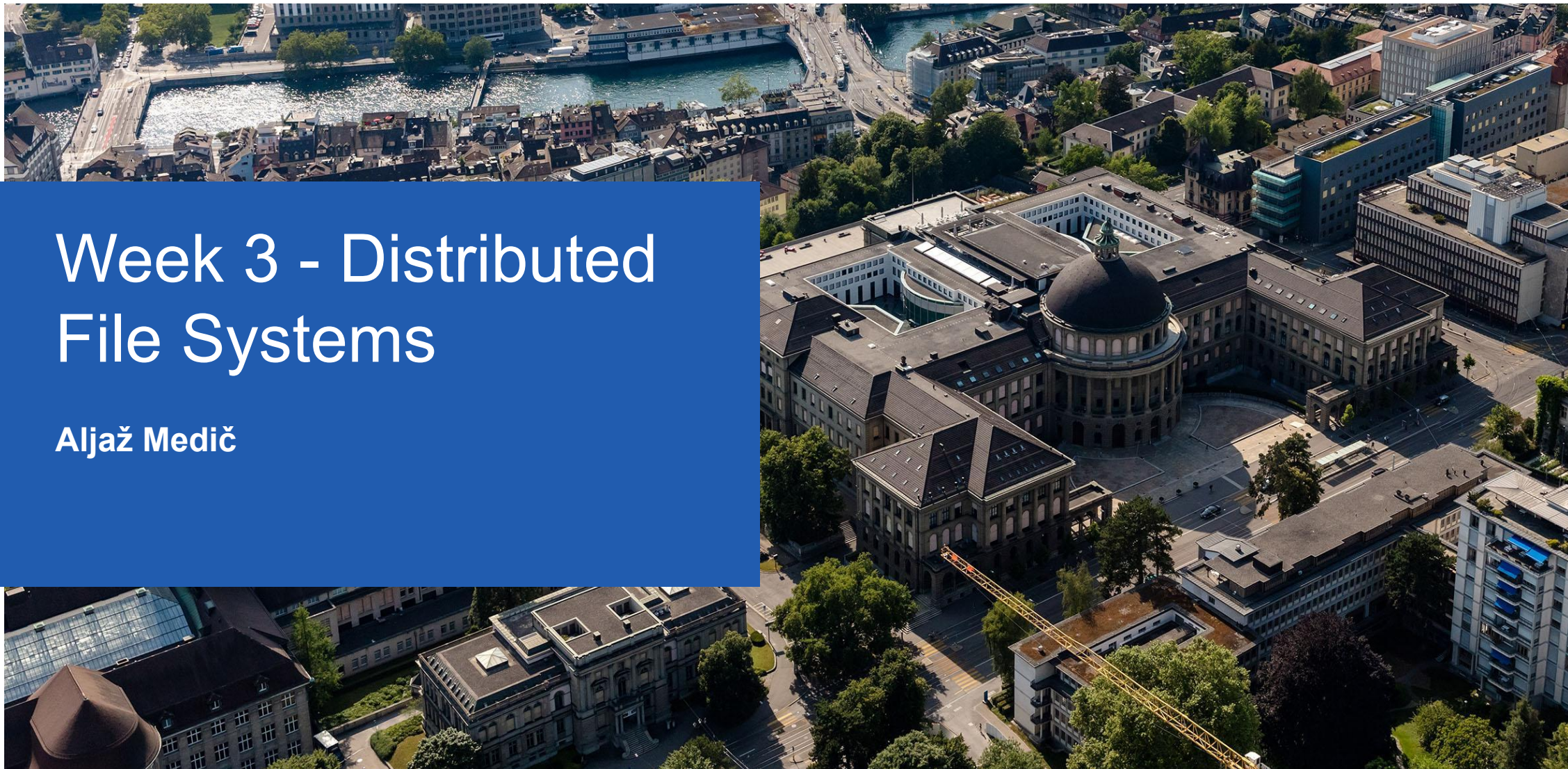


Week 3 - Distributed File Systems

Aljaž Medič



Plan for today

1. Questions about Grade Quiz or Exercise Sheet
2. Quiz
3. Replication policy

QUIZ

HDFS natively organizes files in a hierarchy of directories rather than using a flat key-value model.

True.

A typical HDFS block is around 64 MB to 128 MB, which is much larger than a typical local-filesystem block of 4 KB.

True.

Every block belonging to an HDFS file has exactly the configured block size.

False. (Last block of a file may have less data, unless file size is an exact multiple of block size)

QUIZ

HDFS uses a decentralized peer-to-peer server architecture in which all storage nodes have equivalent roles.

False. (HDFS uses a centralized architecture with a special NameNode and multiple interchangeable DataNodes.)

The NameNode stores the actual contents of HDFS blocks together with their metadata.

False. (The NameNode -> metadata. DataNodes -> block contents.)

Give all three main kinds of information maintained by the NameNode in memory.

1. File namespace (+ ACL) 2. Map: File -> list of block IDs 3. Map block ID -> DataNode list

QUIZ

HDFS block identifiers stored in NameNode metadata are 64-bit identifiers.

True.

A DataNode normally stores each HDFS block as a physical file on its local filesystem.

True.

If the final HDFS block contains only 20 MB of data while the configured block size is 128 MB, the DataNode still allocates a 128 MB local file for it, same as our local filesystems do with 4KB blocks.

False. (They allocate the exact amount, without waste. Still subject to the underlying local filesystem block size)

QUIZ

If the NameNode urgently needs a DataNode to create another replica, it initiates a RPC to DataNode.

False. (It waits for DataNode to send a message and returns the instruction in its response.)

When a block is reported as corrupted, it can temporarily be marked under-replicated while the NameNode arranges asynchronous replication elsewhere.

True.

When writing a replicated block, the client sends the complete block separately to every selected DataNode.

False. (The client sends data to the first DataNode, and the DataNodes forward the data through the replication pipeline.)

QUIZ

An ACK from the replication pipeline signals that the block replica has been successfully persisted to disk.

False. (Pipeline acknowledgements confirm that packets were **received**. Persistence is reported separately to the NameNode with BlockReceived.)

A 128 MB HDFS block is normally streamed over the network as smaller packets rather than being transmitted as one 128 MB network packet.

True.

When reading a file, the client first obtains block and replica-location metadata from the NameNode and then transfers the block contents directly from DataNodes.

True.

QUIZ

To read a replicated HDFS file correctly, the client must download each block from every DataNode holding a replica of that block.

False. (Reads from one DataNode, starting with the closest and falling back to another if necessary.)

The NameNode can return the candidate DataNodes for a block ordered by increasing network distance from the client.

True.

Streaming an HDFS file makes it possible for a client to process a file larger than its working memory.

True.

QUIZ

While a new HDFS file is being written, it is locked so that another client cannot simultaneously write to that same file.

True.

Before releasing the lock on newly written file, the NameNode waits until every block has reached the configured replication factor.

False. (NameNode waits synchronously for minimal replication, then asynchronously reaches the configured replication)

The default HDFS replication factor is three, and it can be changed by the user for individual files.

True.

QUIZ

For every HDFS block, one of its replicas is designated as the main or primary replica.

False. (In HDFS all blocks are logically equal, no primaries.)

For the default replication factor of three, give the replication policy of the three replicas when the writing client runs on a DataNode.

1. Client's own machine 2. DataNode in a different rack 3. Different DataNode in the same rack as #2

In a basic HDFS single-NameNode deployment, the NameNode is a critical single point of failure because the raw blocks alone are insufficient to determine how those blocks should be reassembled into files.

True.

QUIZ

After a NameNode crash, we can obtain the exact latest filesystem state by loading the latest snapshot.

False. (We also have to replay the edit log.)

During a checkpoint, what happens to the snapshot and edit log?

We replay the edit log on top of the snapshot, creating new snapshot. We truncate the edit log.

After a NameNode restarts, where does it recover each part of its metadata from?

From Snapshot/Edit-Log: file namespace+ACL, File-to-blockID. Block Reports: blockID-to-DataNode

QUIZ

A Standby NameNode can reduce failover time because it maintains the relevant NameNode state and can take over after failure.

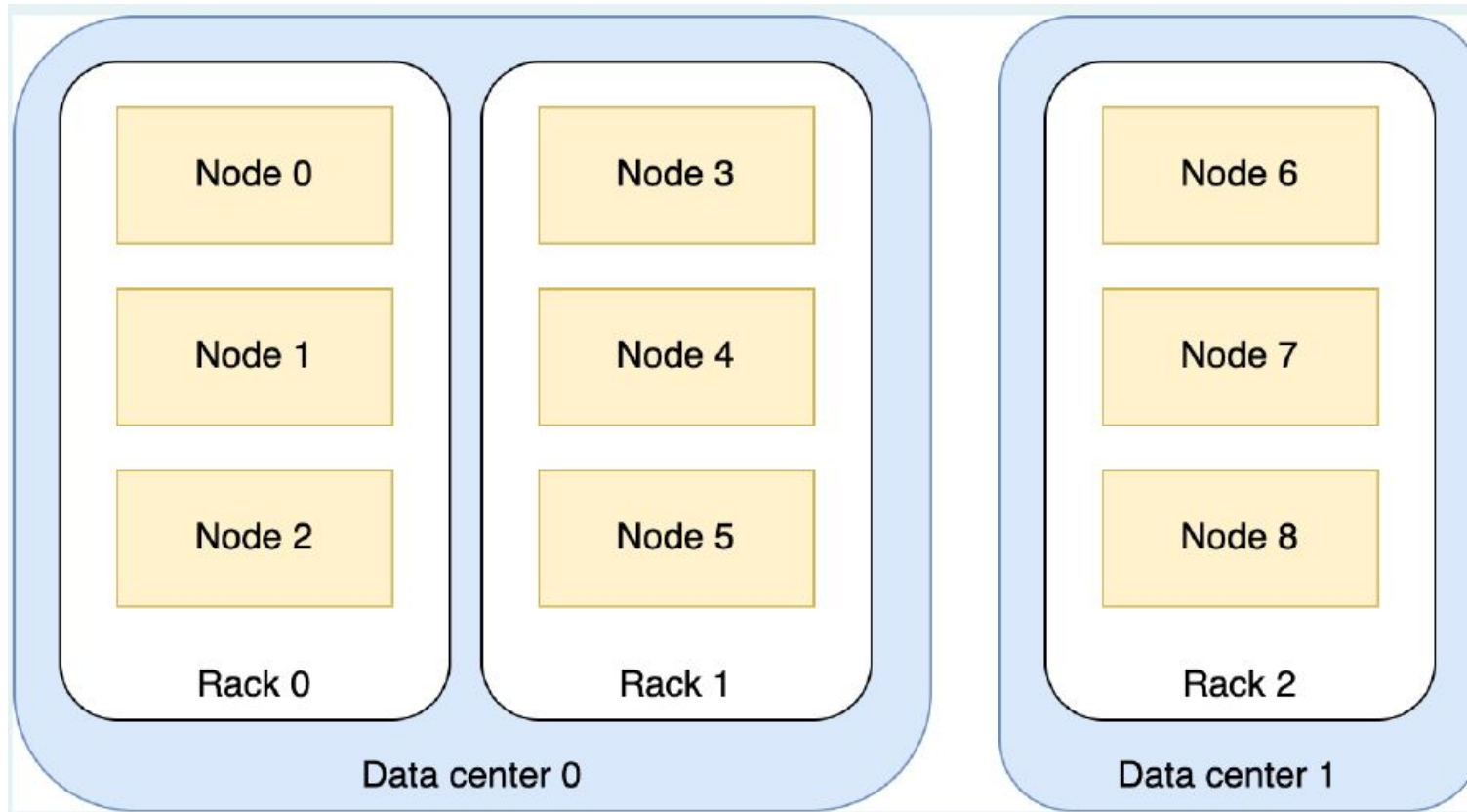
True.

Federated NameNodes allow several NameNodes to operate simultaneously, with different NameNodes managing non-overlapping parts of the file hierarchy.

True.

Replication Policy (HS23, Q19)

HDFS estimates distance between different Nodes with integers. Write the correct distances between the entities below.



Options: 0,1,2,3,4,5,6

- Node 0 and Node 3:
- two processes on Node 7:
- Node 4 and Node 7:
- Node 3 and Node 4:

ETH zürich

See you next week!

Aljaž Medič
amedic@ethz.ch



Slides



Suggestions