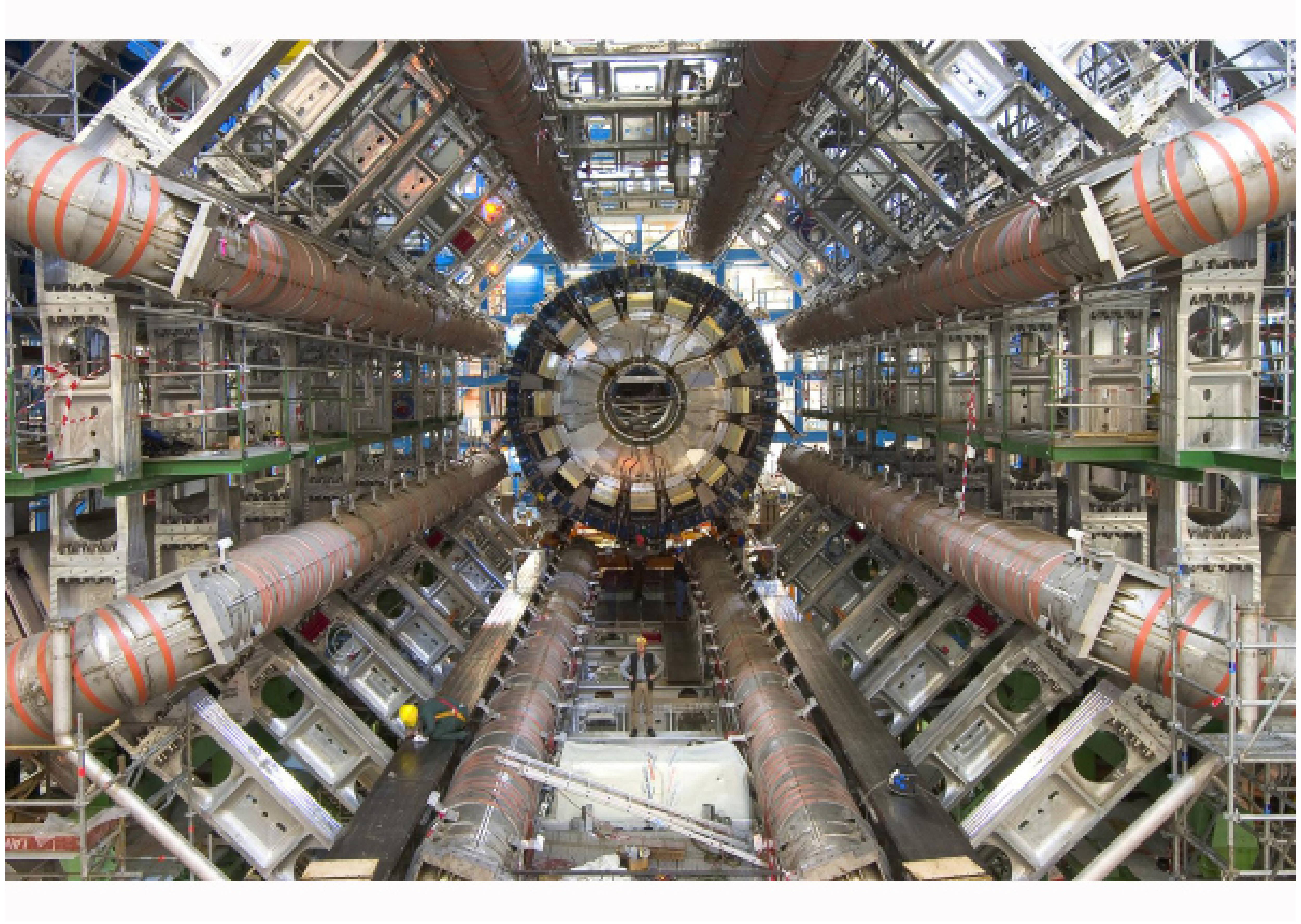
/foo/bar

block 3df2

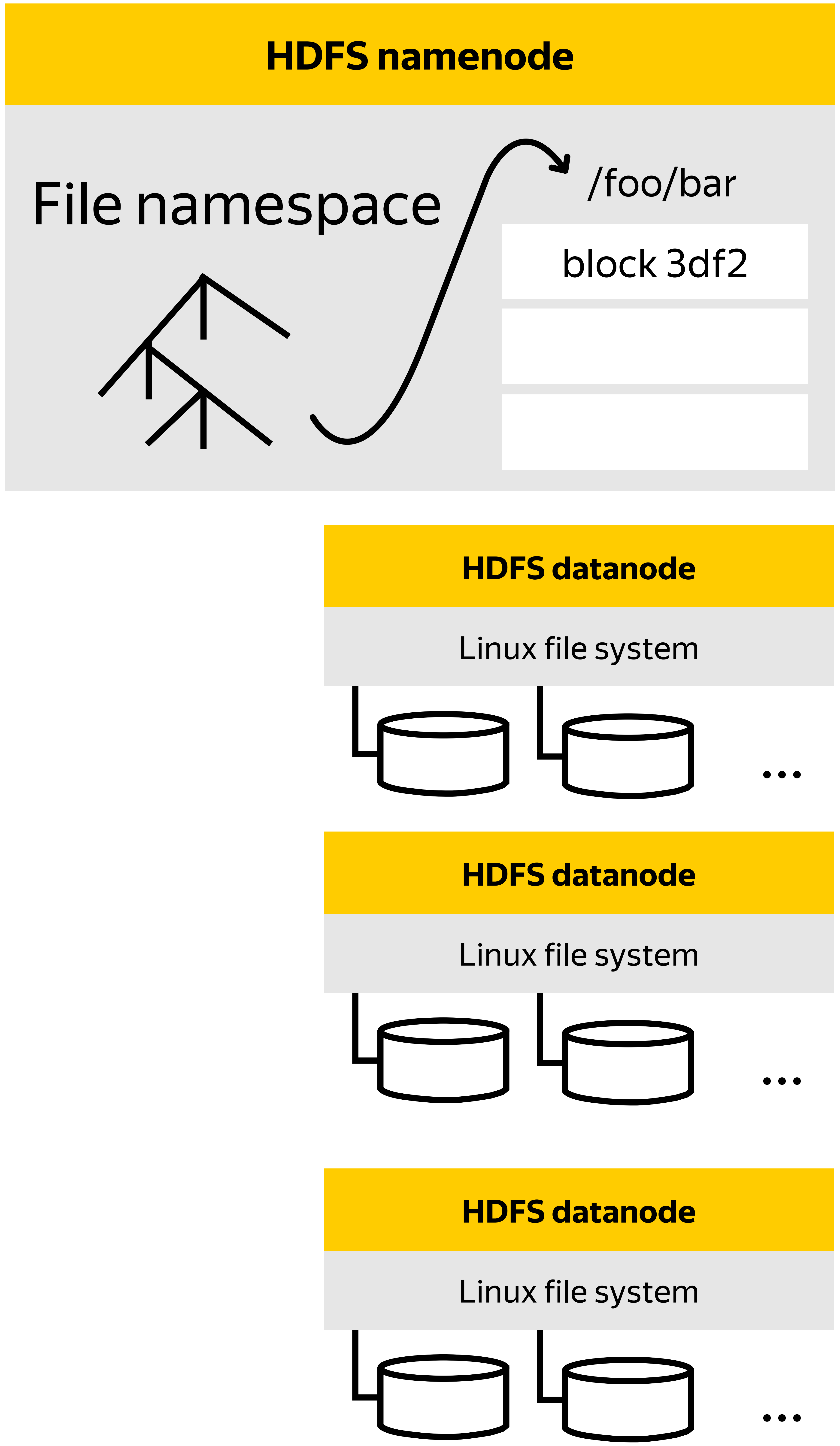
10-100x faster



1 year ~ 10 PB



1 year ~ 10 PB



HDFS namenode

File namespace

/foo/bar

block 3df2

HDFS datanode

Linux file system

...

HDFS datanode

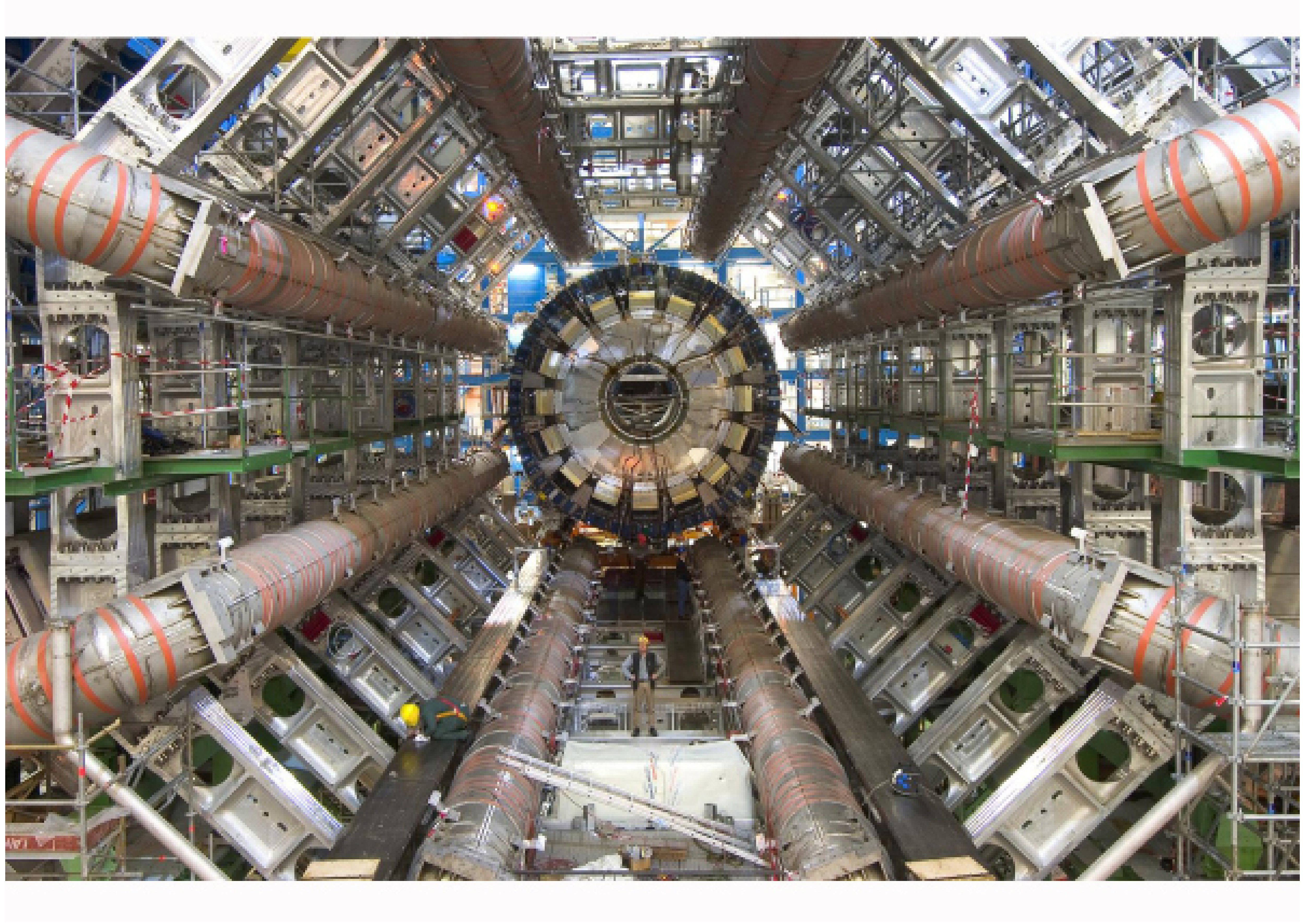
Linux file system

...

HDFS datanode

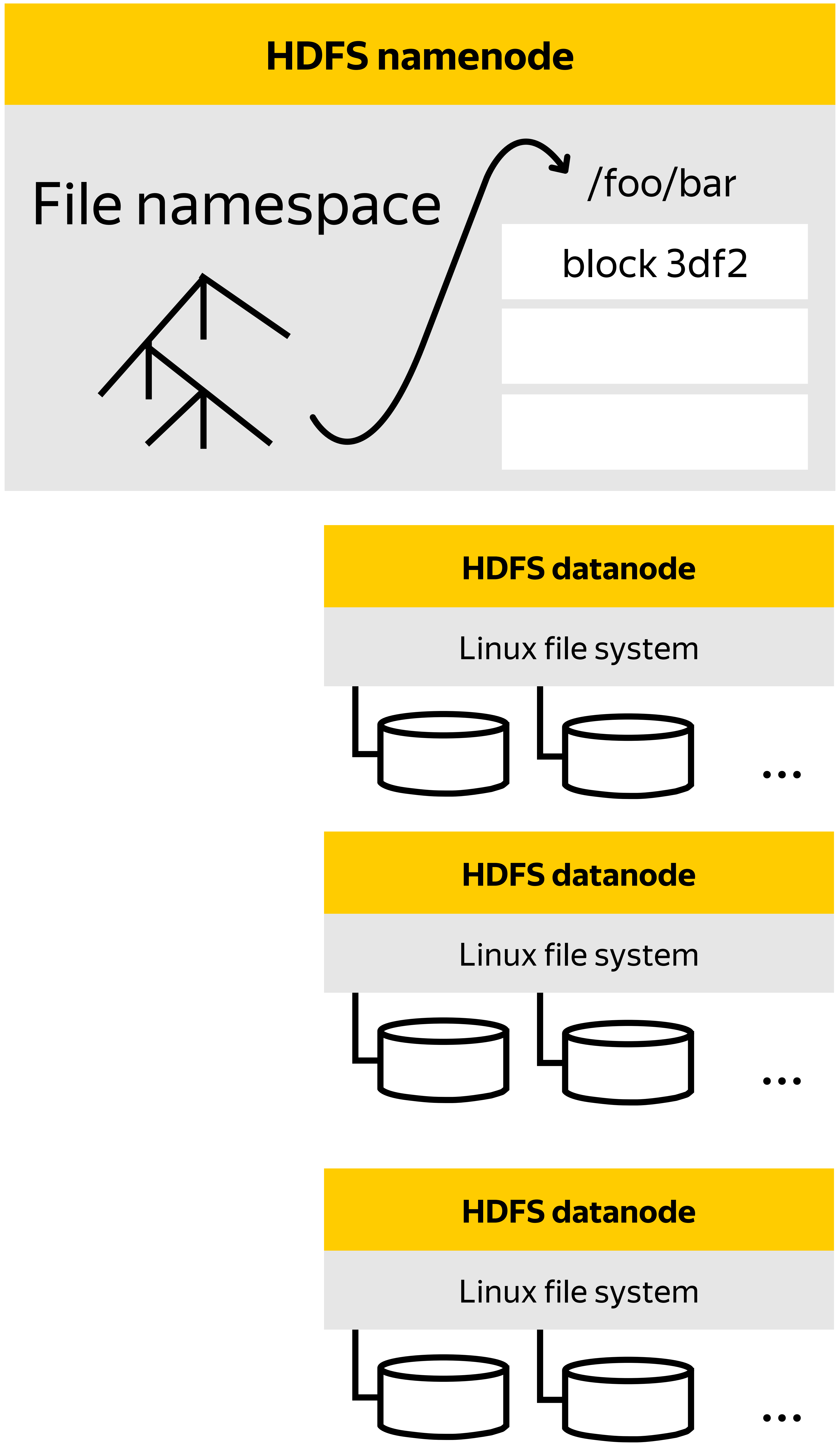
Linux file system

...



1 year ~ 10 PB

10 PB / 2 TB * 3 ~ 15 k



HDFS namenode

File namespace

/foo/bar

block 3df2

HDFS datanode

Linux file system

...

HDFS datanode

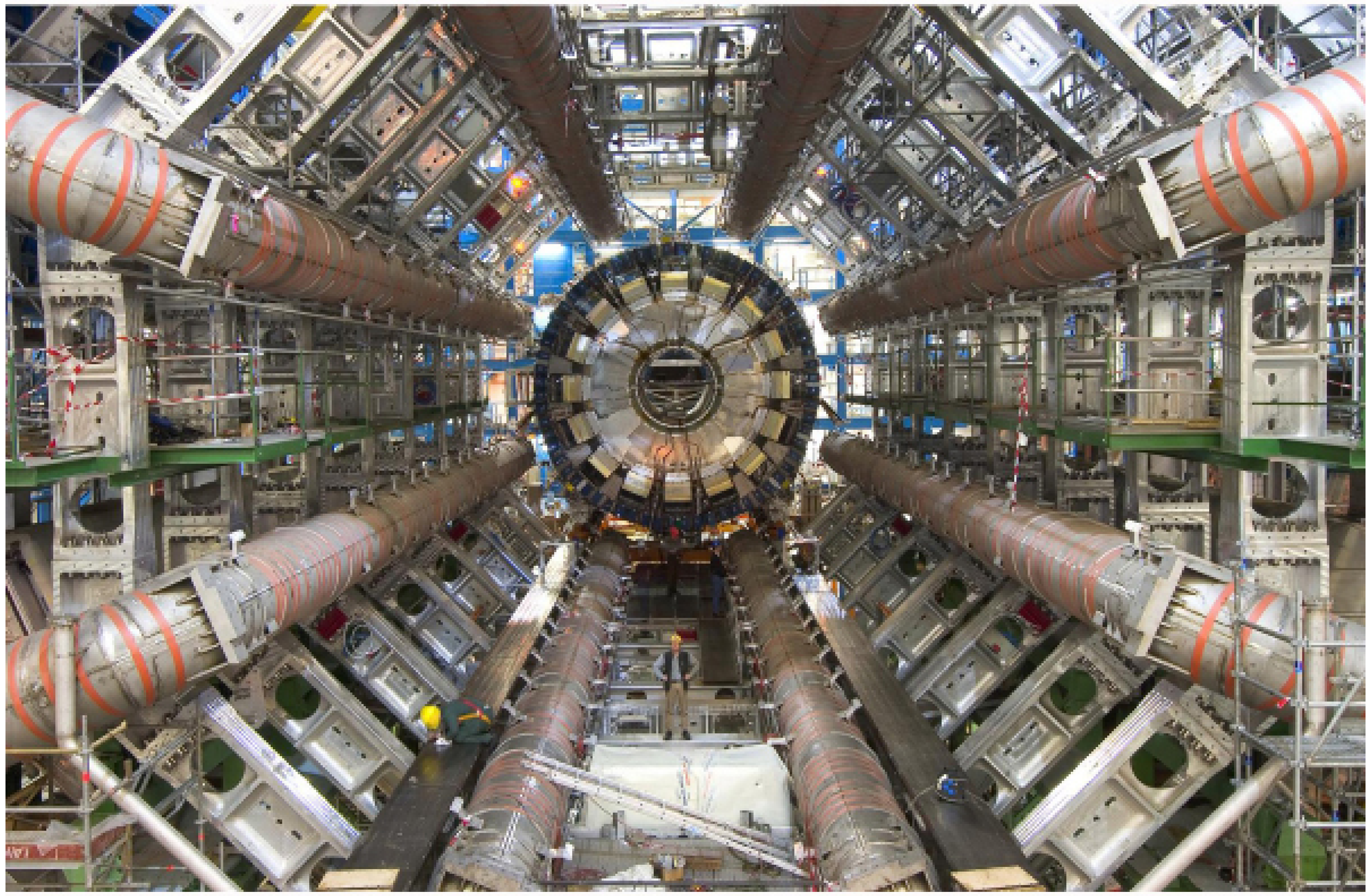
Linux file system

...

HDFS datanode

Linux file system

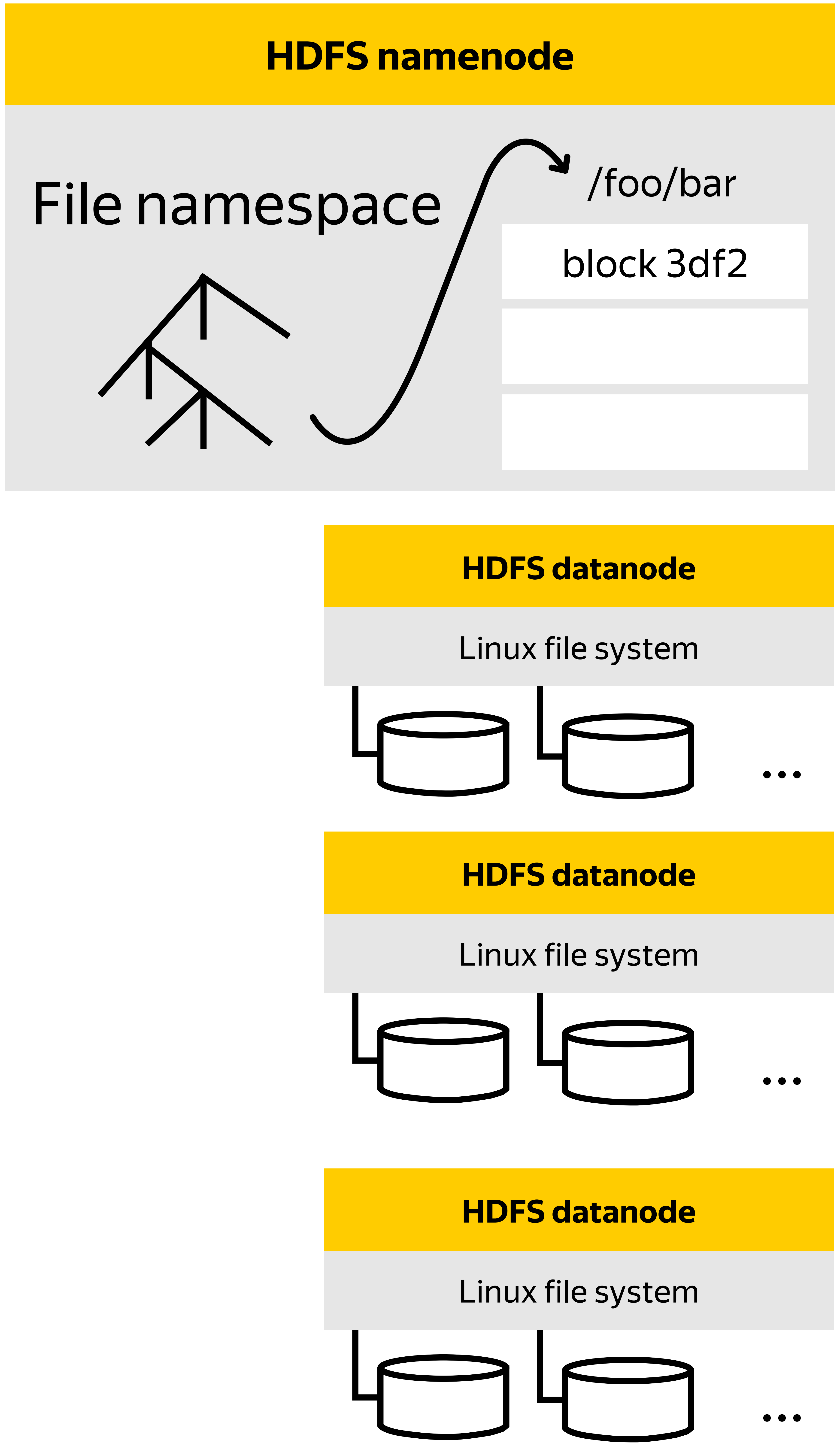
...



1 year ~ 10 PB

10 PB / 2 TB * 3 ~ 15 k

RAM ???



HDFS namenode

File namespace

/foo/bar

block 3df2

HDFS datanode

Linux file system

...

HDFS datanode

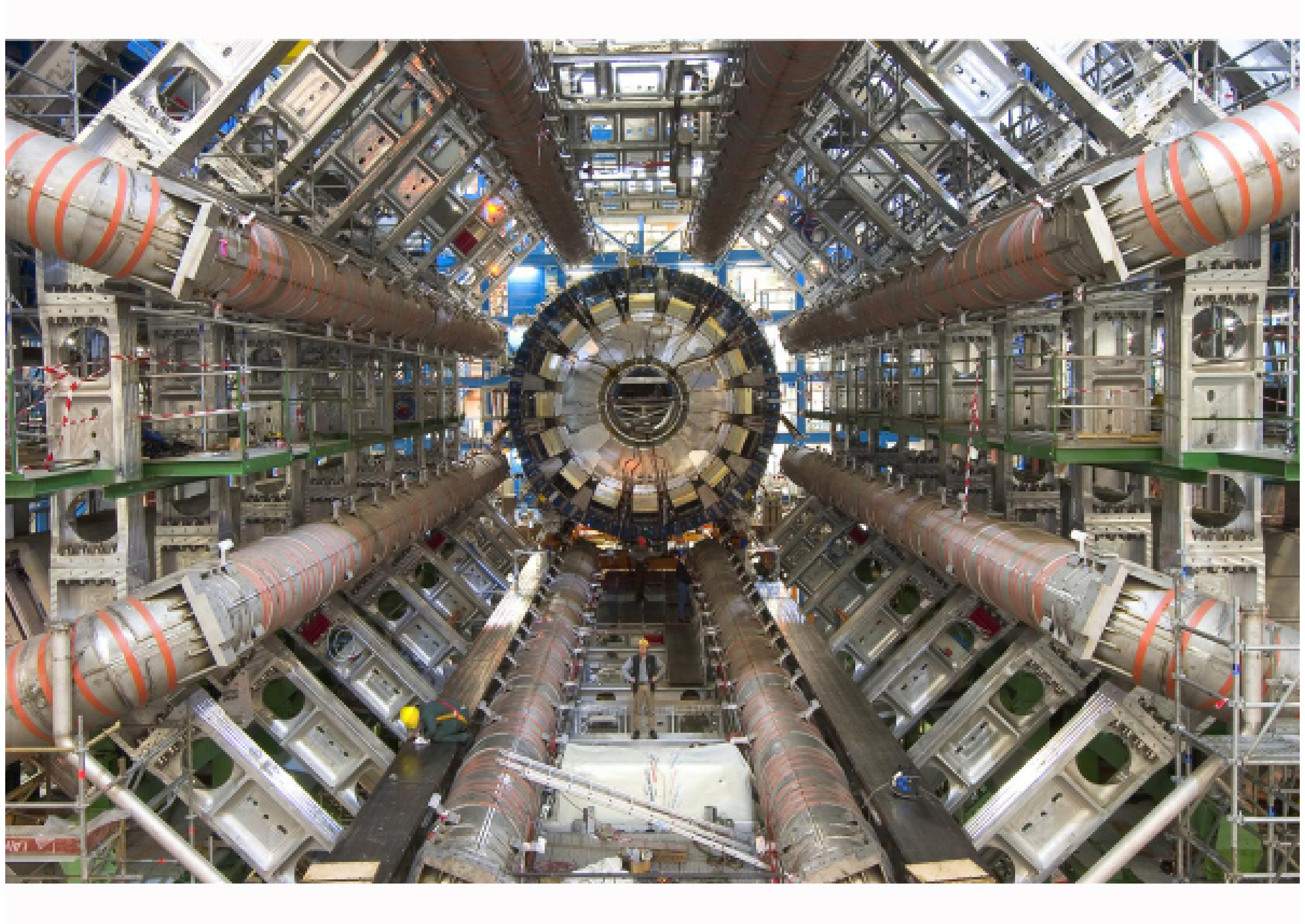
Linux file system

...

HDFS datanode

Linux file system

...

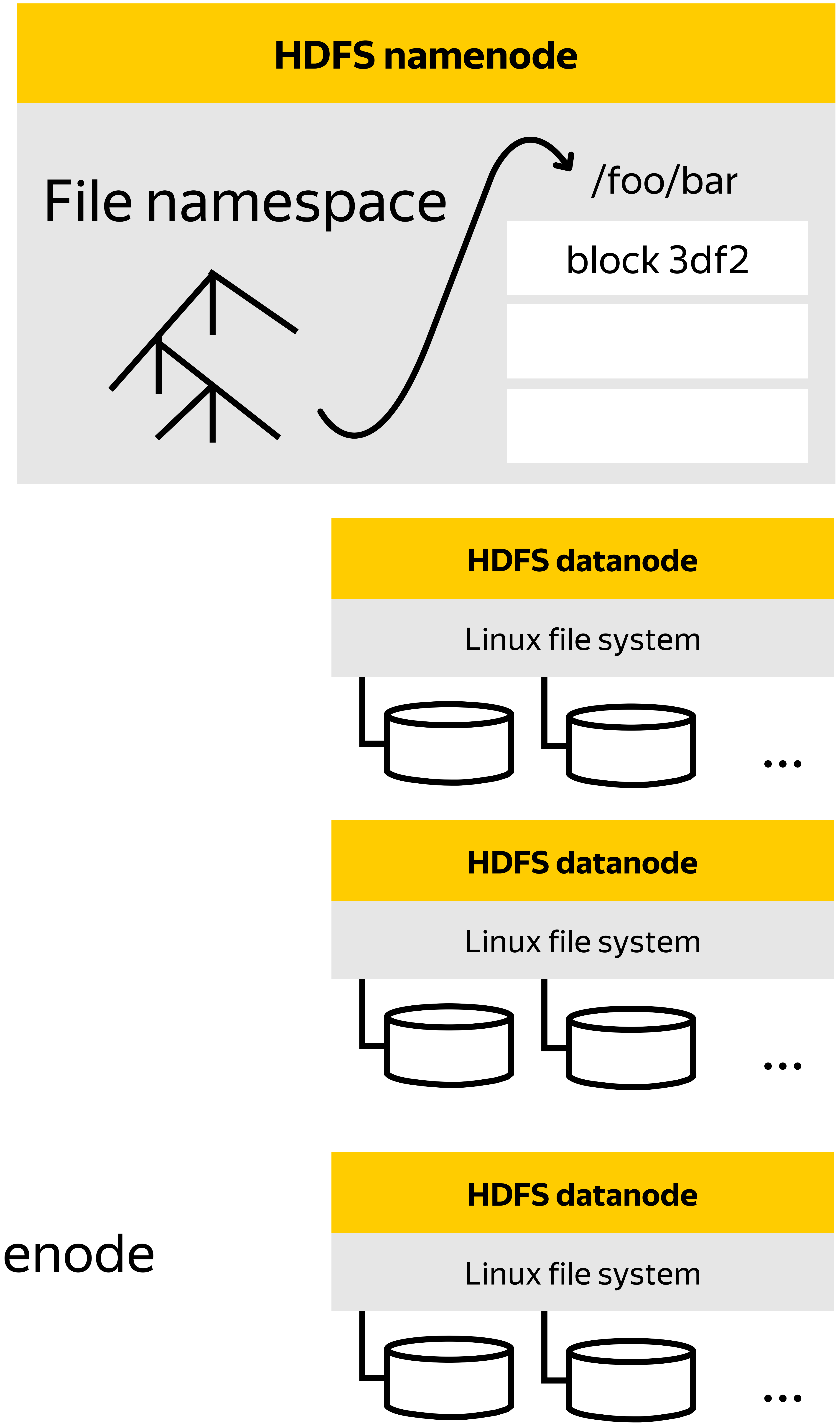


1 year ~ 10 PB

10 PB / 2 TB * 3 ~ 15 k

150 B - average block size on Namenode

https://issues.apache.org/jira/browse/HADOOP-1687



**HDFS namenode**

File namespace

/foo/bar

block 3df2

**HDFS datanode**

Linux file system

...

**HDFS datanode**

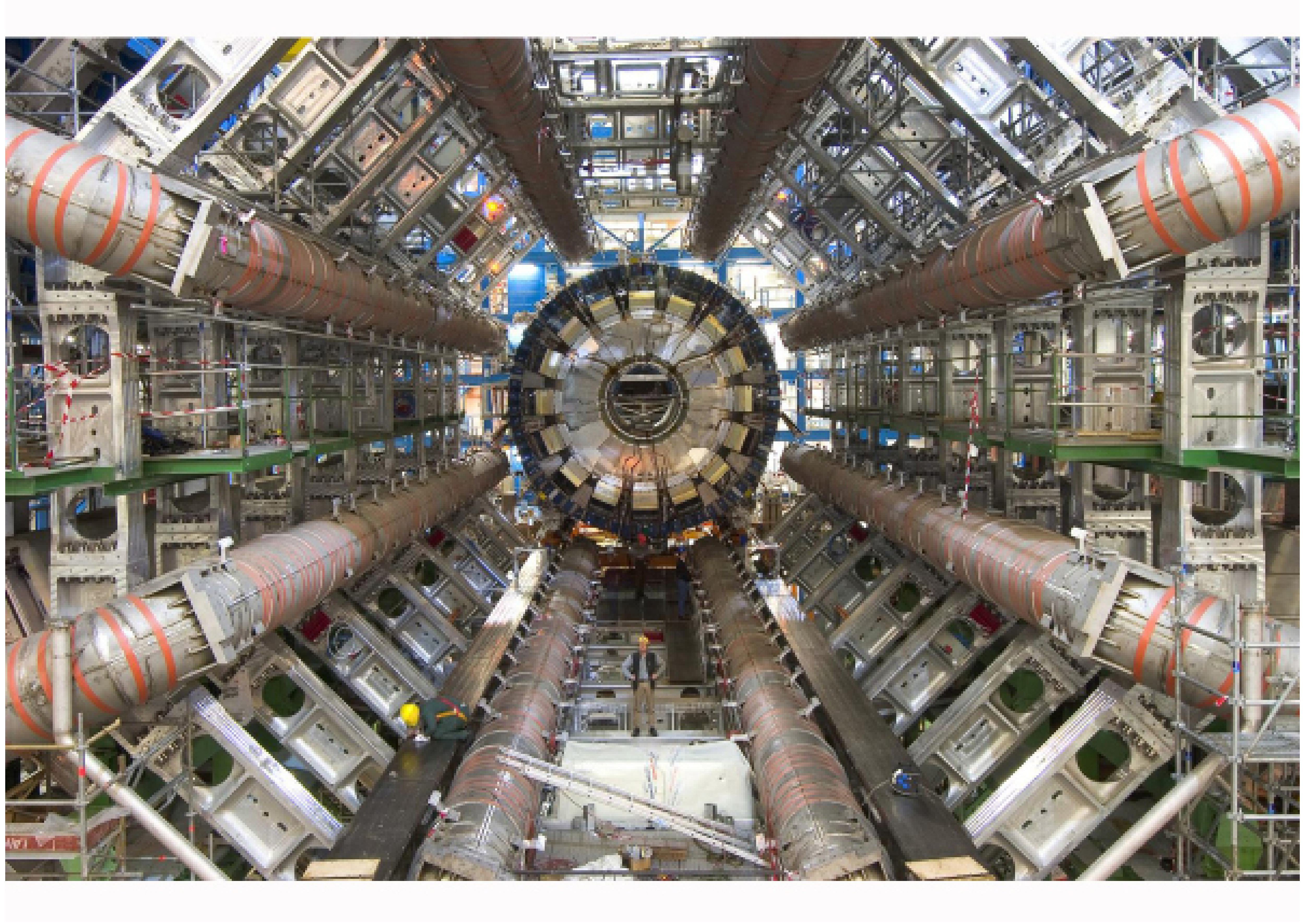
Linux file system

...

**HDFS datanode**

Linux file system

...

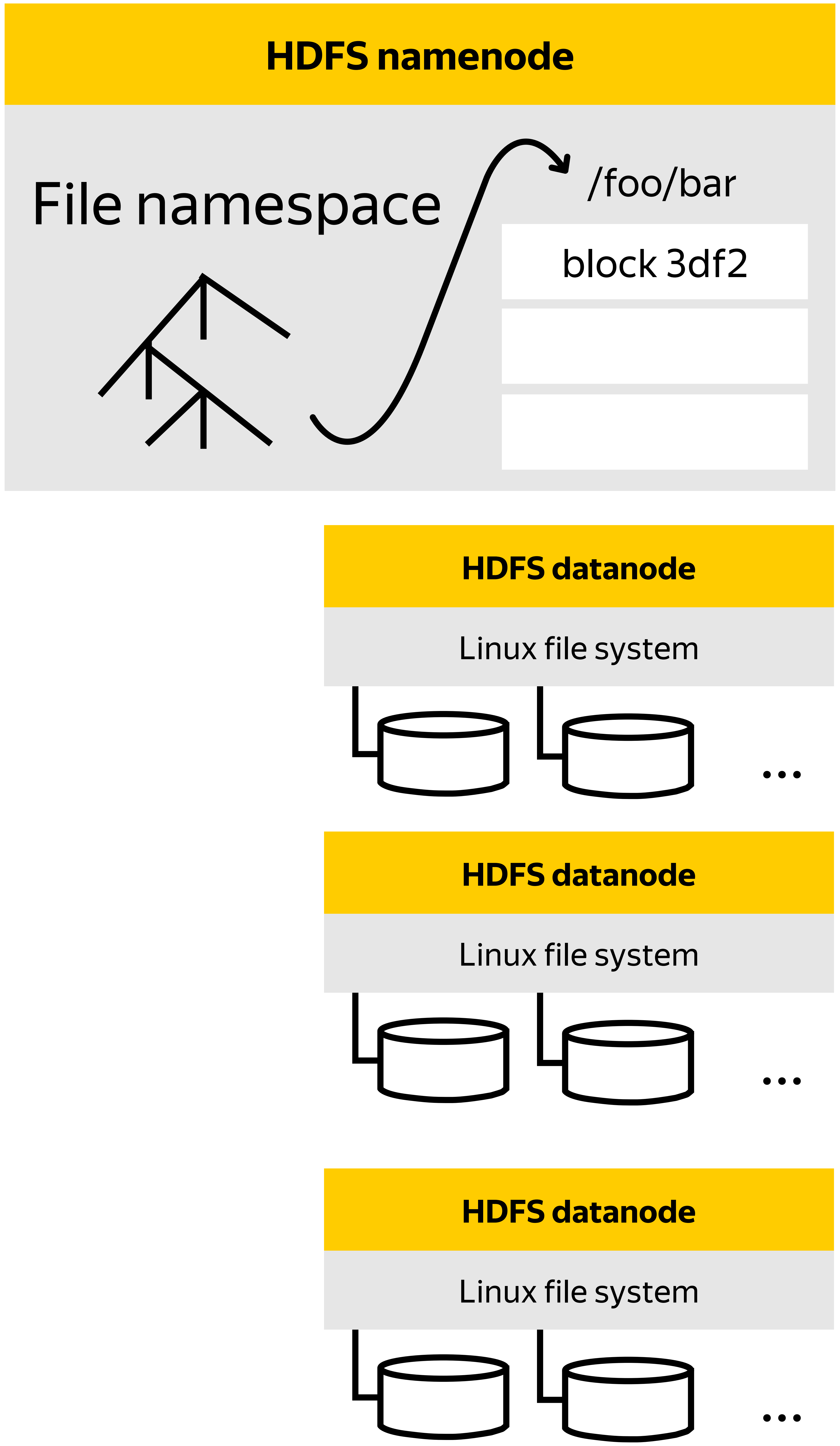


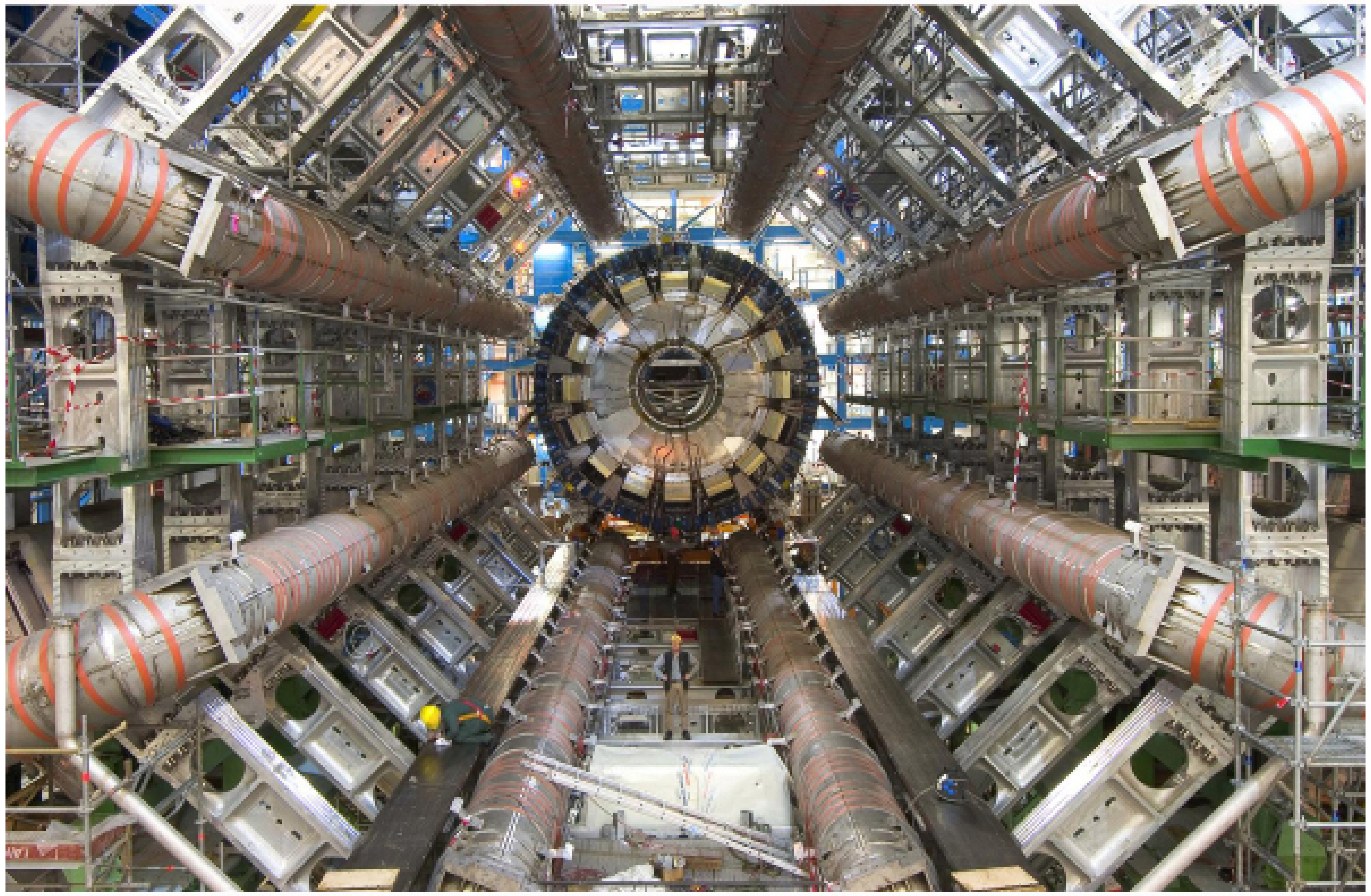
1 year ~ 10 PB

10 PB / 2 TB * 3 ~ 15 k

RAM

10 PB / 128 MB * 3 * 150 B ~ 35 GB



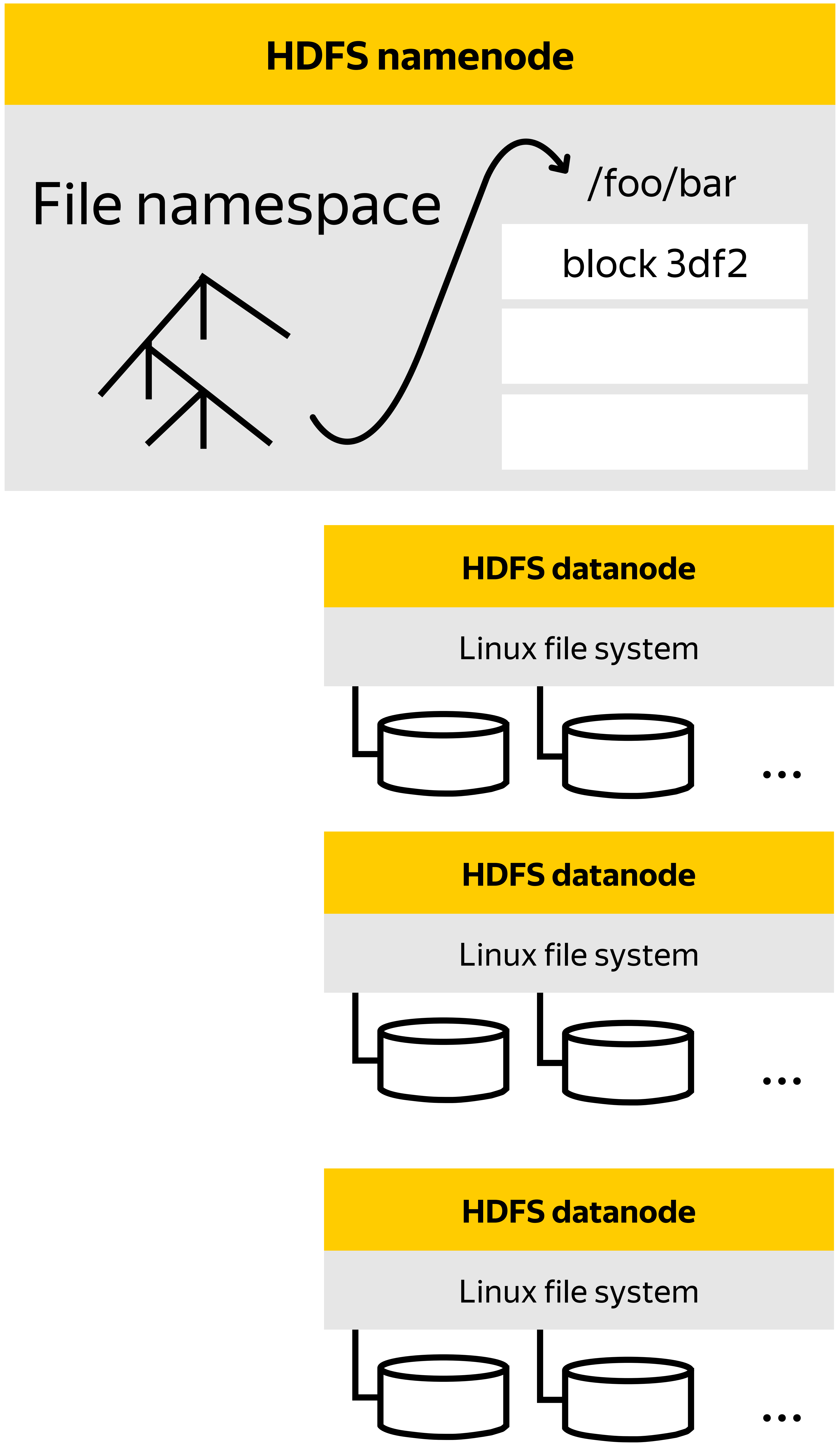


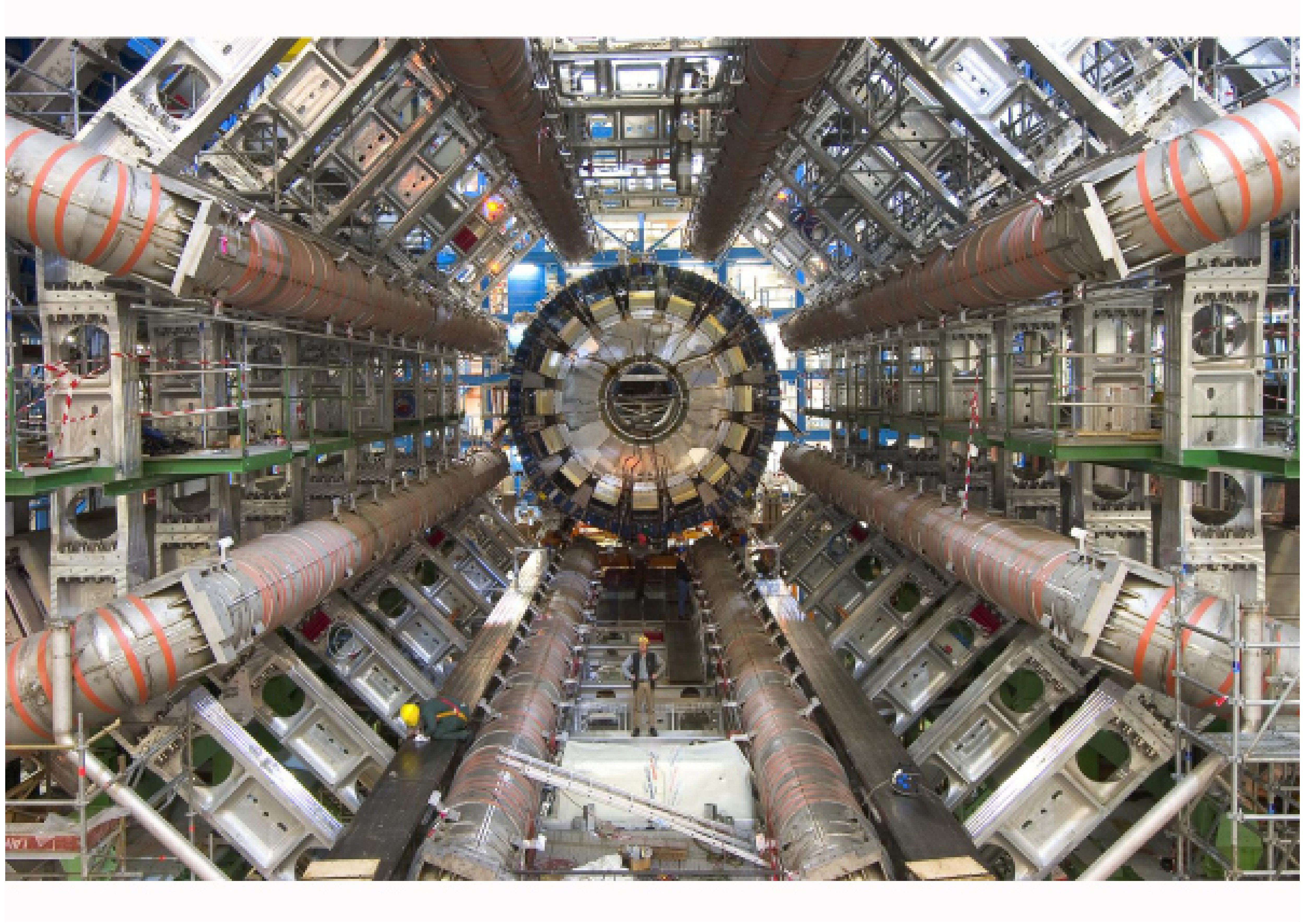
1 year ~ 10 PB

10 PB / 2 TB * 3 ~ 15 k

RAM 10 PB / 128 MB * 3 * 150 B ~ 35 GB

small files problem





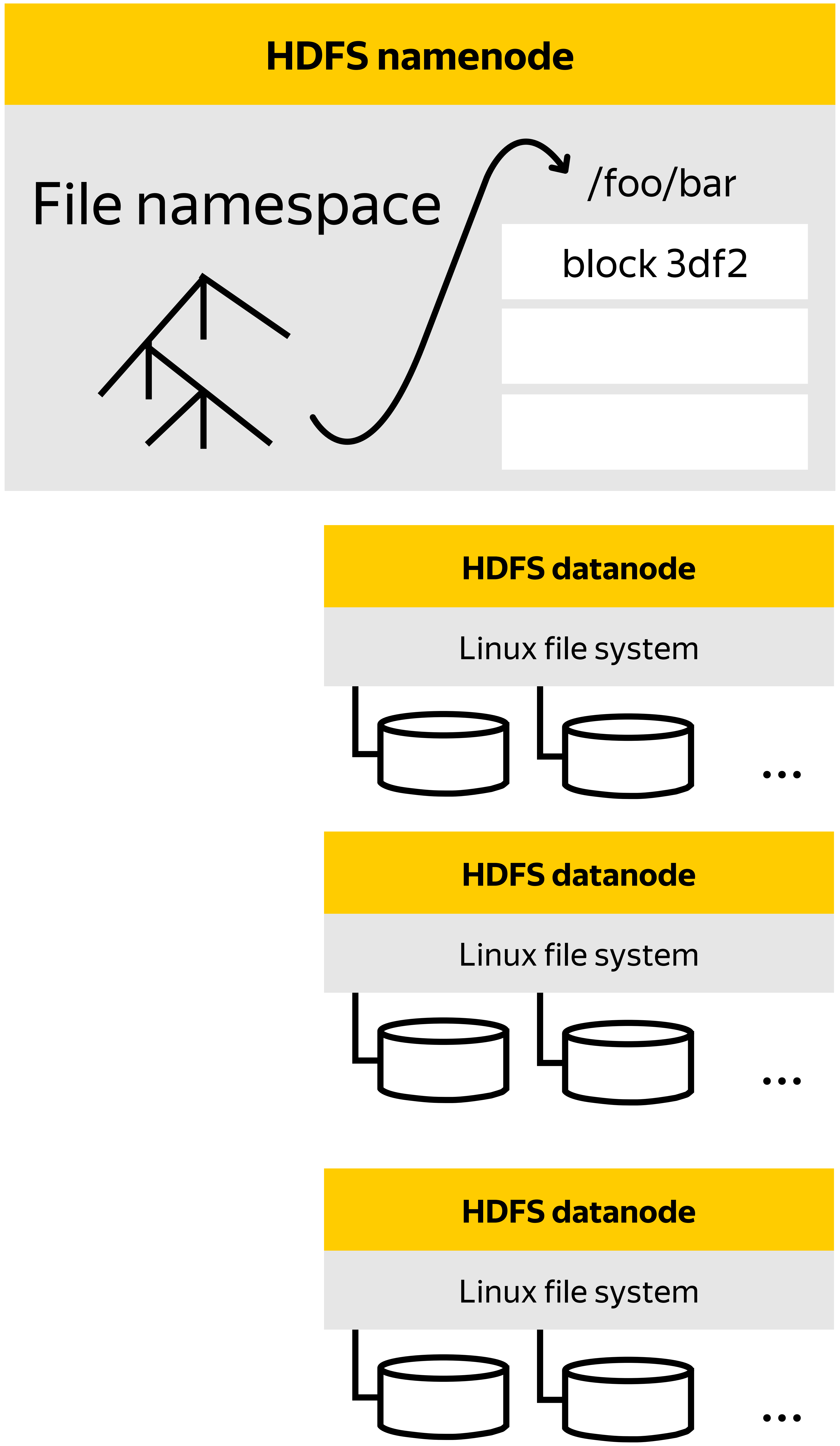
1 year ~ 10 PB

10 PB / 2 TB * 3 ~ 15 k

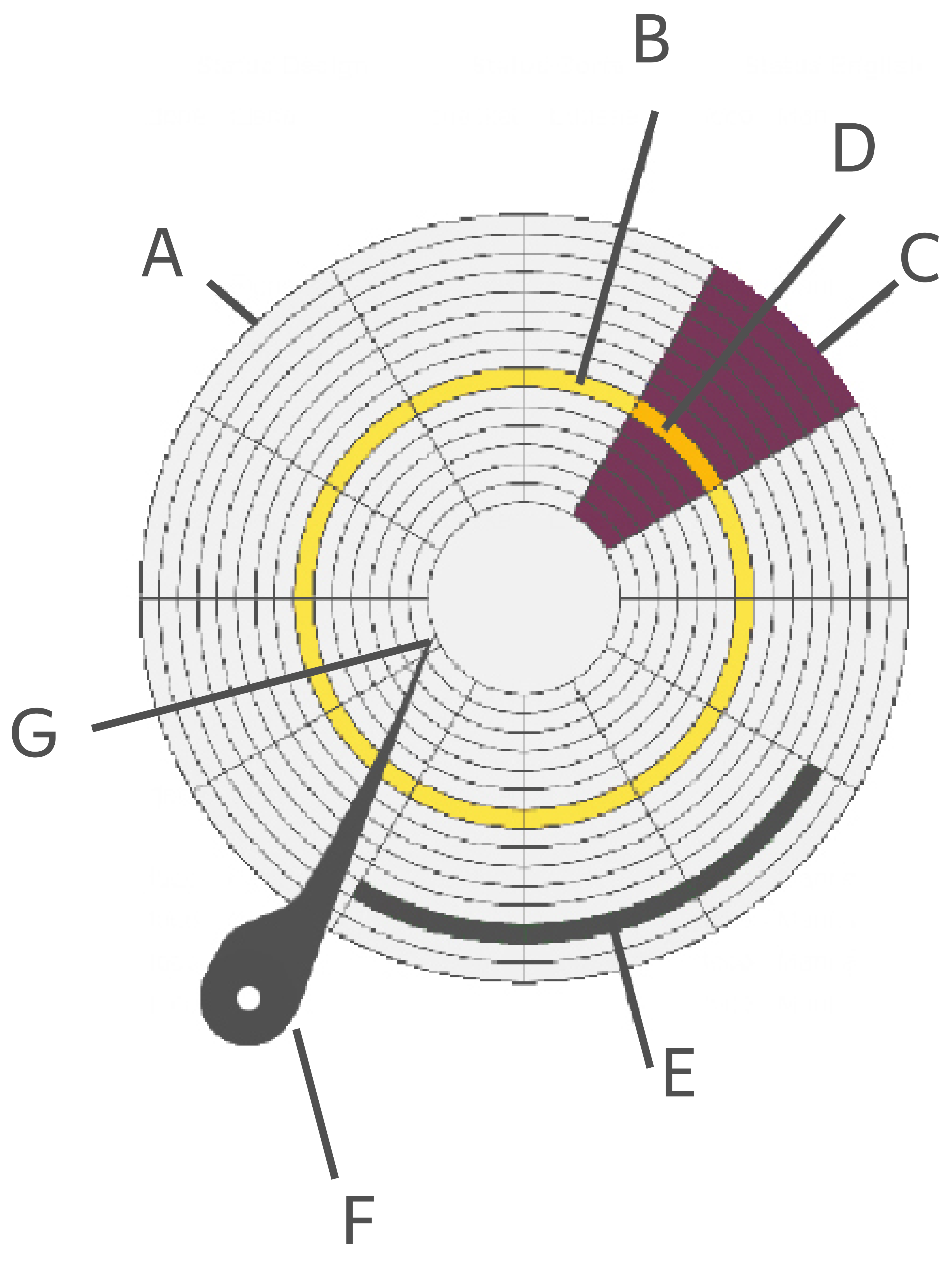
RAM

10 PB / **128 MB** * 3 * 150 B ~ 35 GB

**default block size**

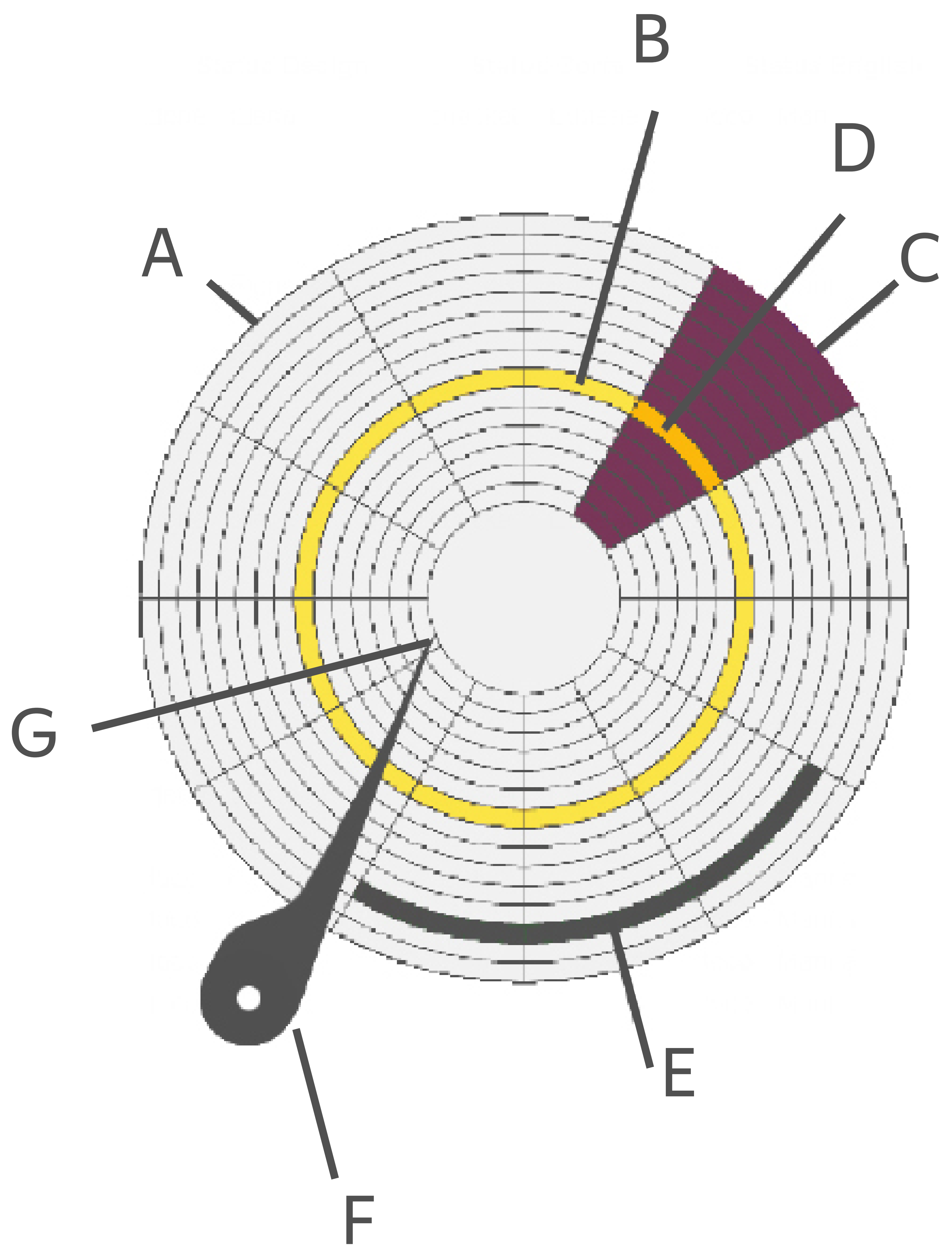


# Default Block Size



A — Platter
B — Track
C — Disk Sector
D — Track Sector
E — Cluster
F — Actuator Arm
G — Head

# Default Block Size

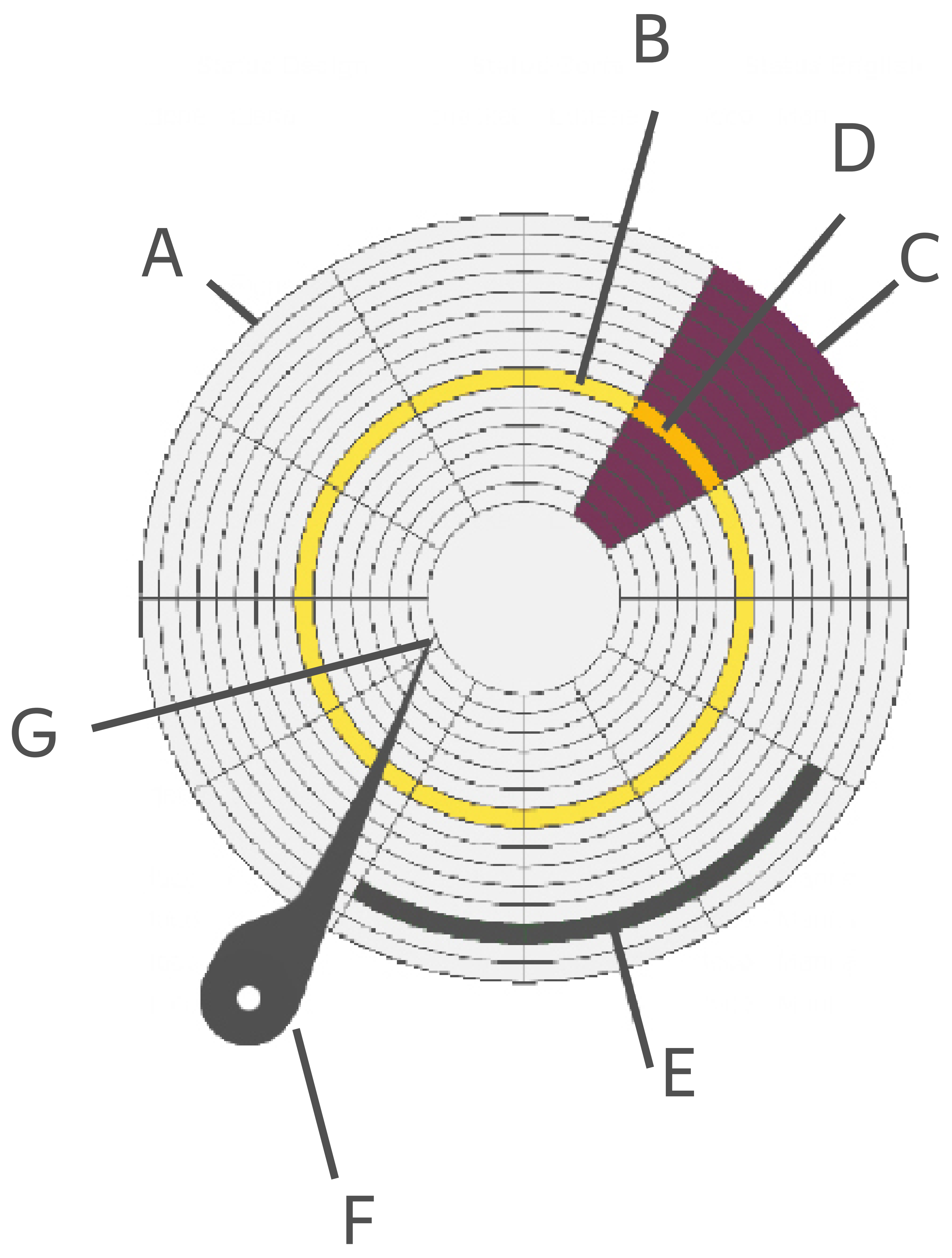


Samsung 940 PRO SSD:

* reading speed - 3.5 GB/sec

* 128 MB - 30-40 ms

# Default Block Size



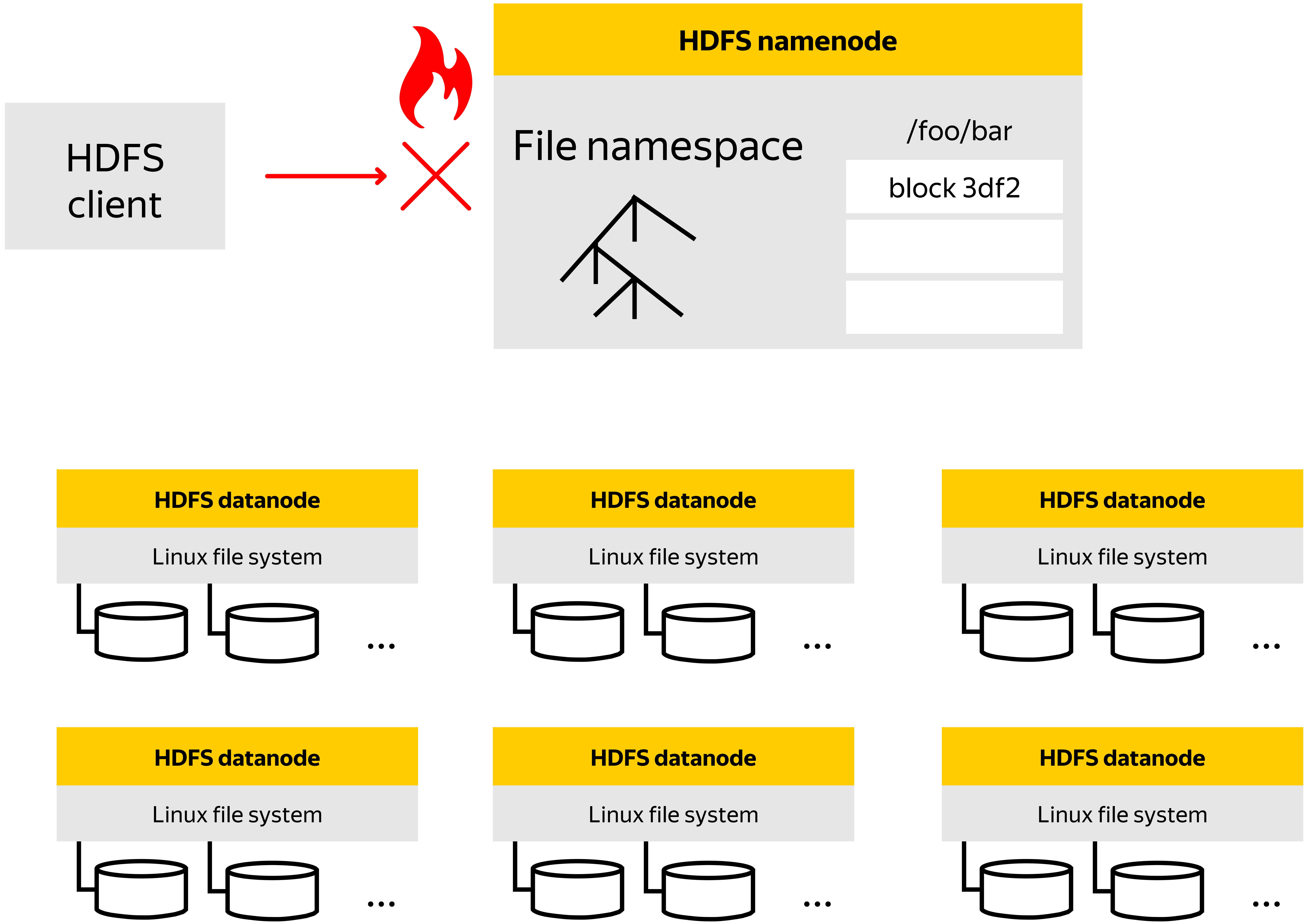
Samsung 940 PRO SSD:

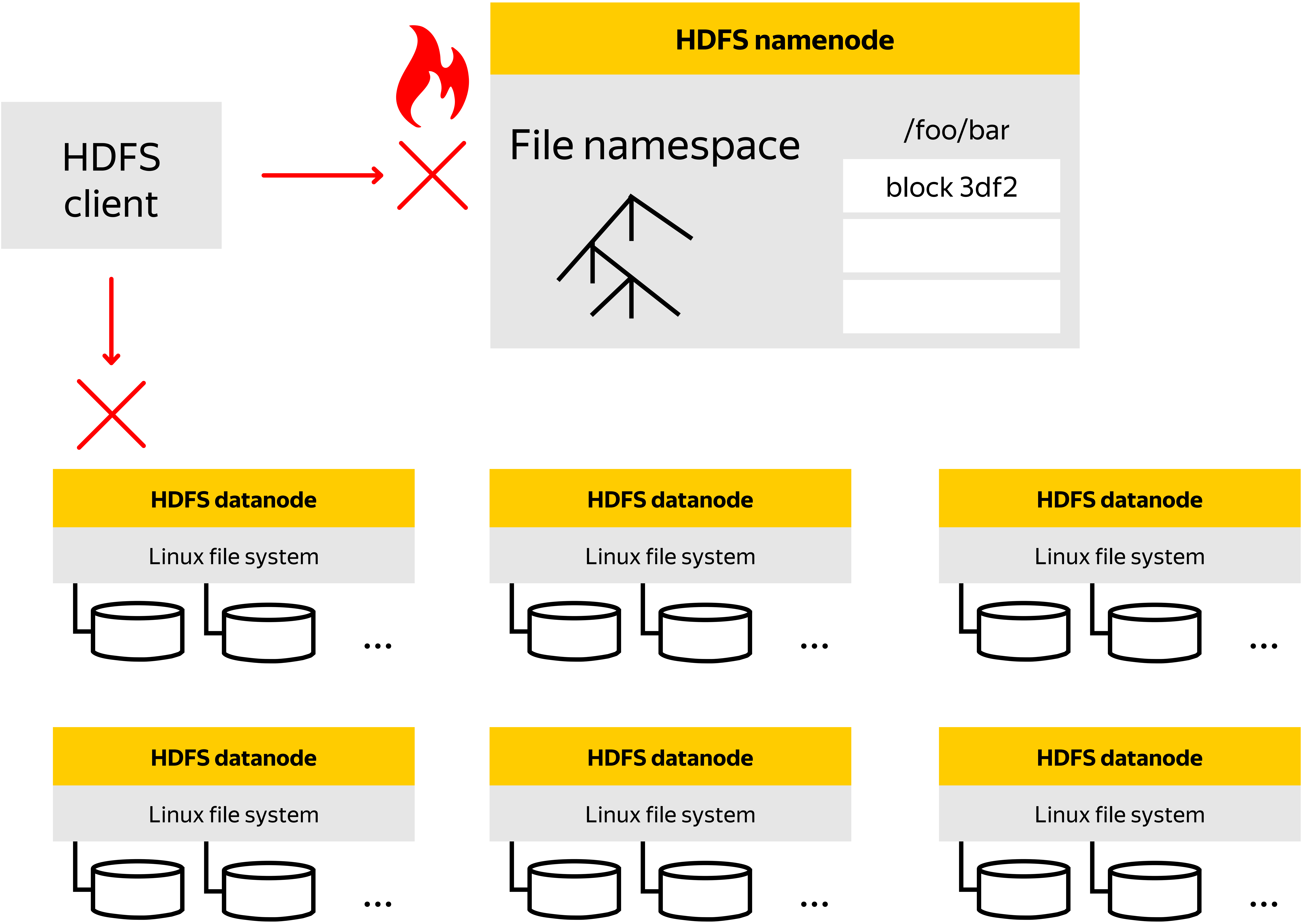
* reading speed - 3.5 GB/sec

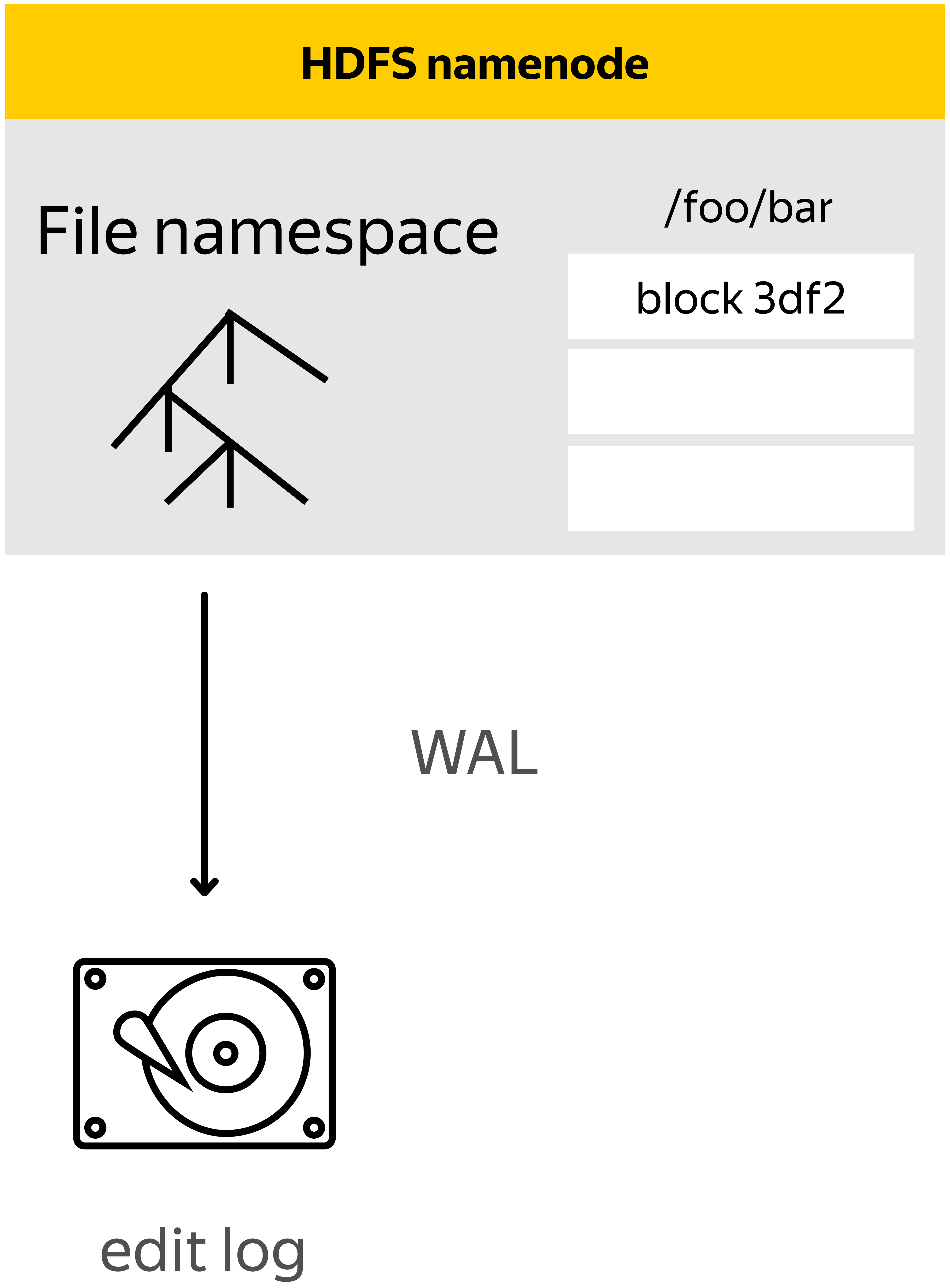
* 128 MB - 30-40 ms

seek time: 0.2-0.8 ms

1%


HDFS client
HDFS namenode
File namespace
/foo/bar
block 3df2
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...


HDFS namenode
HDFS client
File namespace
/foo/bar
block 3df2
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...
HDFS datanode
Linux file system
...



HDFS namenode

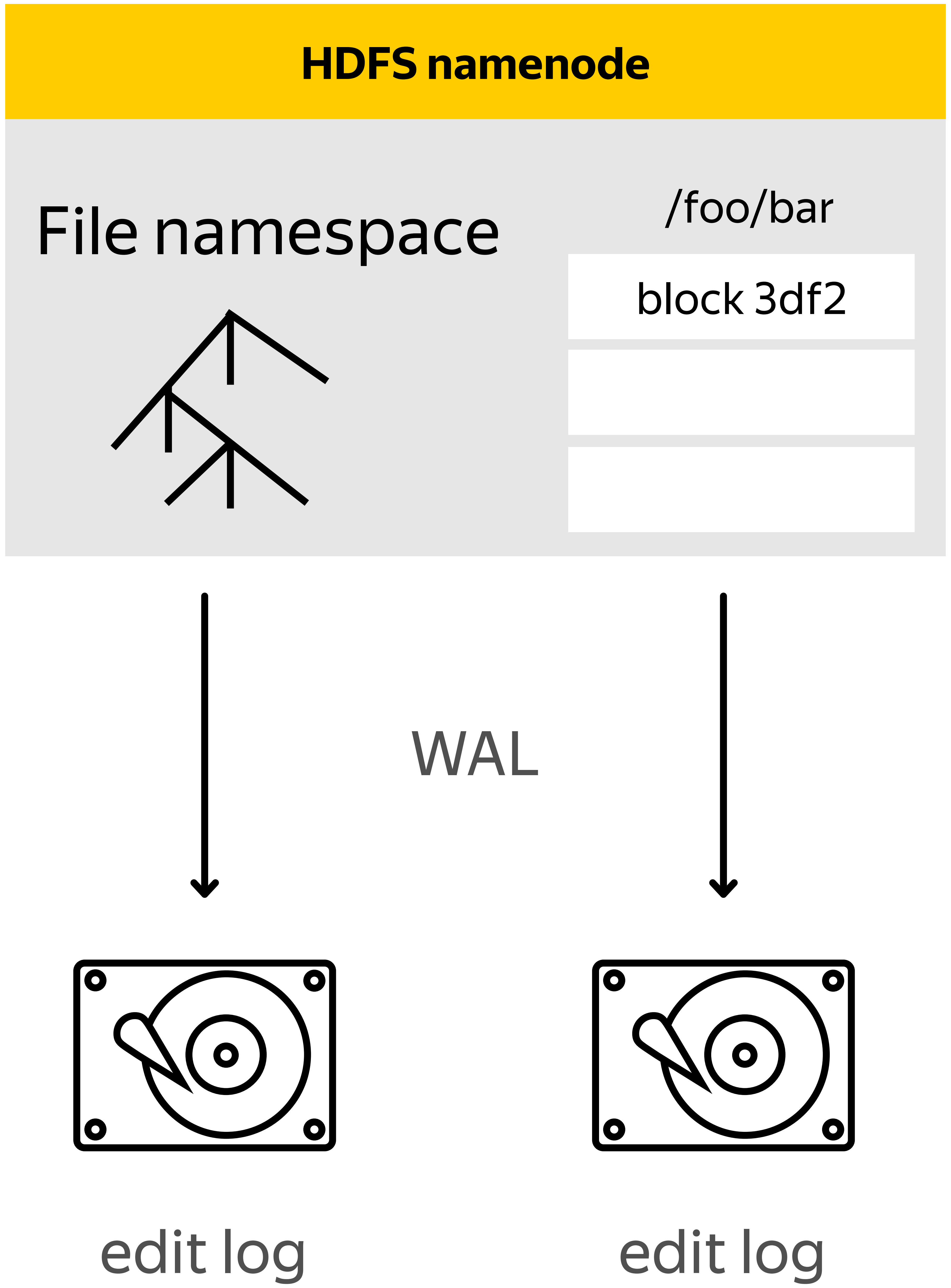
File namespace

/foo/bar

block 3df2

WAL

edit log



HDFS namenode

File namespace

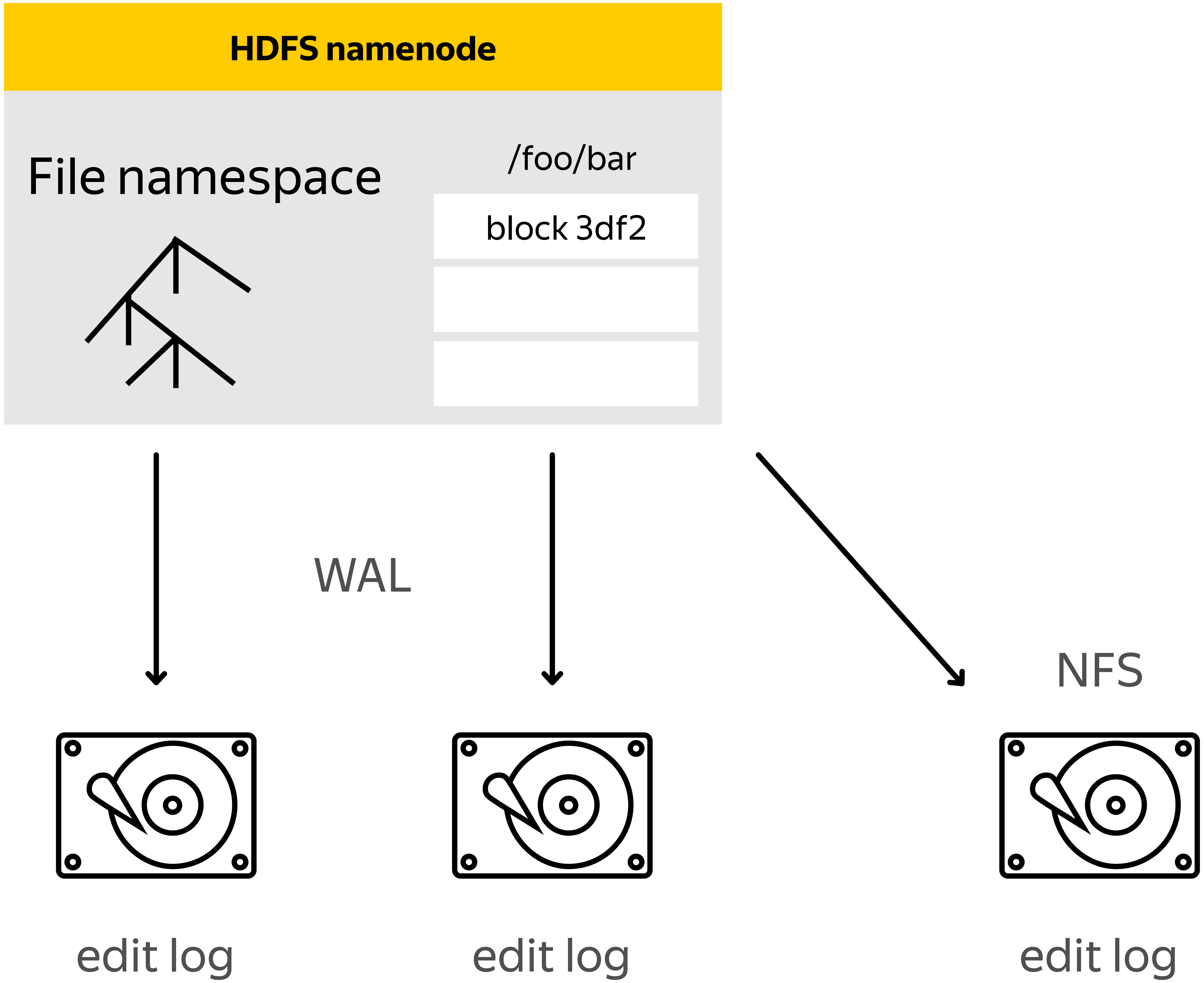
/foo/bar

block 3df2

WAL

edit log

edit log



HDFS namenode

File namespace

/foo/bar

block 3df2

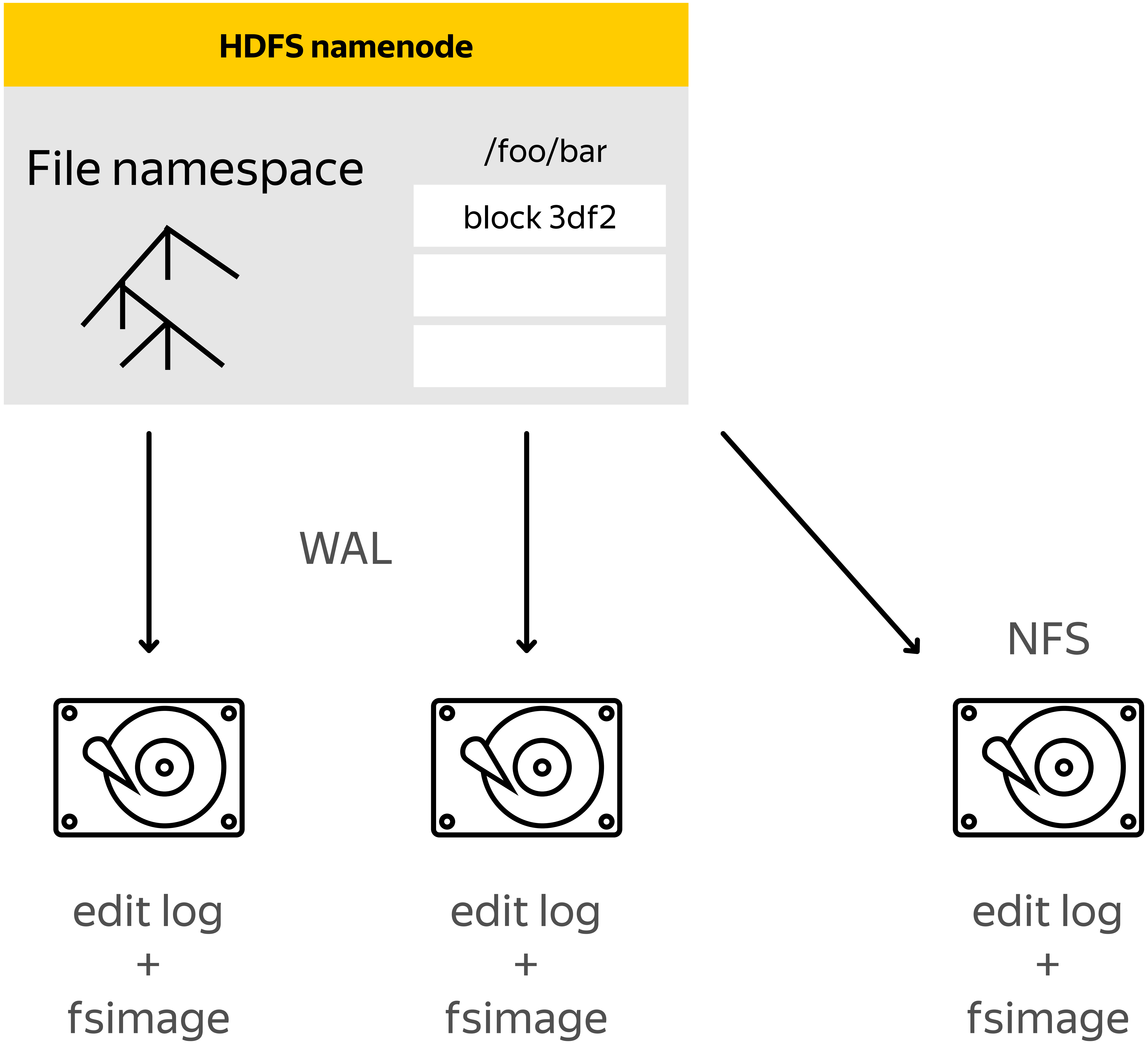
WAL

NFS

edit log

edit log

edit log



HDFS namenode

File namespace

/foo/bar

block 3df2

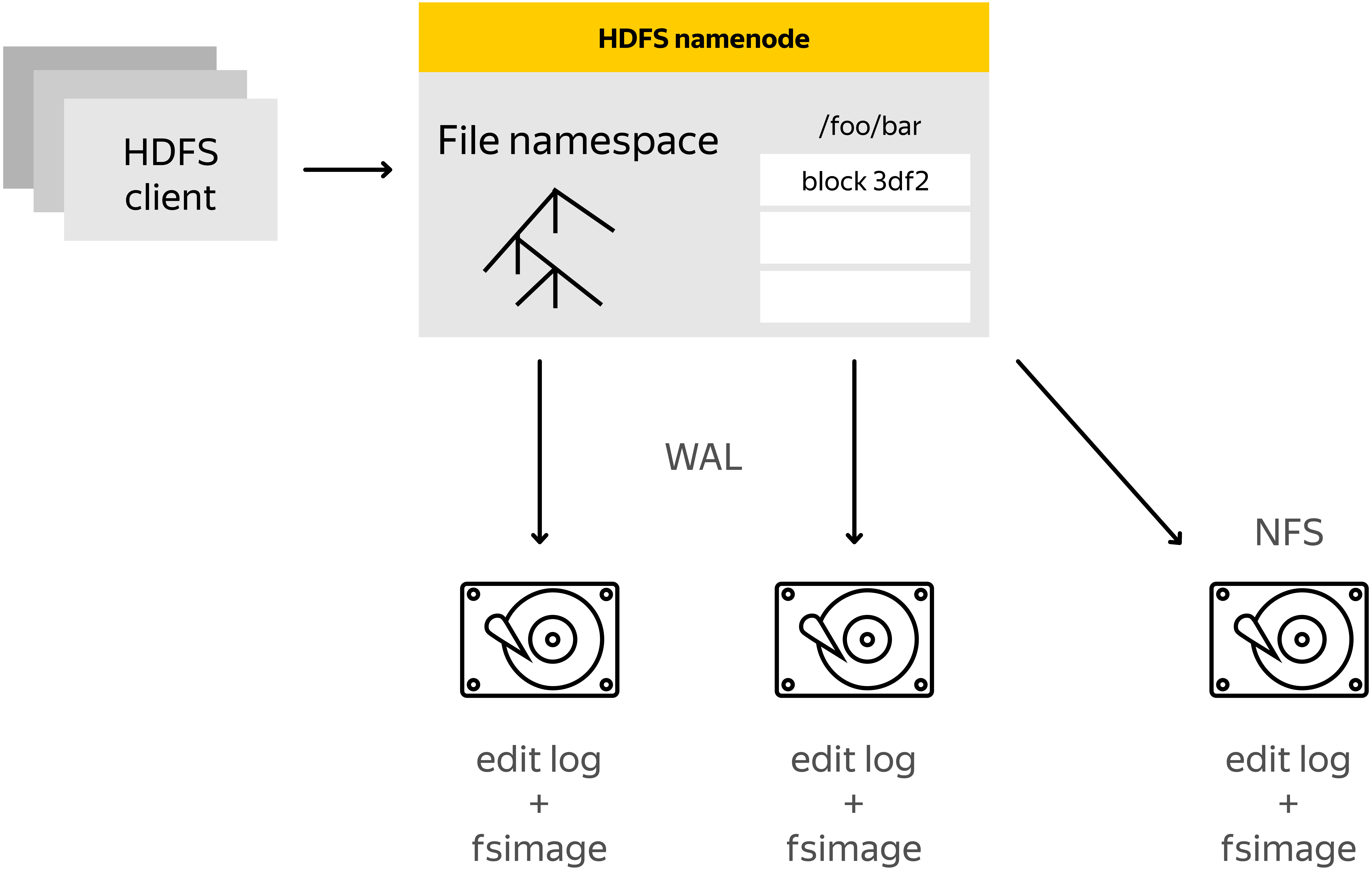
WAL

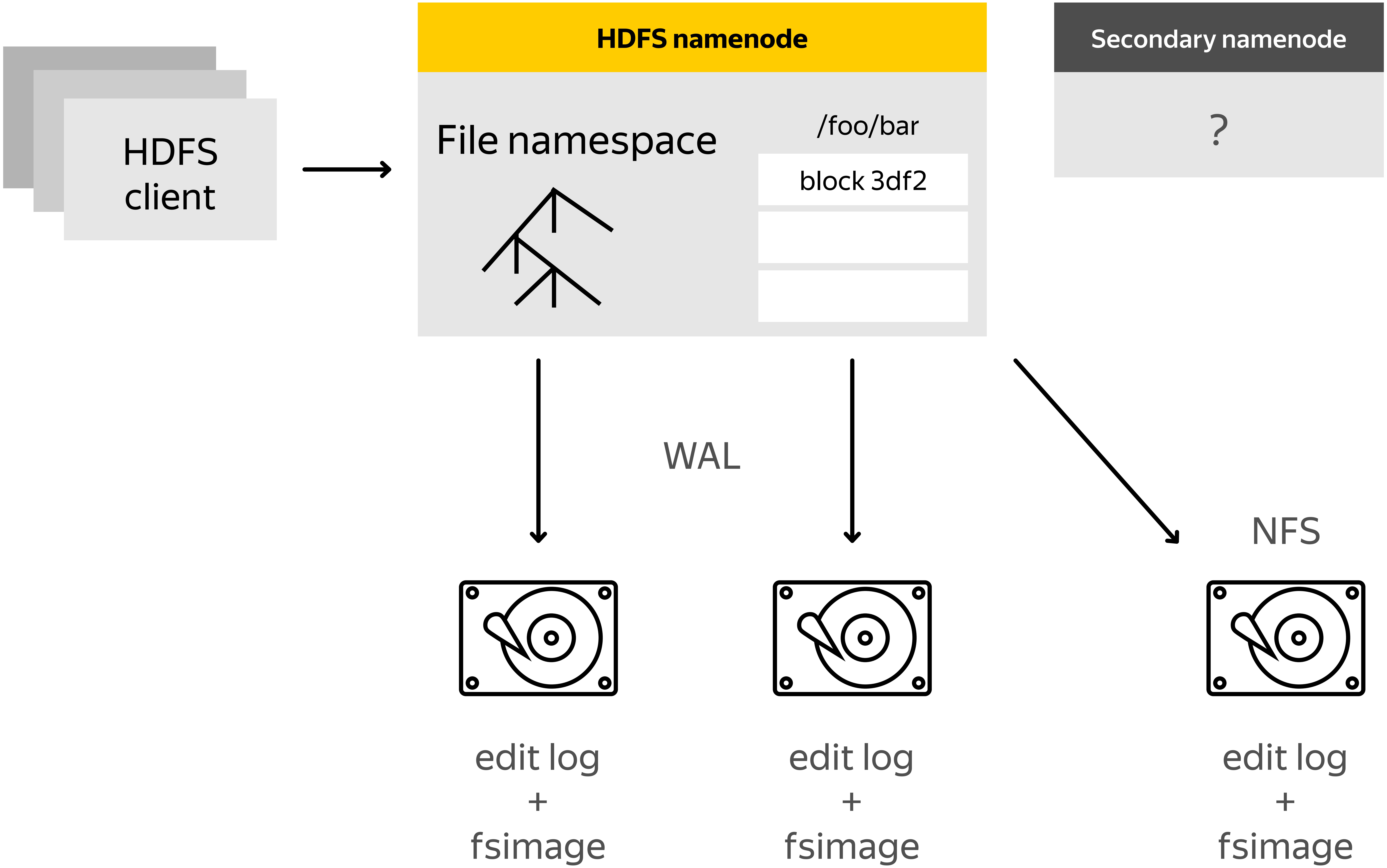
NFS

edit log
+
fsimage

edit log
+
fsimage

edit log
+
fsimage


HDFS client
HDFS namenode
File namespace
/foo/bar
block 3df2
WAL
NFS
edit log + fsimage
edit log + fsimage
edit log + fsimage



HDFS client

HDFS namenode

File namespace

/foo/bar

block 3df2

Secondary namenode

?

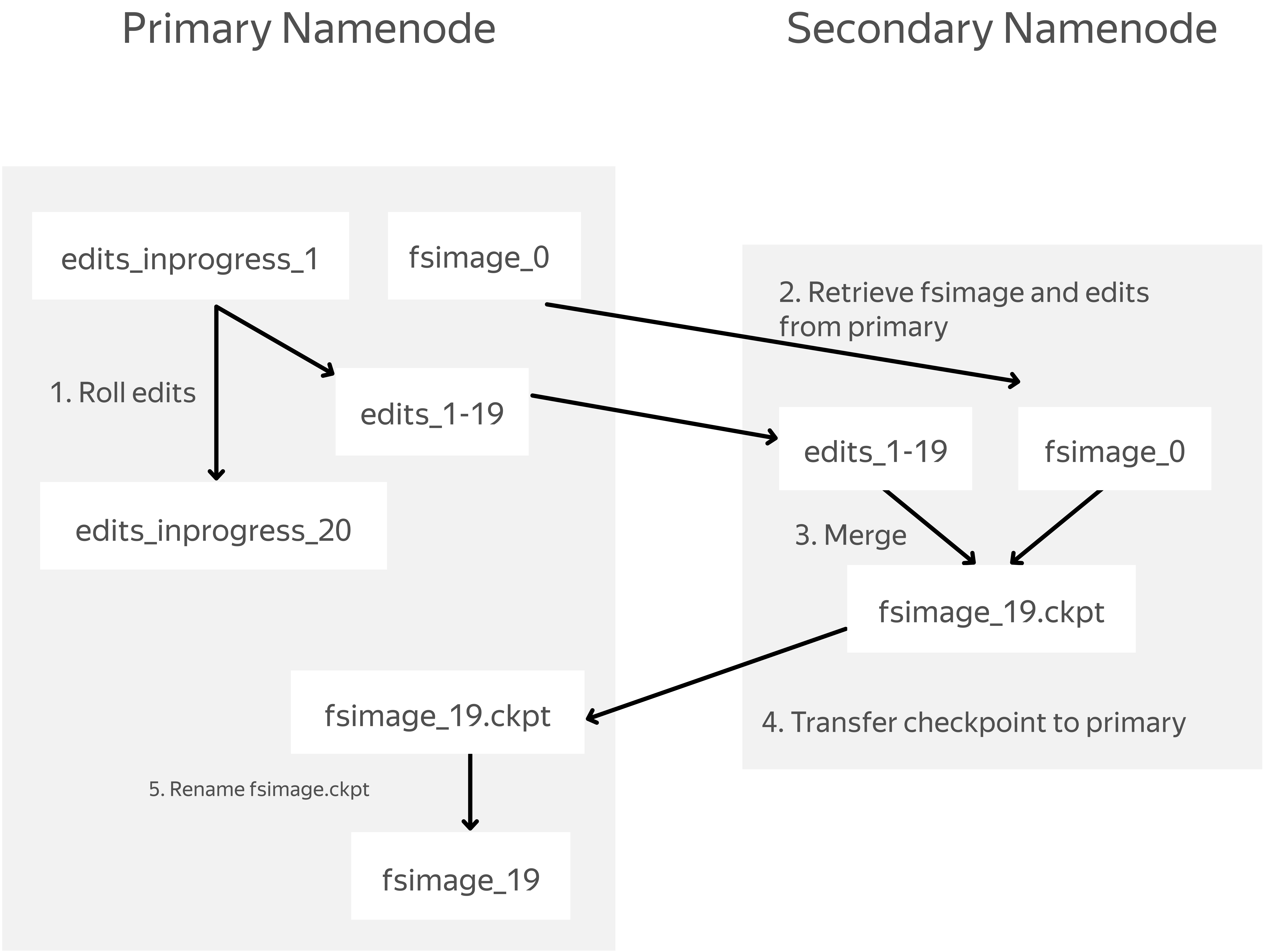
WAL

NFS

edit log
+
fsimage

edit log
+
fsimage

edit log
+
fsimage



Primary Namenode

Secondary Namenode

edits_inprogress_1

fsimage_0

1. Roll edits

edits_1-19

edits_inprogress_20

2. Retrieve fsimage and edits from primary

edits_1-19

fsimage_0

3. Merge

fsimage_19.ckpt

4. Transfer checkpoint to primary

fsimage_19.ckpt

5. Rename fsimage.ckpt

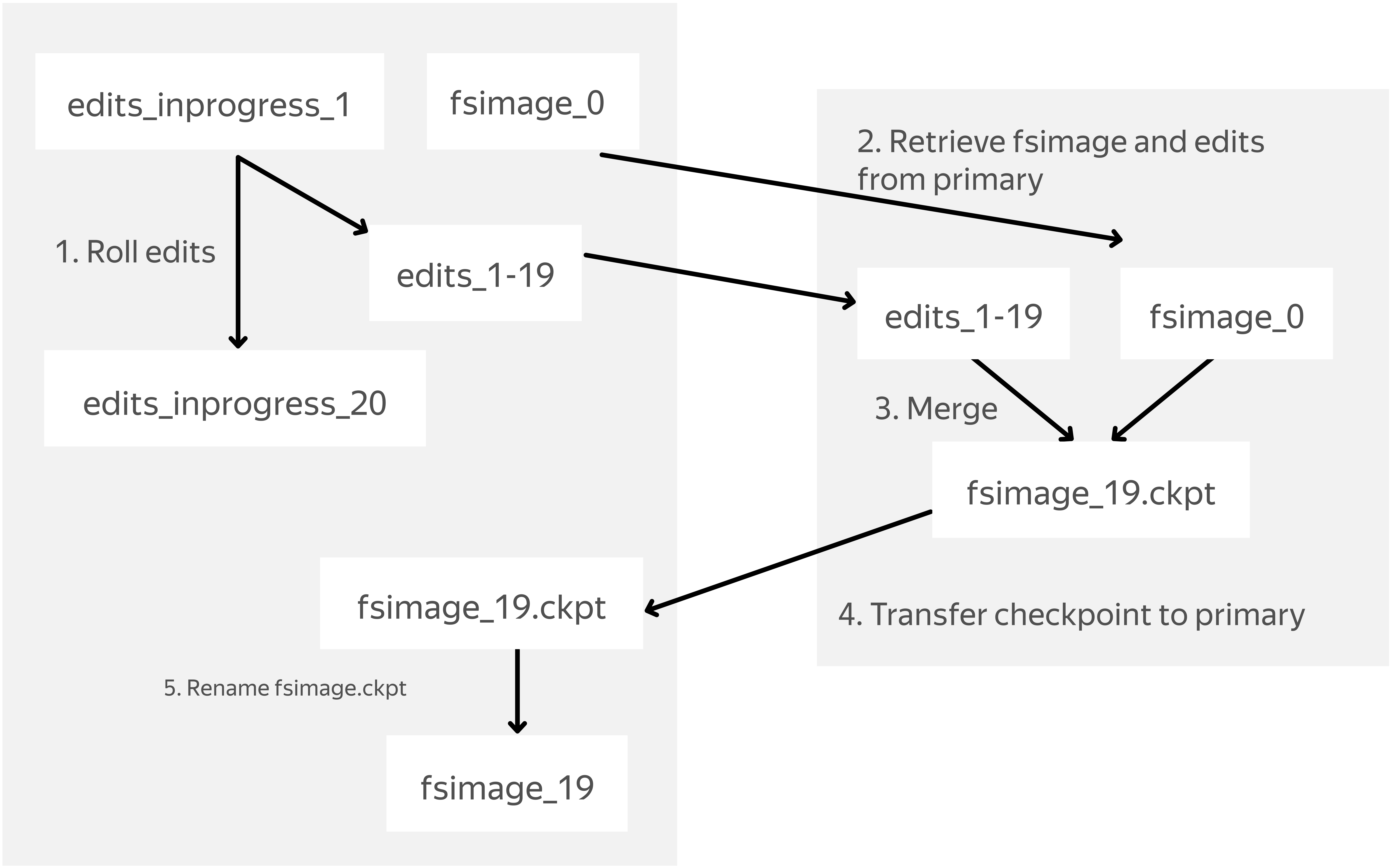
fsimage_19

Primary Namenode

Secondary Namenode

= Checkpoint Namenode

≠ Backup Node



edits_inprogress_1

fsimage_0

1. Roll edits

edits_1-19

edits_inprogress_20

2. Retrieve fsimage and edits from primary

edits_1-19

fsimage_0

3. Merge

fsimage_19.ckpt

4. Transfer checkpoint to primary

fsimage_19.ckpt

5. Rename fsimage.ckpt

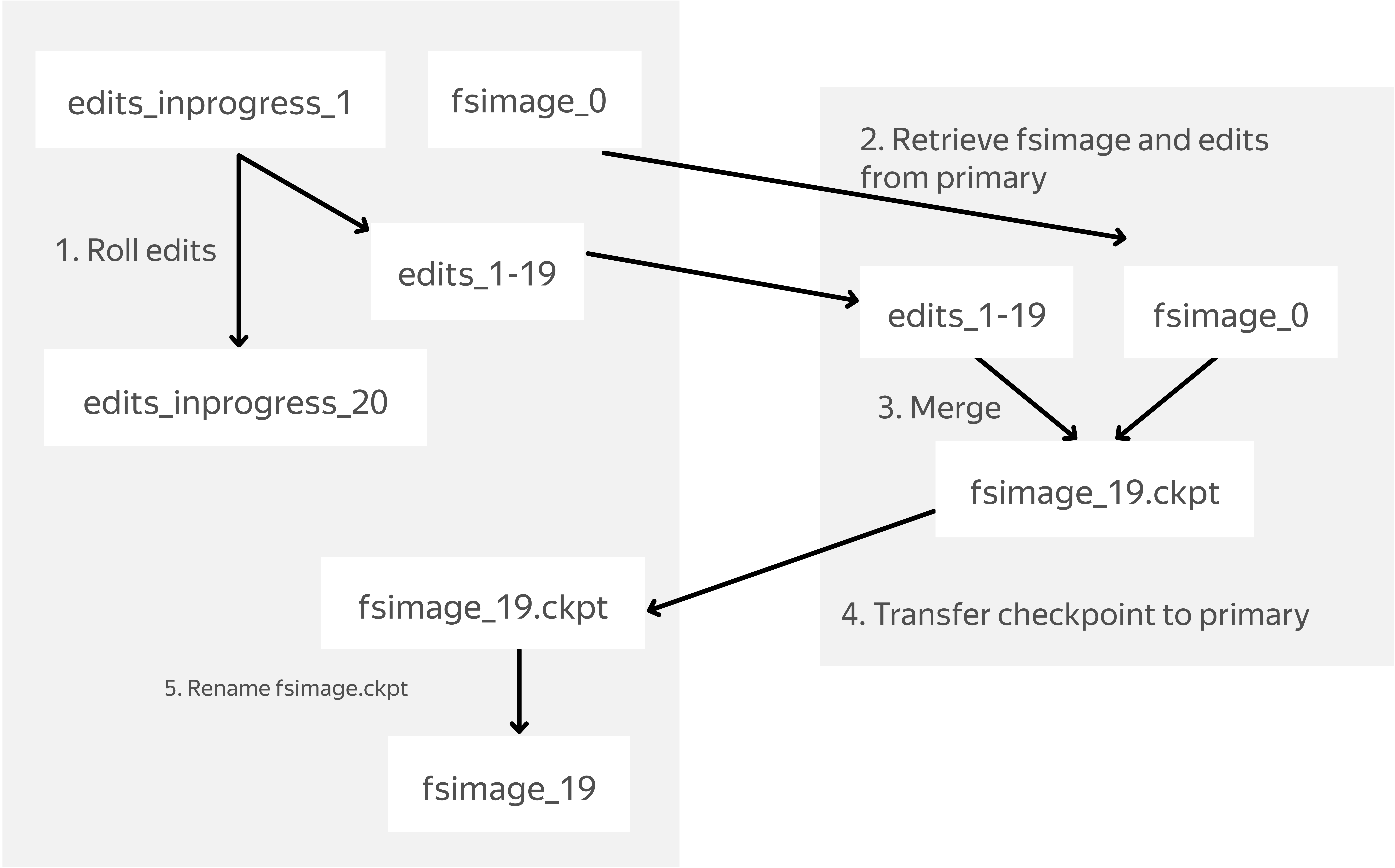
fsimage_19

## Primary Namenode

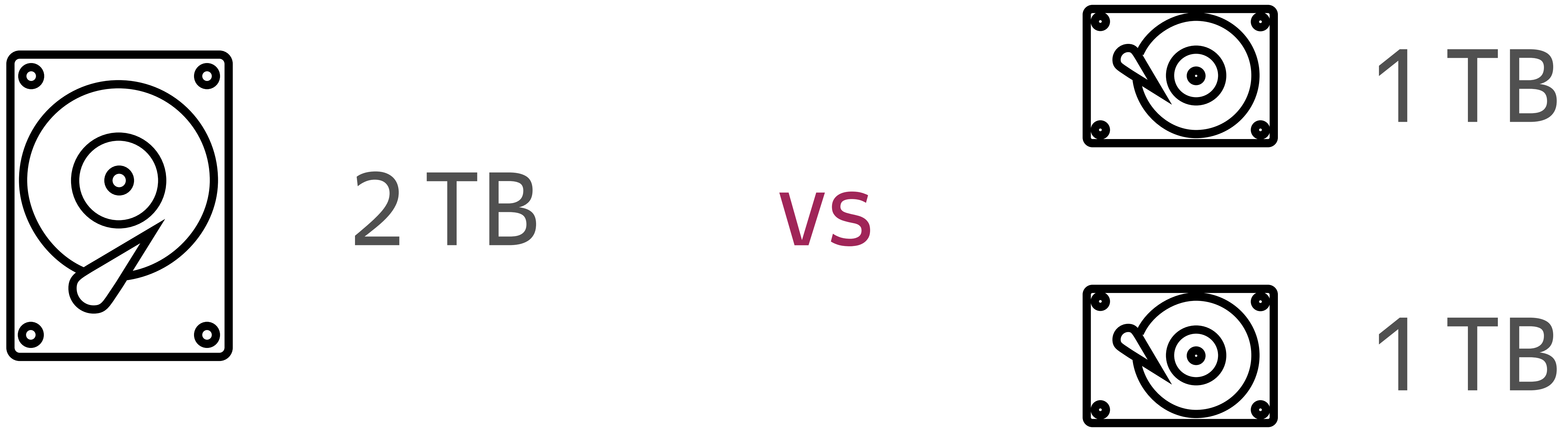
## Secondary Namenode

= Checkpoint Namenode

≠ Backup Node



https://issues.apache.org/jira/browse/HADOOP-4539



2 TB

vs

1 TB

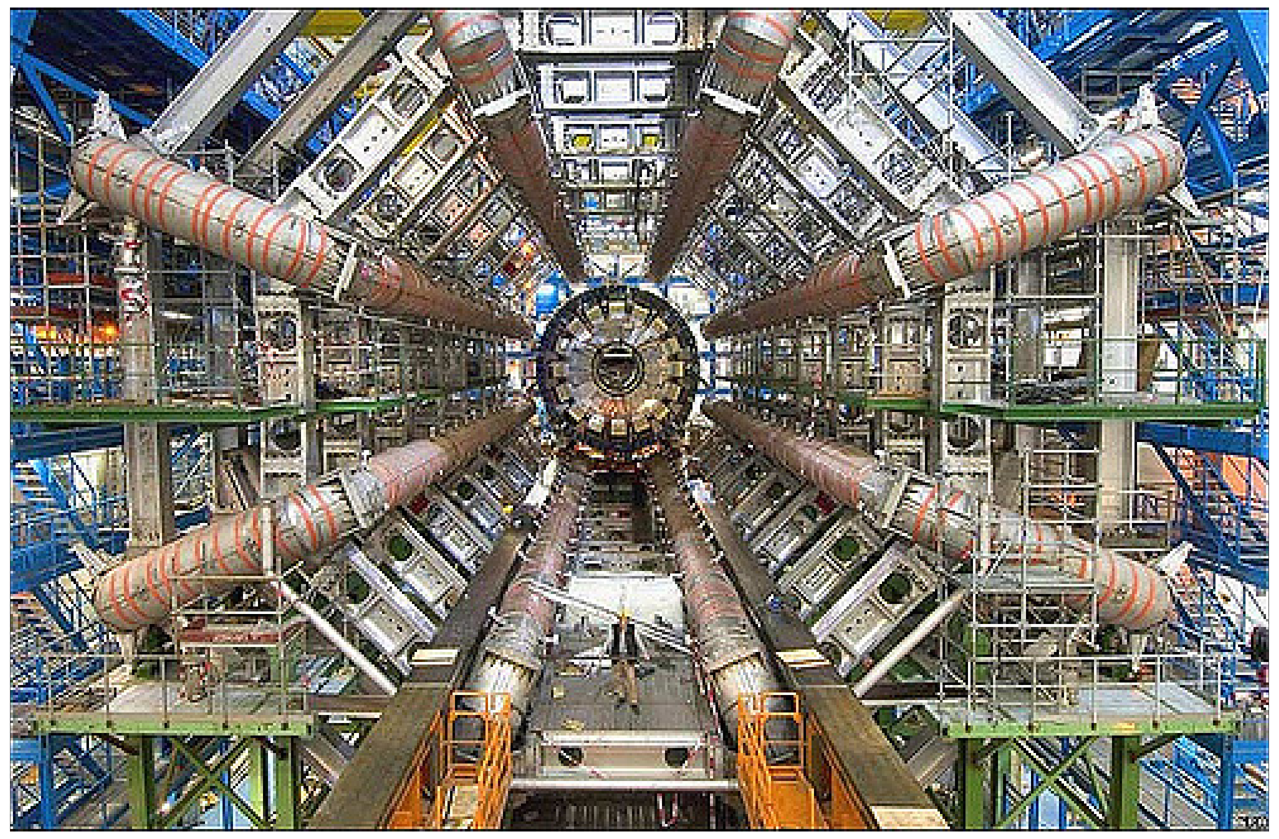
1 TB

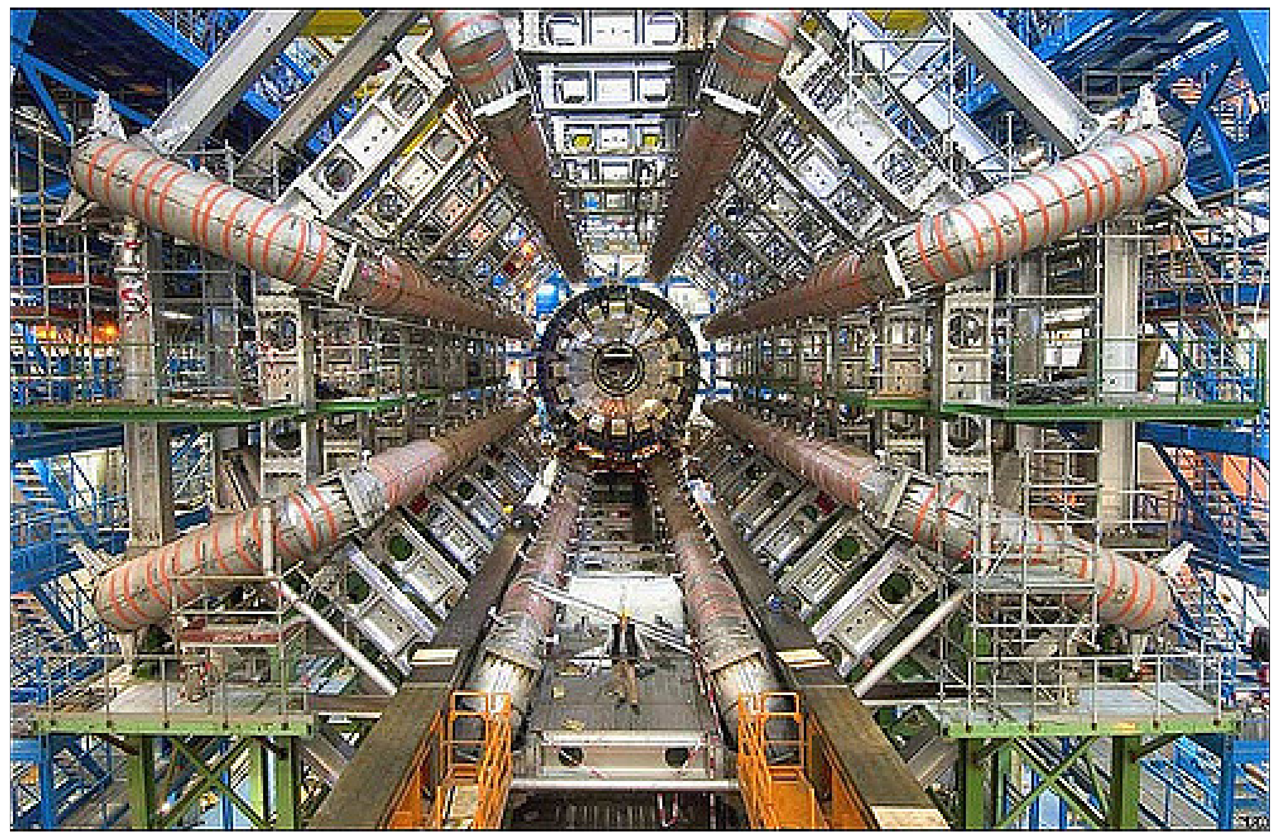
2 TB

vs

1 TB

1 TB





1 year ~ 10 PB

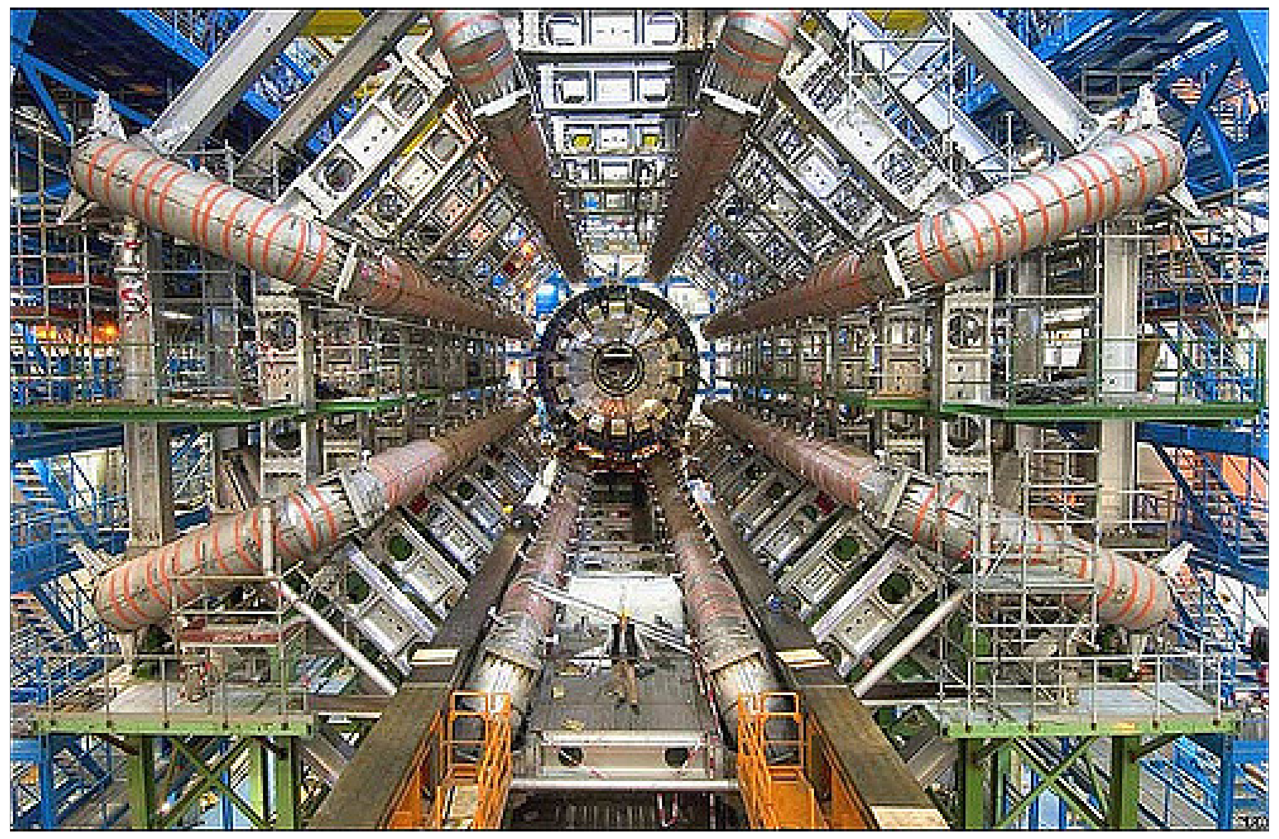
1 year ~ 5 PB + 5 PB

2 TB

vs

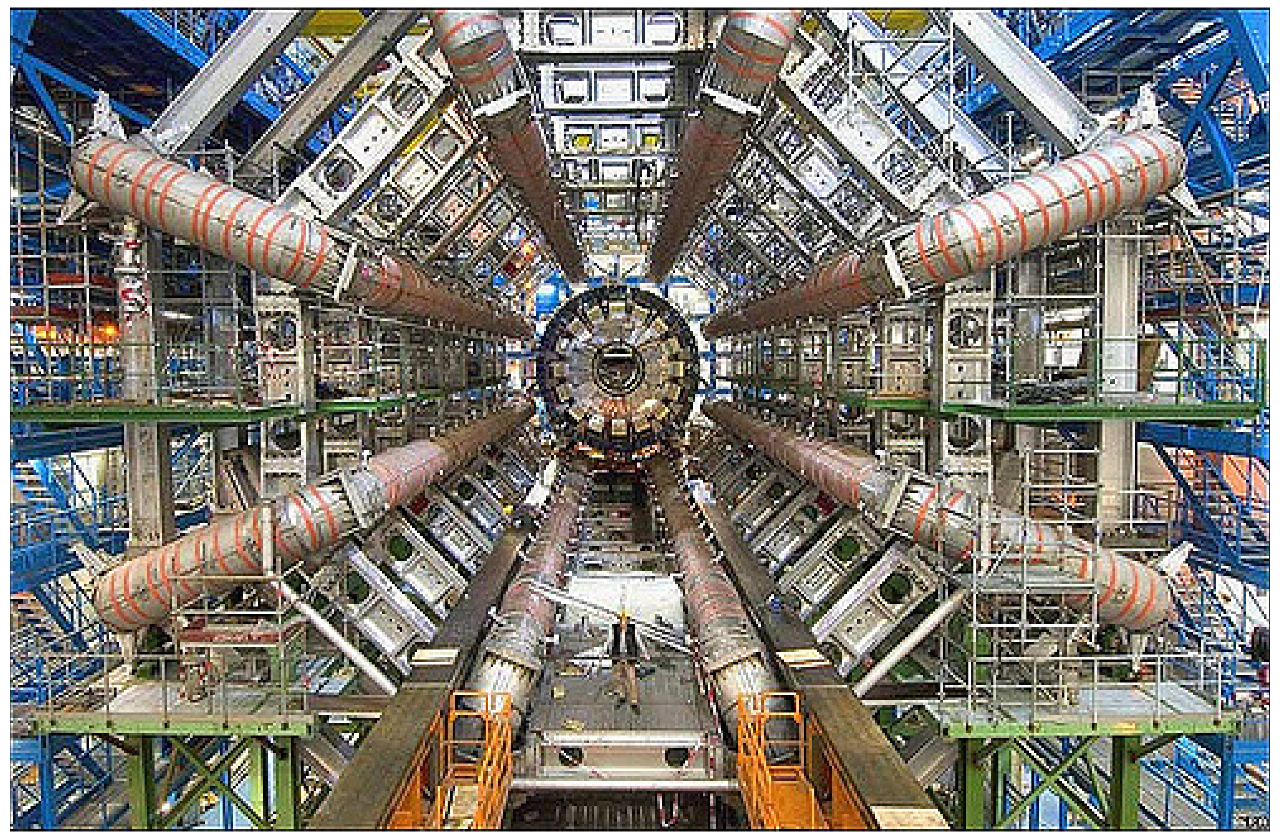
1 TB

1 TB



1 year ~ 10 PB

35 days



1 year ~ 5 PB + 5 PB

17.5 days

# Summary

# Summary

- you can **explain and reason about** HDFS Namenode architecture (RAM; fsimage + edit log; block size)

# Summary

- you can **explain and reason about** HDFS Namenode architecture (RAM; fsimage + edit log; block size)
- you can **estimate** required resources for a Hadoop cluster

# Summary

- you can **explain and reason about** HDFS Namenode architecture (RAM; fsimage + edit log; block size)
- you can **estimate** required resources for a Hadoop cluster
- you can **explain** what small files problem is and where a bottleneck is

# Summary

- you can **explain and reason about** HDFS Namenode architecture (RAM; fsimage + edit log; block size)
- you can **estimate** required resources for a Hadoop cluster
- you can **explain** what small files problem is and where a bottleneck is
- you can **list differences** between different types of Namenodes (Secondary / Checkpoint / Backup)

**Big**DATAteam