SINE NOMINE ASSOCIATES

# Overcoming Ubik Limitations

Marcio Barbosa

2019 OpenAFS Workshop

# AGENDA

SINE NOMINE ASSOCIATES

- Election
- Recovery
- Limitations
- Reads-during-sync
- Transactions
- Read-transaction
- Write-transaction
- Limitations
- Reads-during-commit
- Other fixes

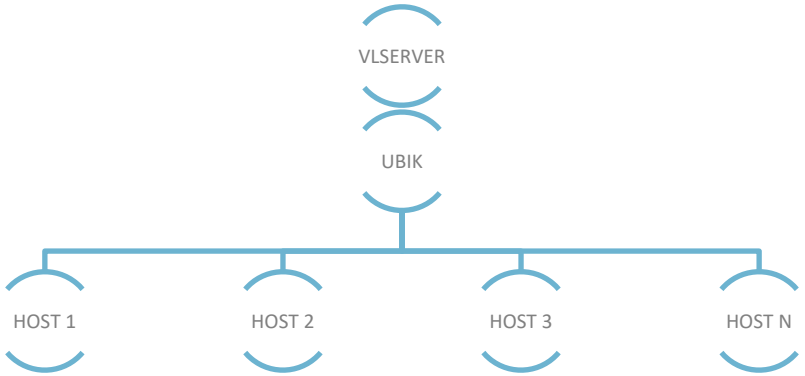
SINE NOMINE
ASSOCIATES

# ELECTION

# ELECTION

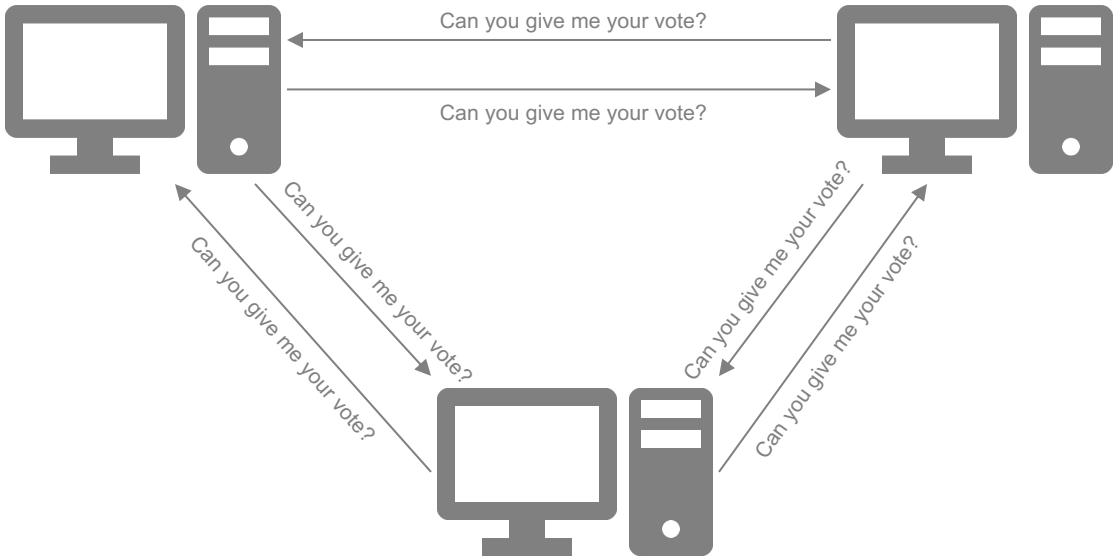
SINE NOMINE ASSOCIATES

- A coordinator is elected by sending out beacon packets;
- A beacon implicitly asks the recipient to vote for the sender;
- Site with more votes is elected the synchronization-site;
- Coordinator will periodically attempt to extend its mandate;
- Voter that replied positively will not vote for another site before a timeframe;



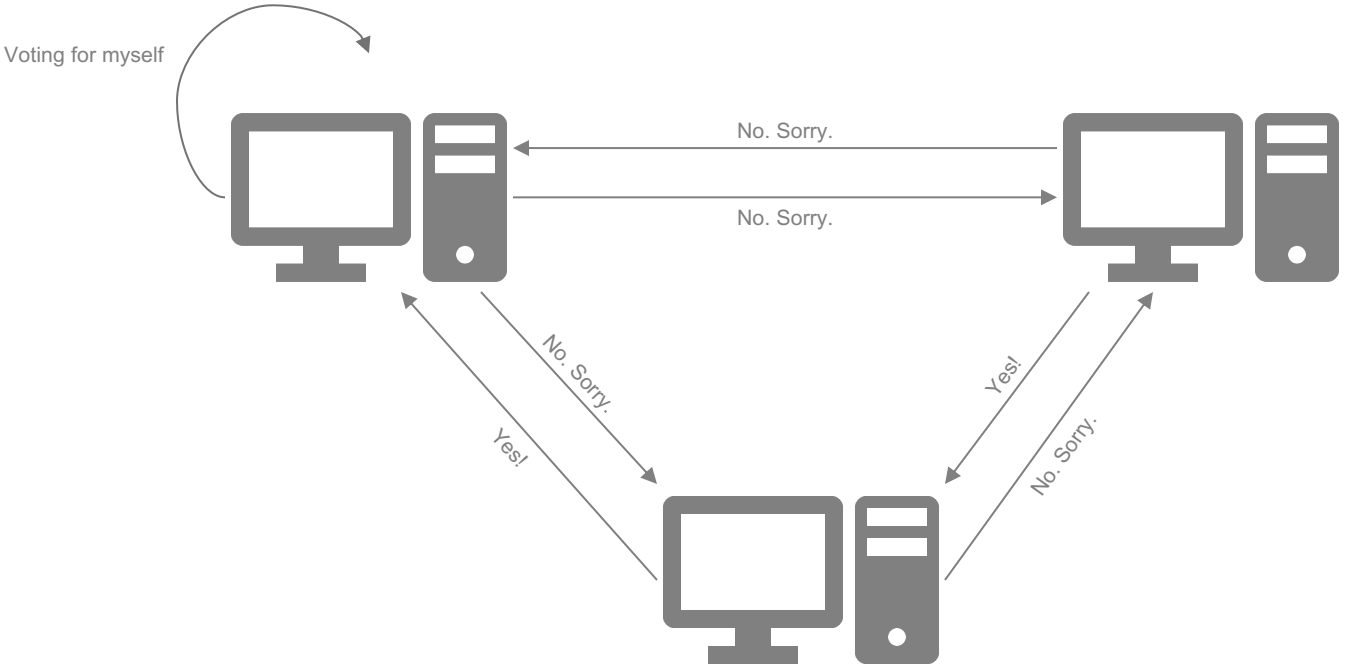
# ELECTION: OVERVIEW

SINE NOMINE ASSOCIATES



SINE NOMINE ASSOCIATES

# ELECTION: OVERVIEW



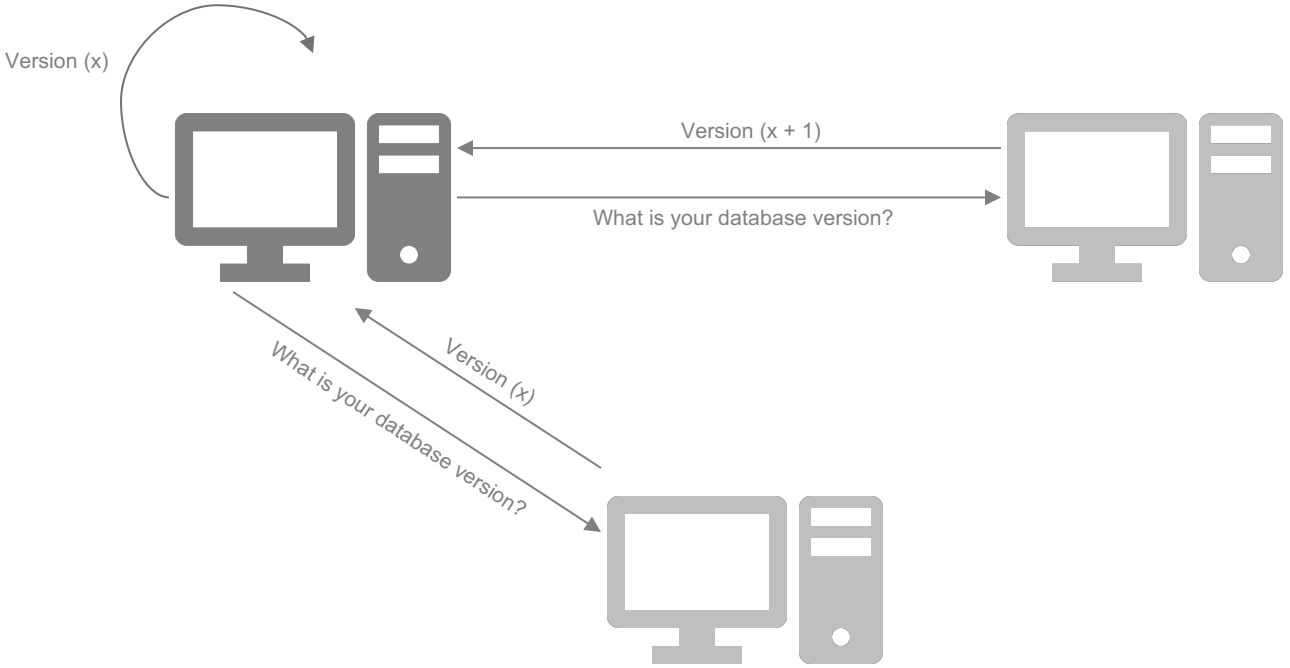

# RECOVERY

# RECOVERY

SINE NOMINE ASSOCIATES

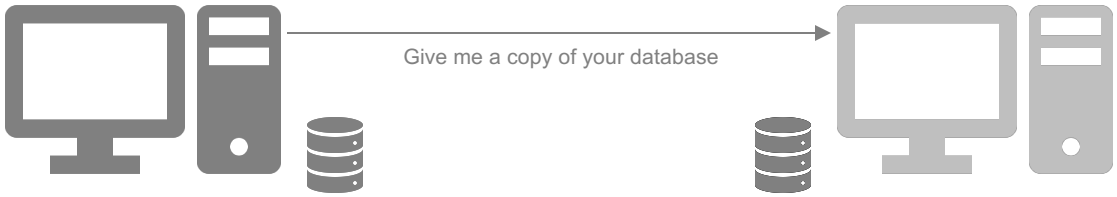
- Recovery procedure is executed immediately after becoming sync-site:
  - Sync-site contacts all servers and determines the latest version;
  - Sync-site updates its local database to the latest version;
  - Coordinator relabels the database as the first version during his mandate;
  - Sync-site updates all remote databases to the latest version;

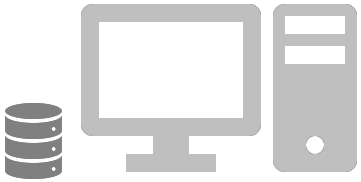
# RECOVERY: OVERVIEW



# RECOVERY: OVERVIEW

SINE NOMINE ASSOCIATES



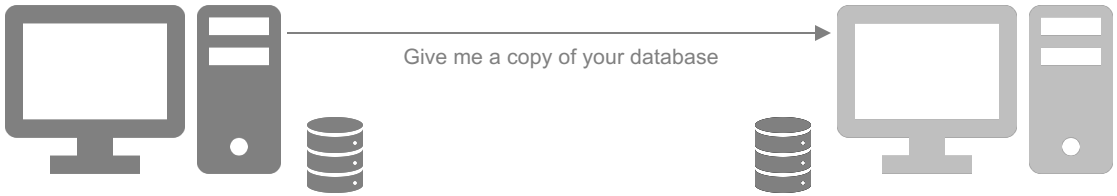



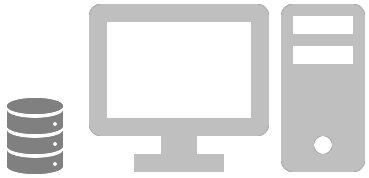
# RECOVERY: LIMITATIONS

SINE NOMINE ASSOCIATES

- Read-transactions not allowed;
- Write-transactions not allowed;
- In other words, sites involved are not available during this phase;
- Why? Because the live database is being replaced;
- Current version does not replace the database directly;
  - Received database is stored in a temporary file;
  - Temporary file replaces live database;

# RECOVERY: OVERVIEW



# READS-DURING-SYNC

SINE NOMINE ASSOCIATES

- Scenario 1: site is sending a copy of the database;
  - Local database is not being modified;
  - Writes must be blocked;
  - There is no reason to block reads;
  - Allow reads but block writes;
- Scenario 2: site is receiving a copy of the database;
  - Received data is stored in a temporary file;
  - Live database will (eventually) be replaced by this temporary file;
  - Writes must be blocked;
  - Reads can be allowed until the replacement-phase;
  - Replacement-phase blocks new reads;
  - Replacement-phase aborts read-transactions;


# TRANSACTIONS

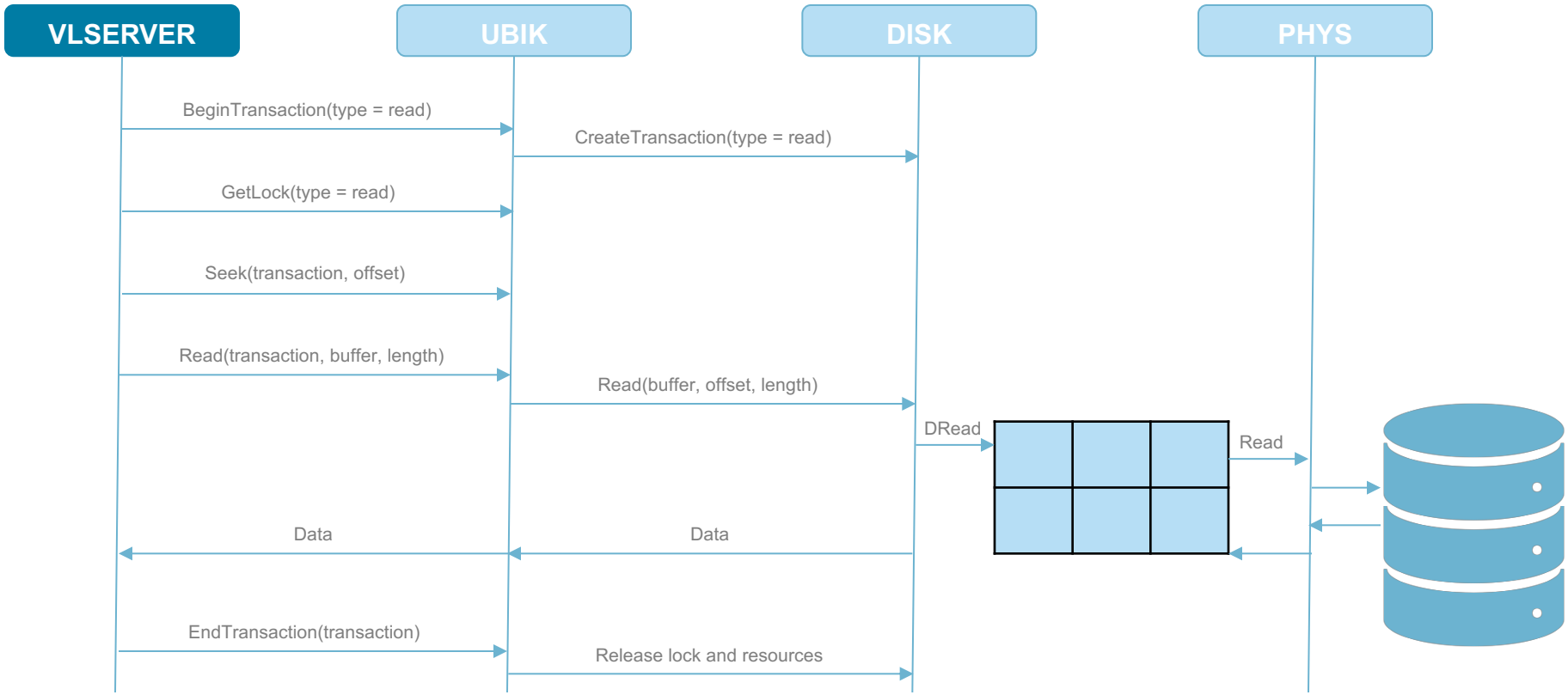
# READ-TRANSACTIONS

SINE NOMINE ASSOCIATES

- Read-transactions;
  - Designed to handle a high number of read transactions;
  - Executed by any server in the quorum;
  - Can handle multiple read-transactions at the same time;
  - Locking is done locally to the server receiving the request;
  - Reads data from a database under a transaction;
  - The parameters to read are a transaction, a buffer and a length;
  - It functions like the Unix read system call;
  - Reads length bytes from the current file position into the specified buffer;

# READ-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



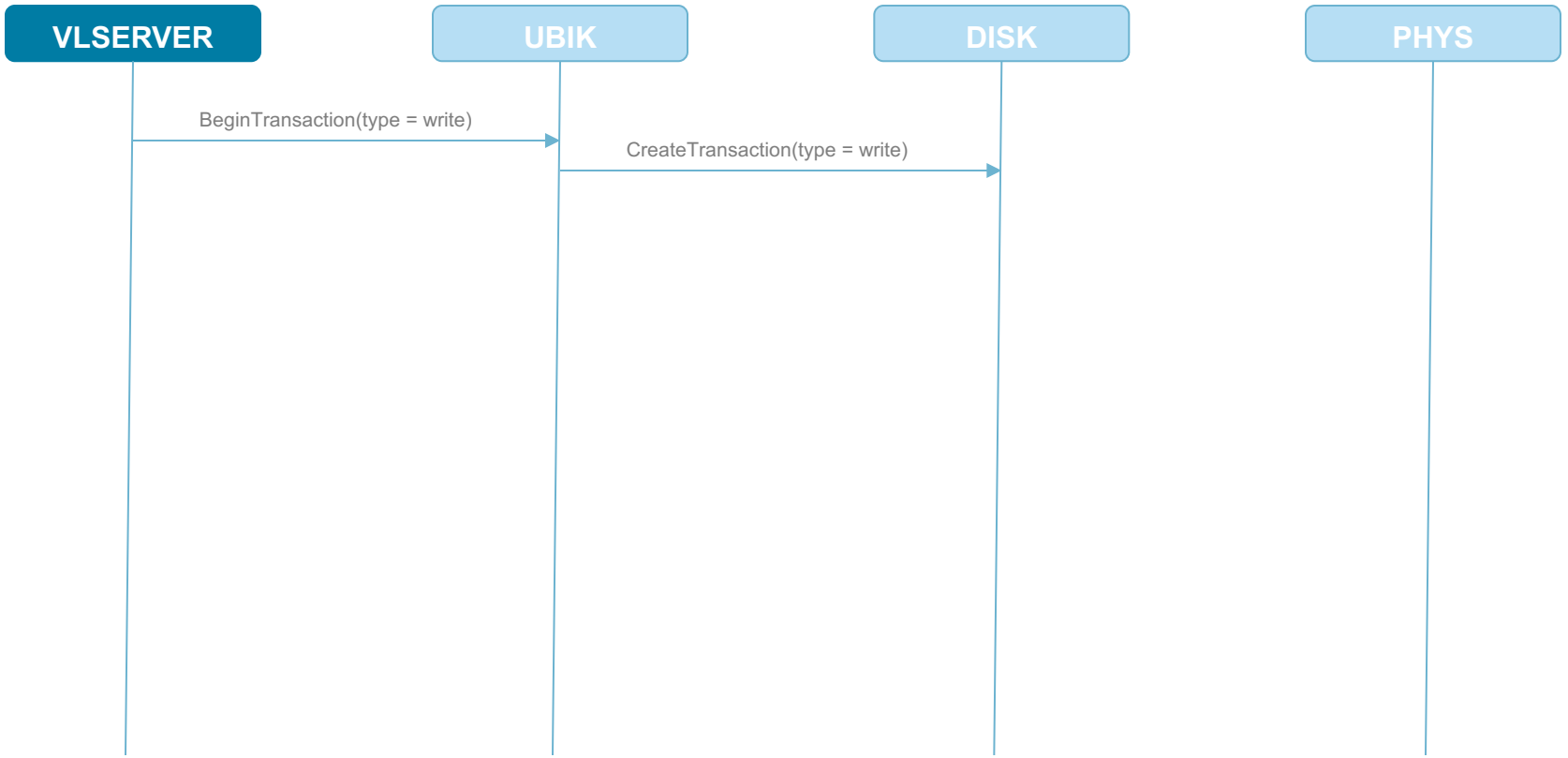
# WRITE-TRANSACTIONS

SINE NOMINE ASSOCIATES

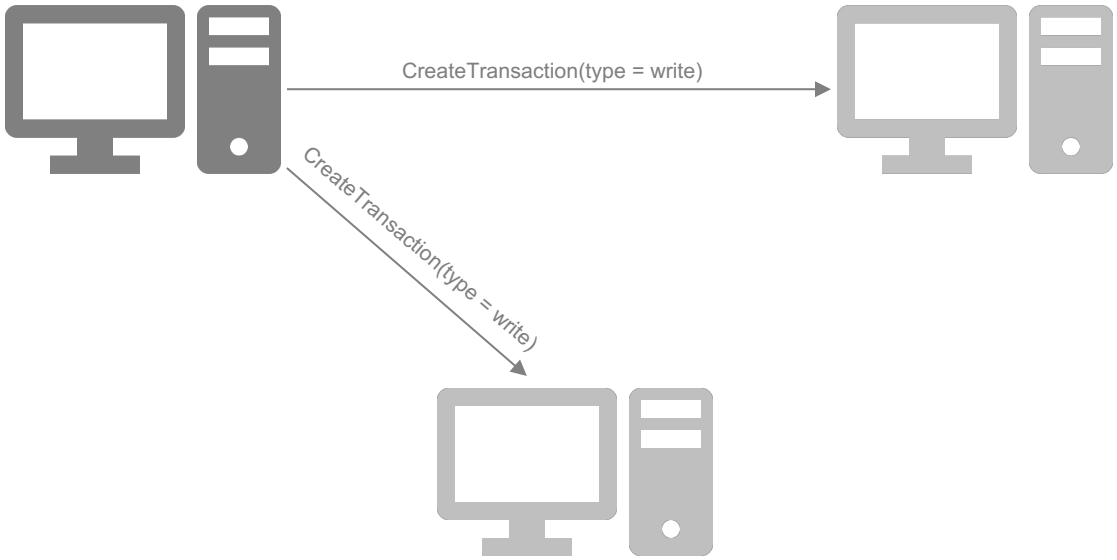
- Write-transactions;
  - Not designed to handle a high number of write transactions;
  - Executed by every server in the quorum;
  - Can only handle one write-transaction at a specific time;
  - Locking is done globally;
  - Writes data to a database under a transaction;
  - The parameters to write are a transaction, a buffer and a length;
  - It functions like the Unix read system call;
  - Writes length bytes at the current file position from the specified buffer;

# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

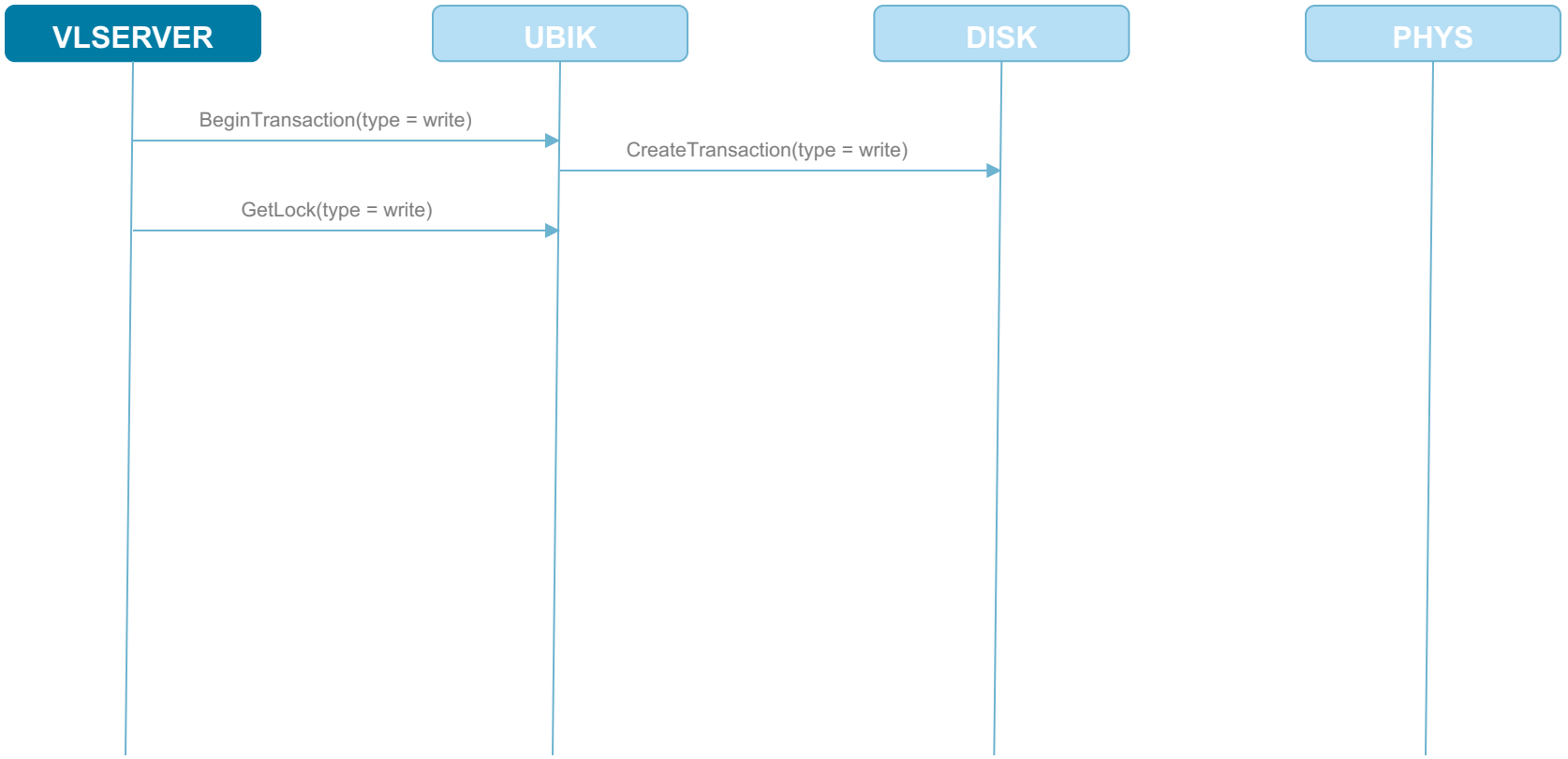


# WRITE-TRANSACTIONS (SIMPLIFICATION)

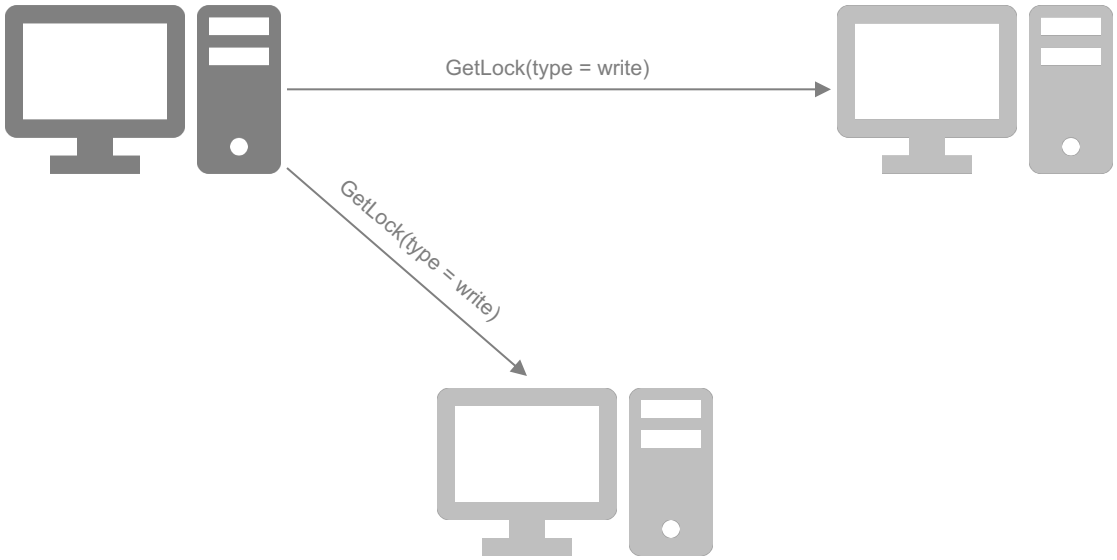


# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

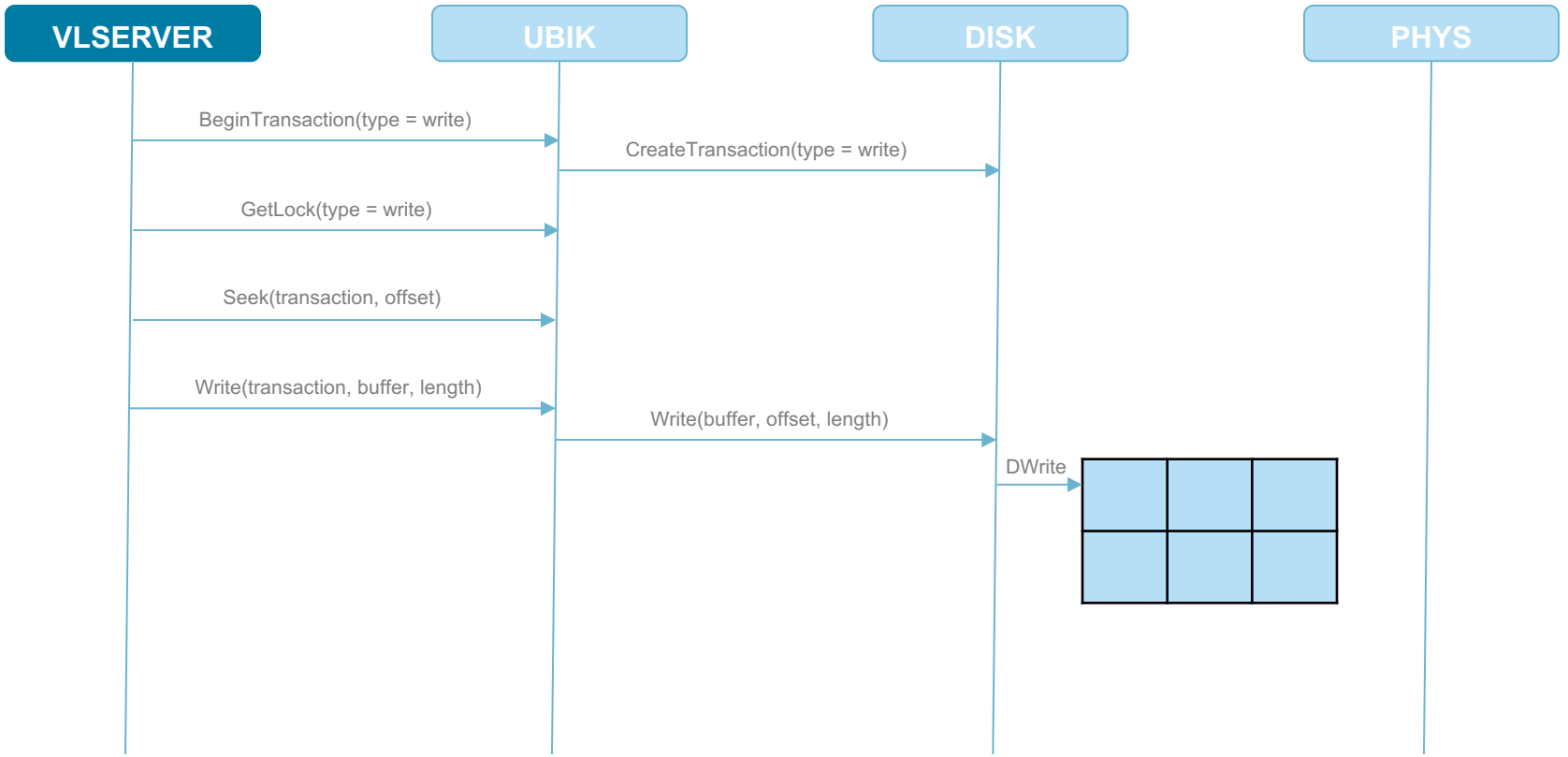


# WRITE-TRANSACTIONS (SIMPLIFICATION)



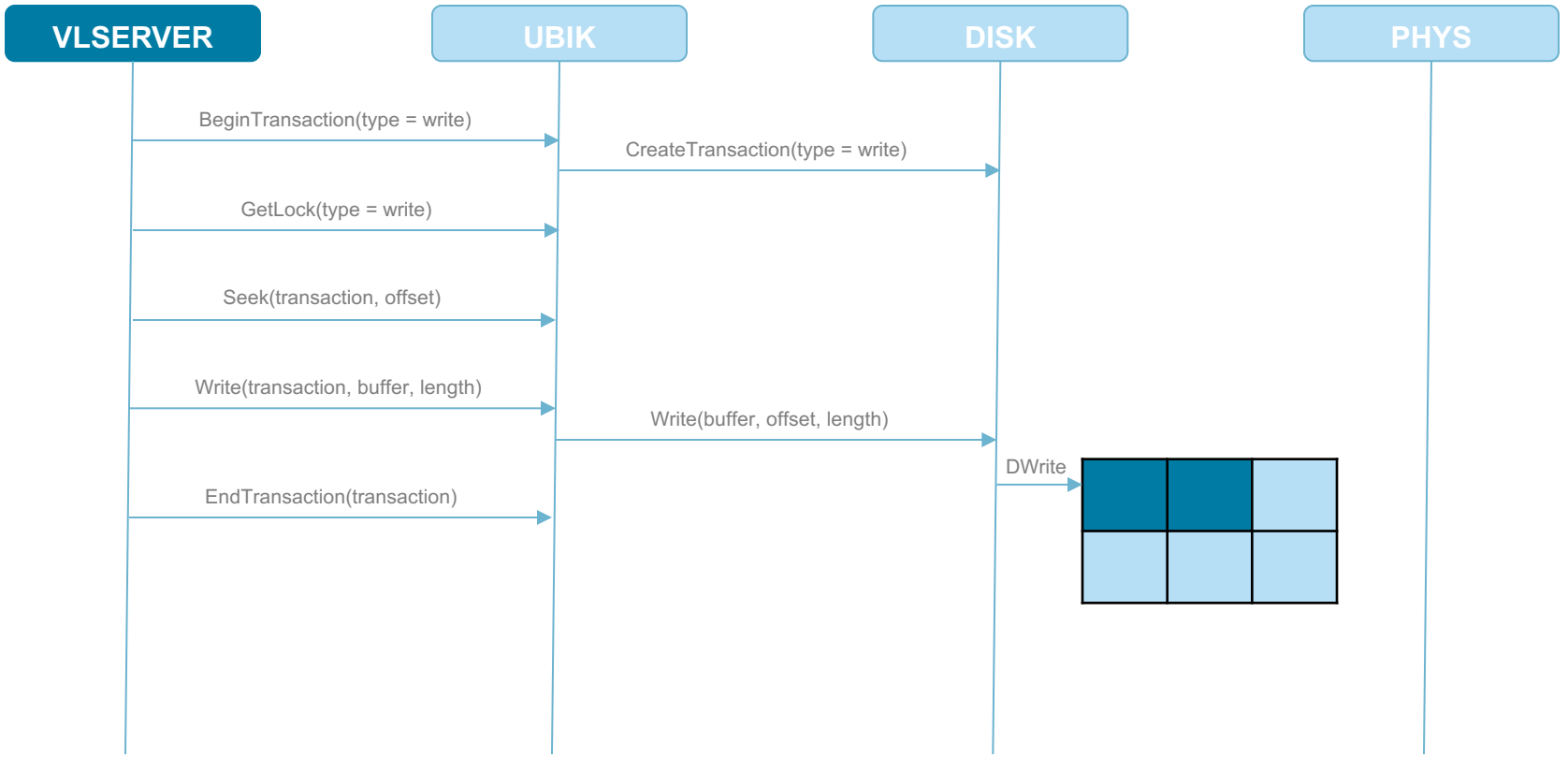
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



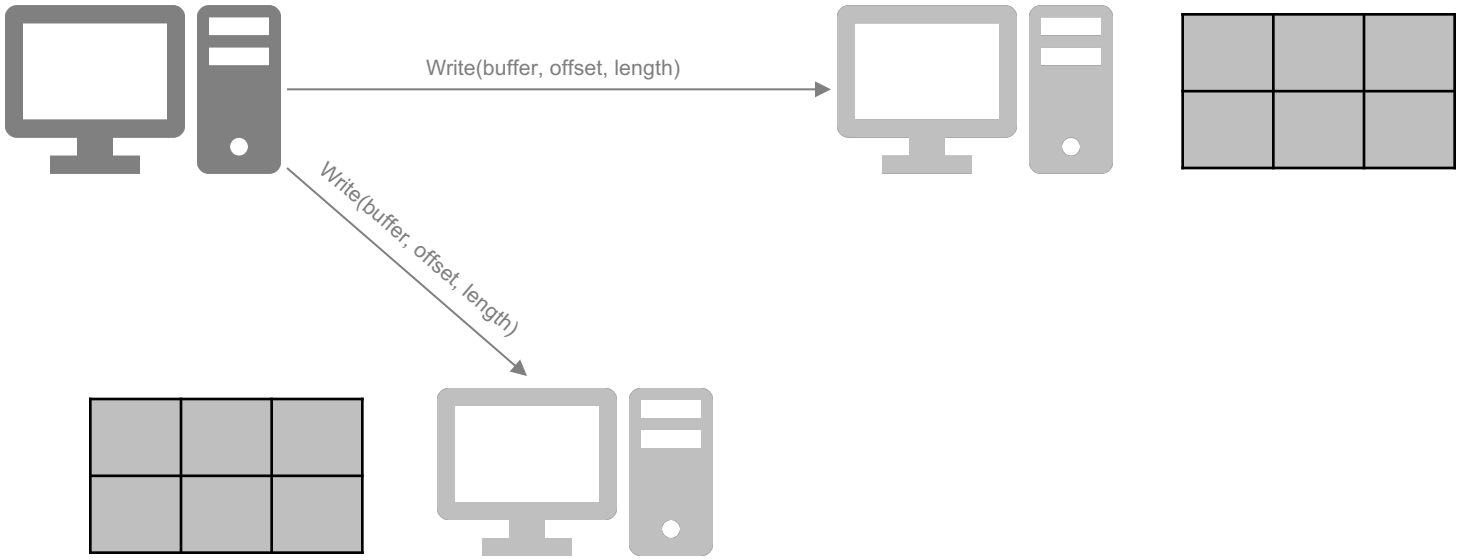
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

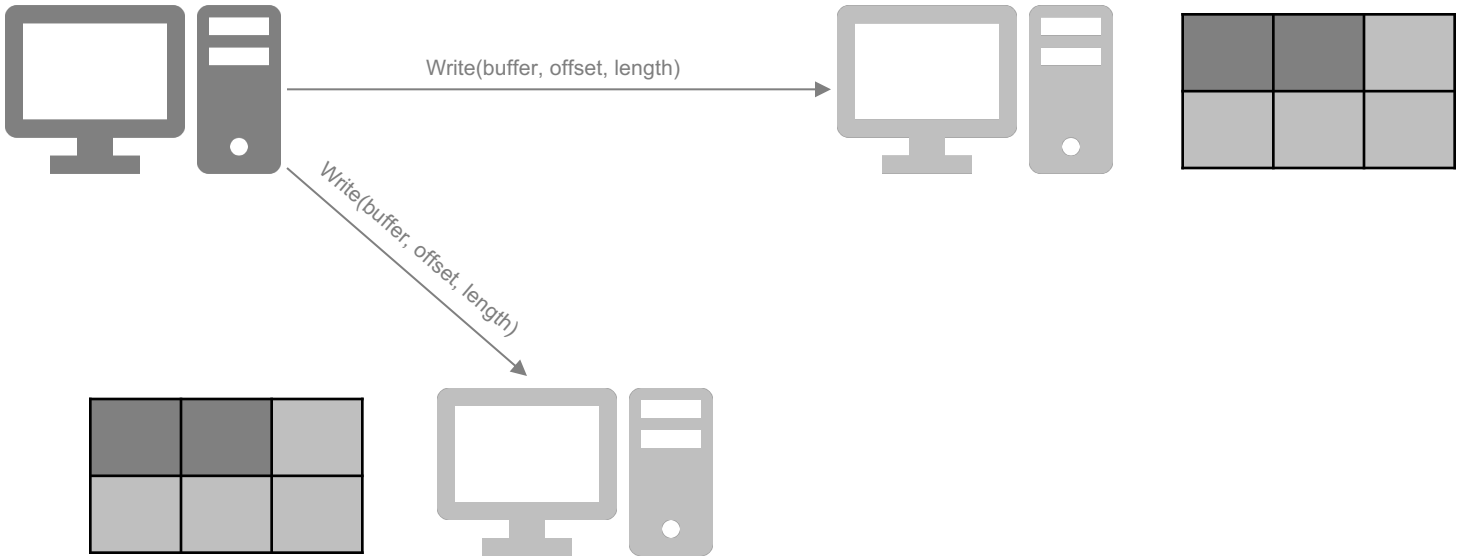


# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

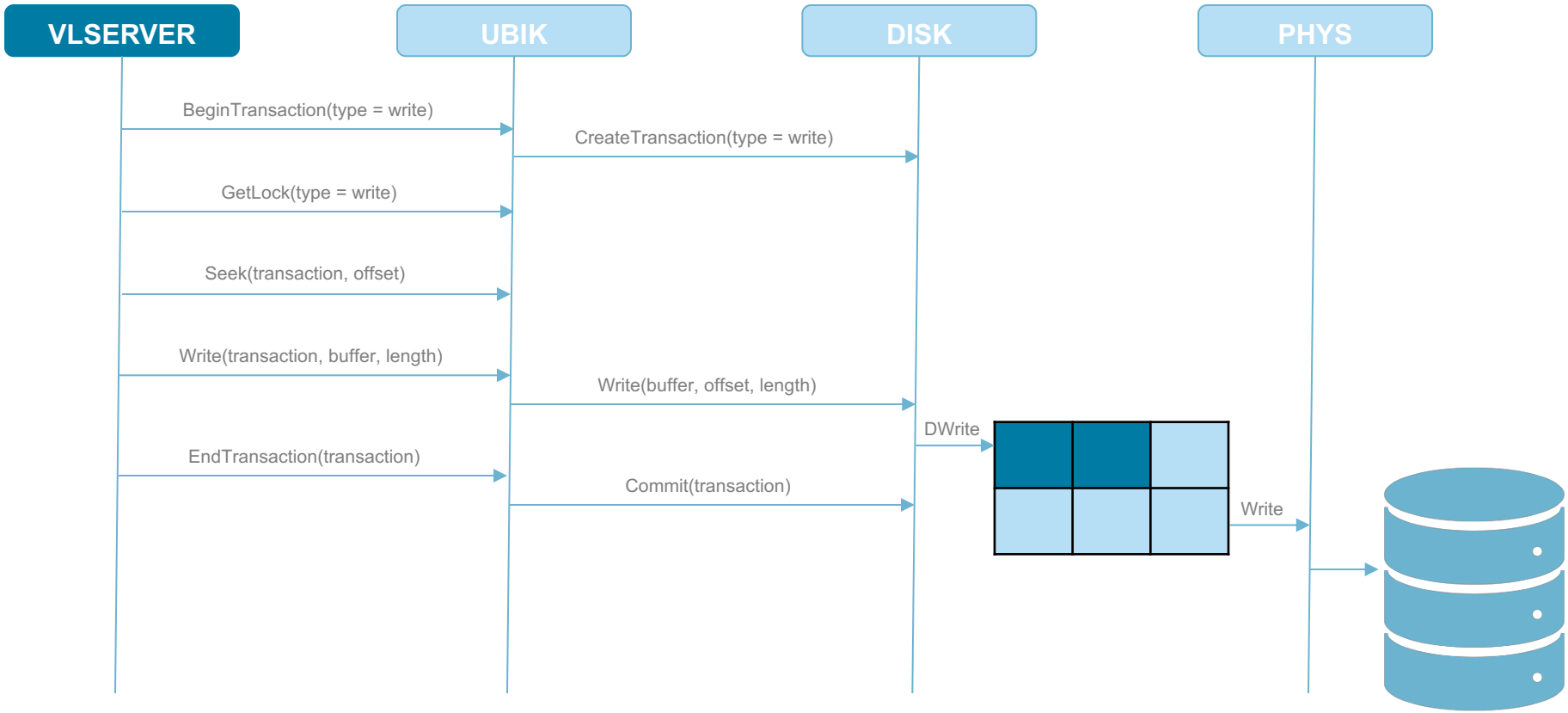


# WRITE-TRANSACTIONS (SIMPLIFICATION)



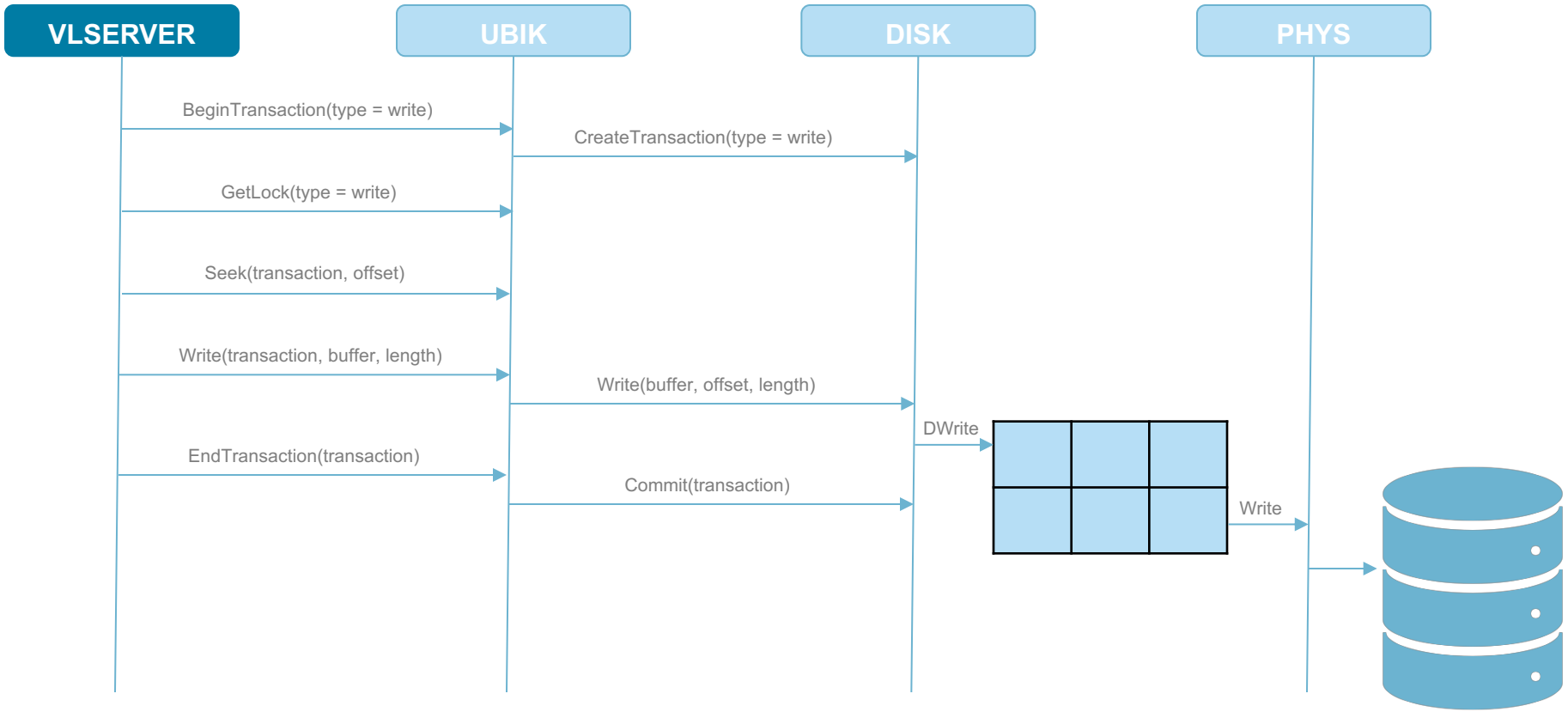
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



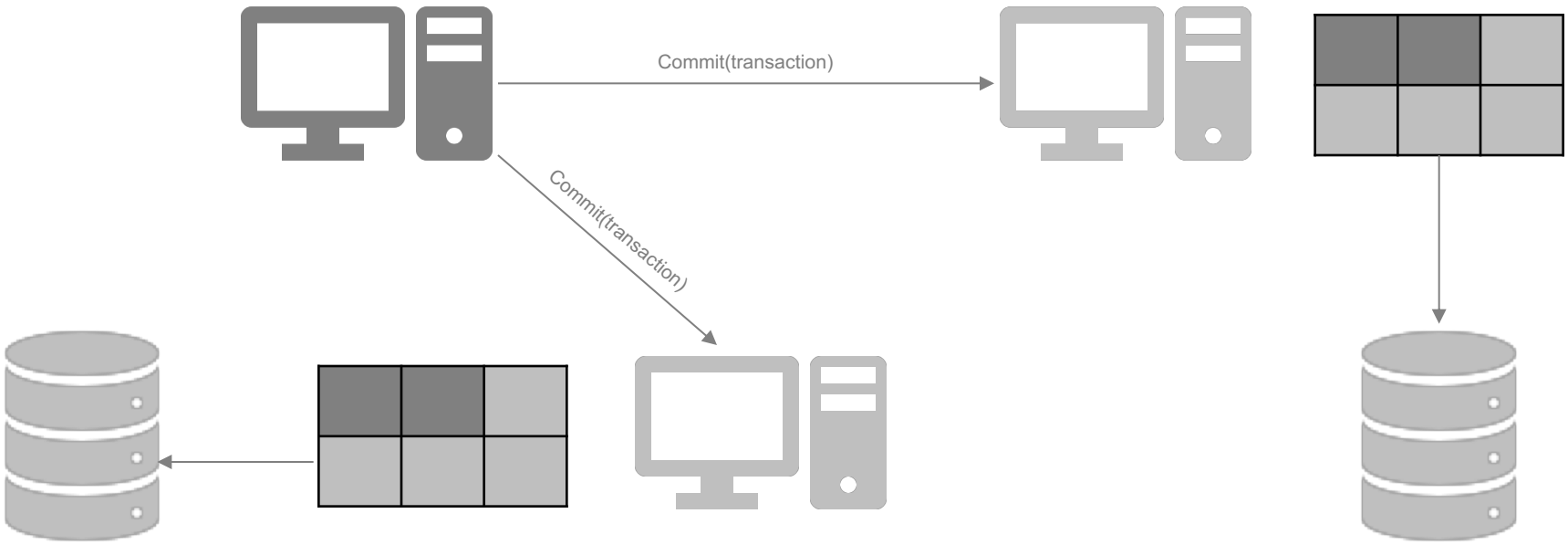
SINE NOMINE ASSOCIATES

# WRITE-TRANSACTIONS (SIMPLIFICATION)



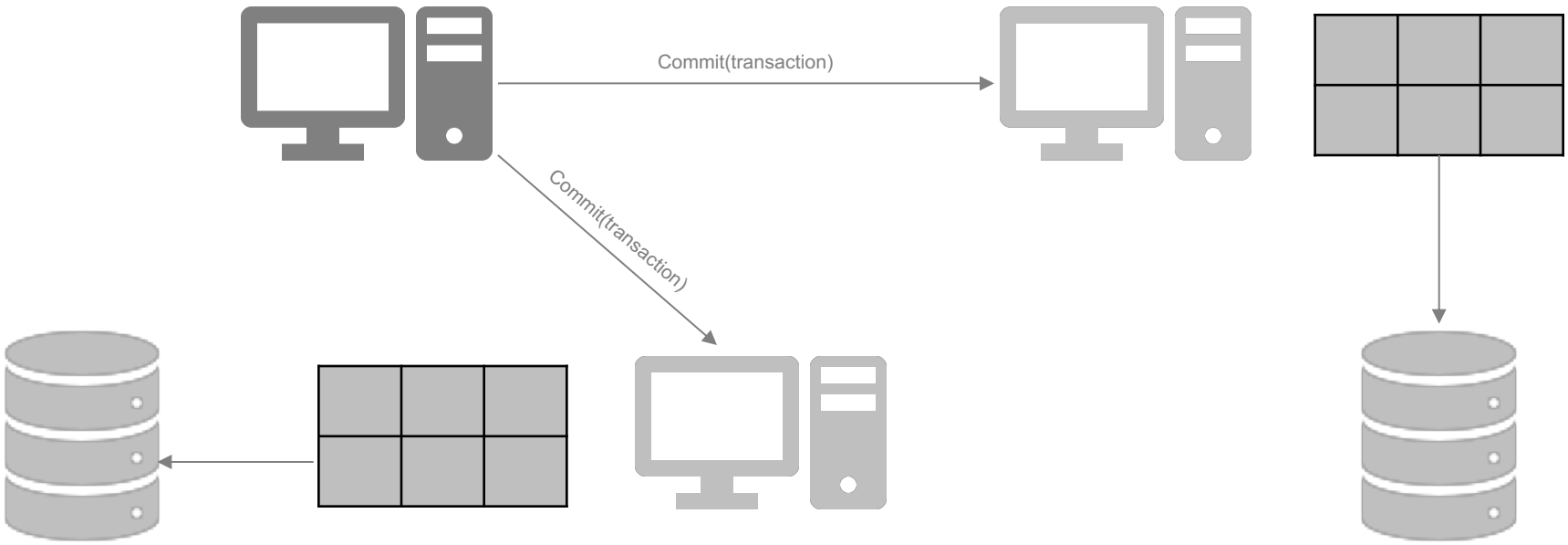
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



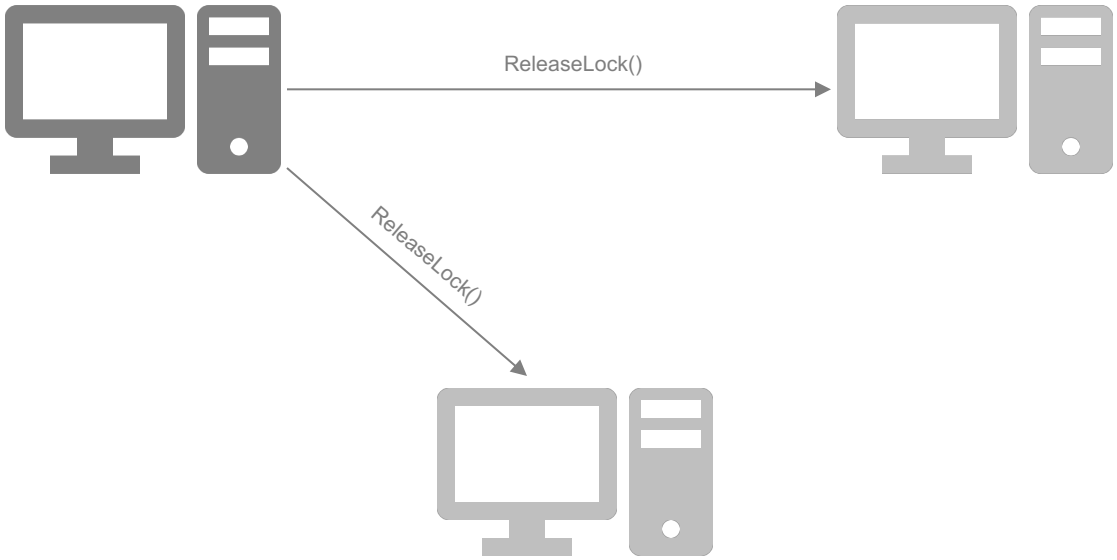
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



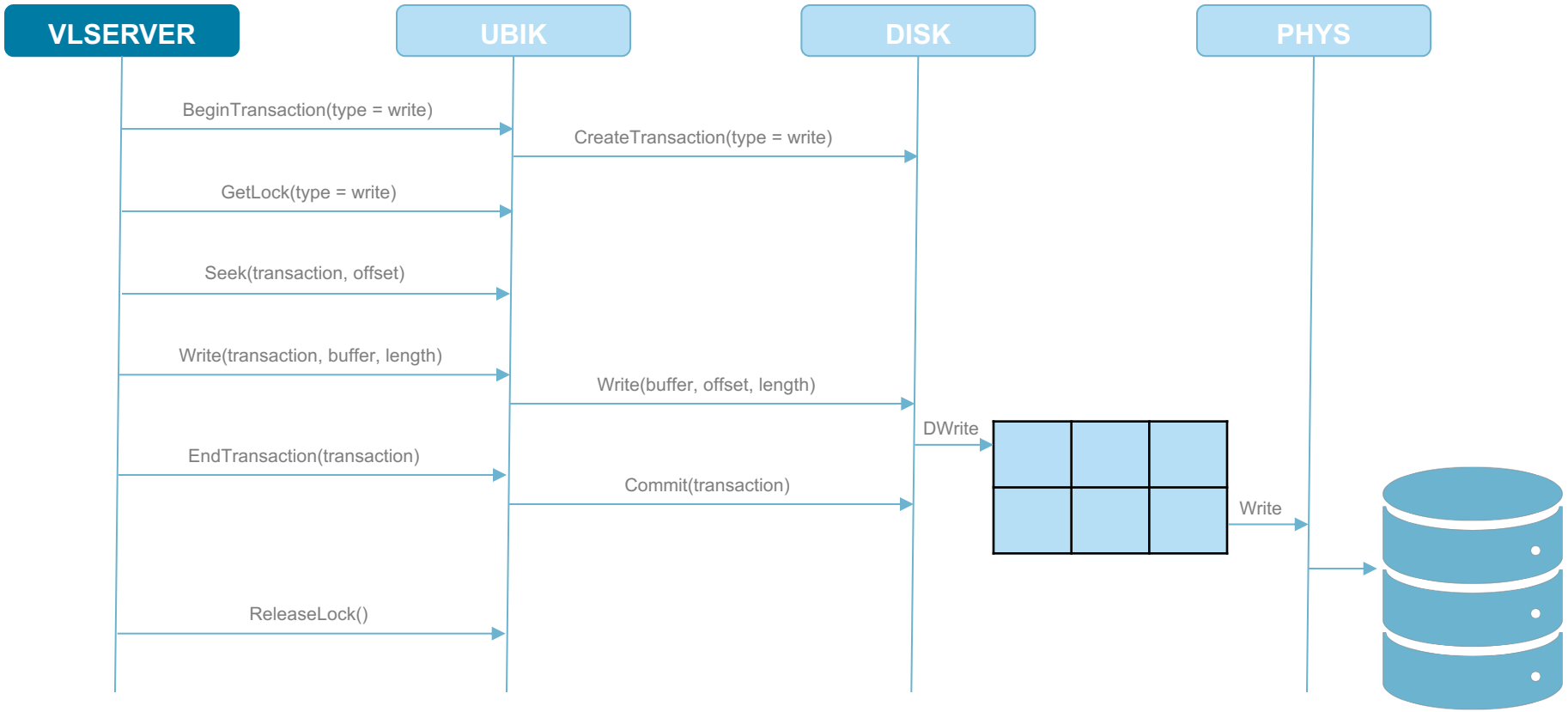
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



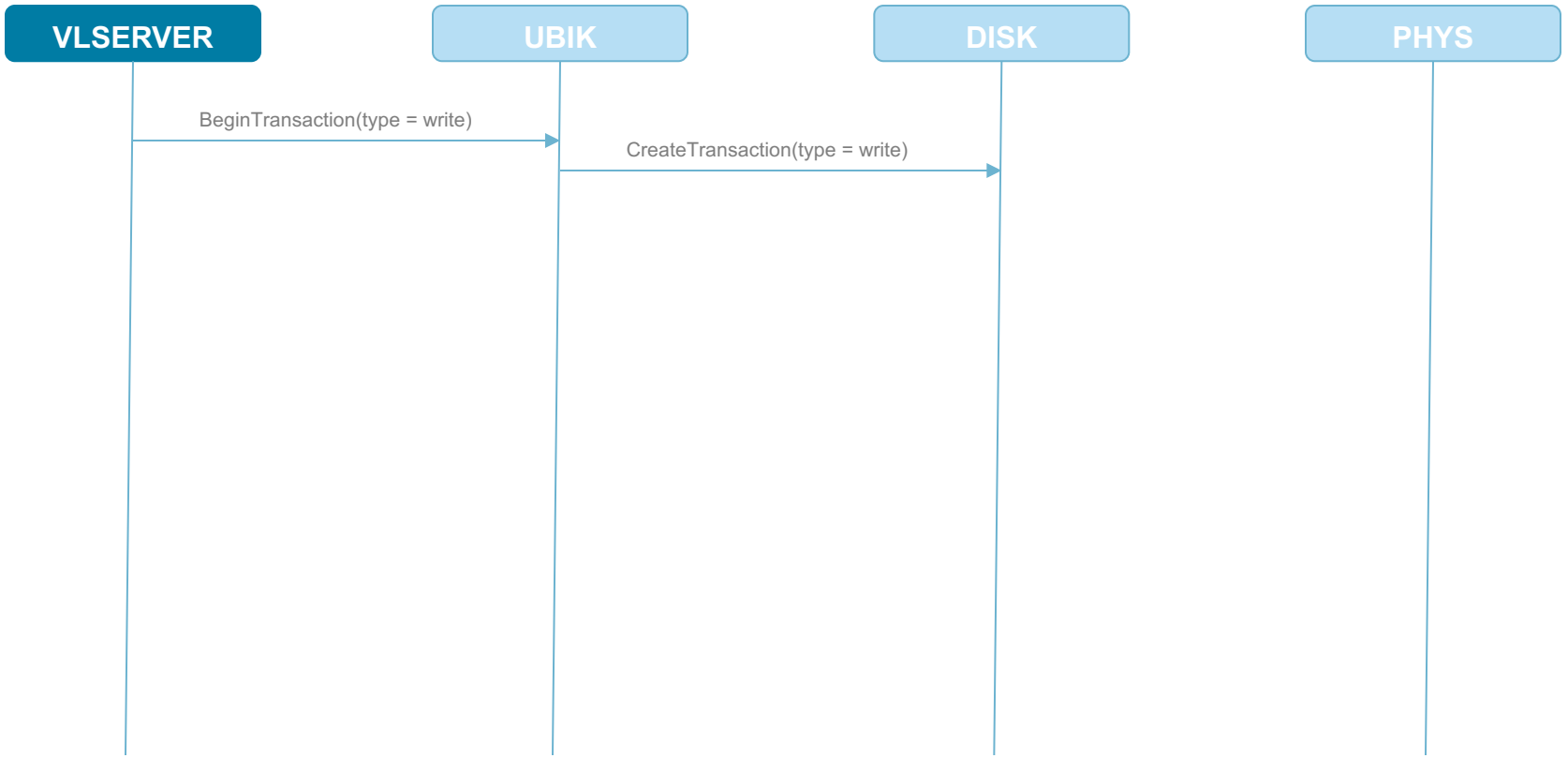
# WRITE-TRANSACTIONS

SINE NOMINE ASSOCIATES

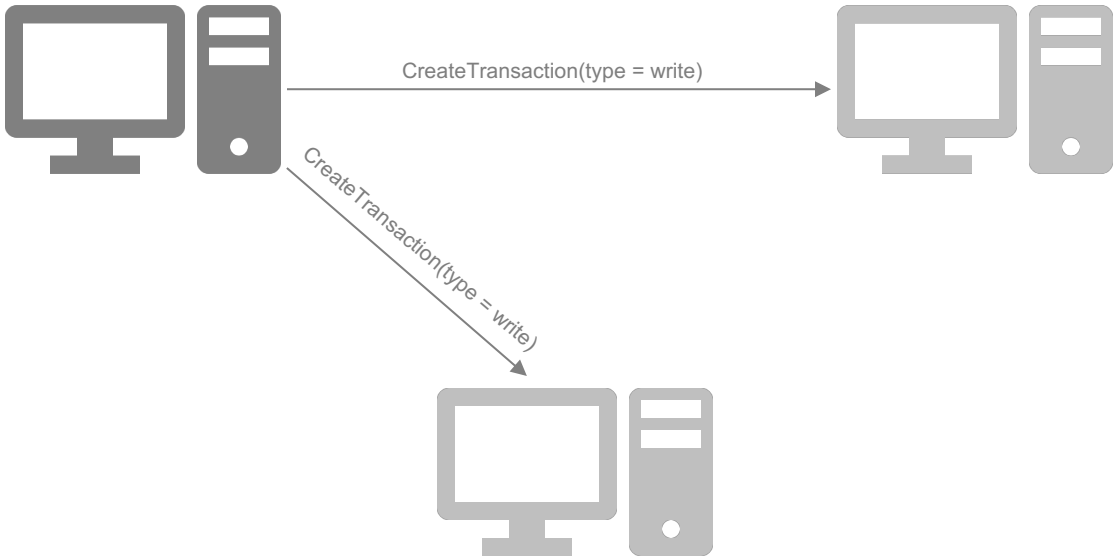
- Can be slow (communication cost);
- Read-transactions not allowed during this process;
- Whole cell is blocked;
- How can we alleviate this problem?
  - Allowing reads-during-write;

# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

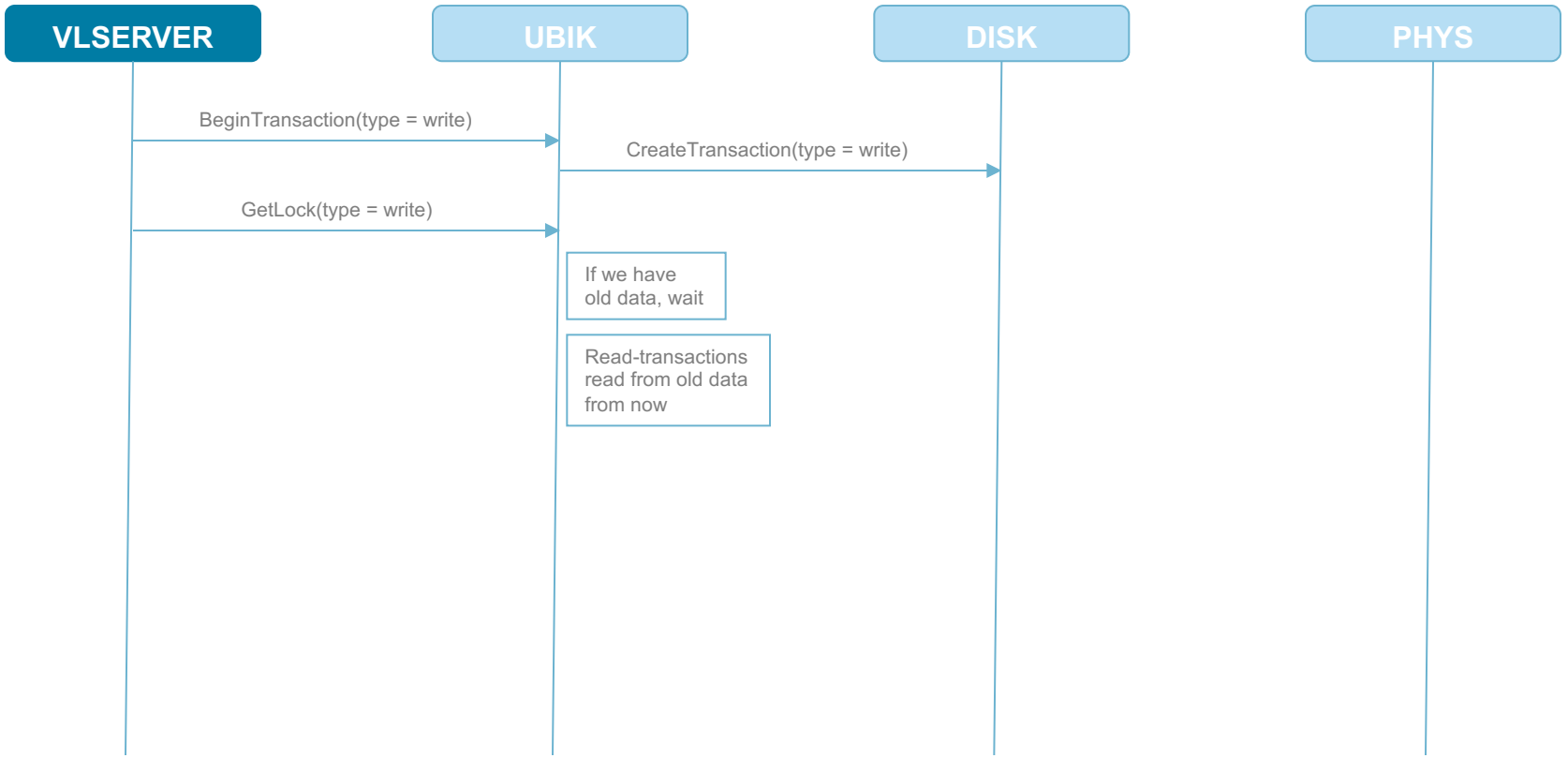


# WRITE-TRANSACTIONS (SIMPLIFICATION)



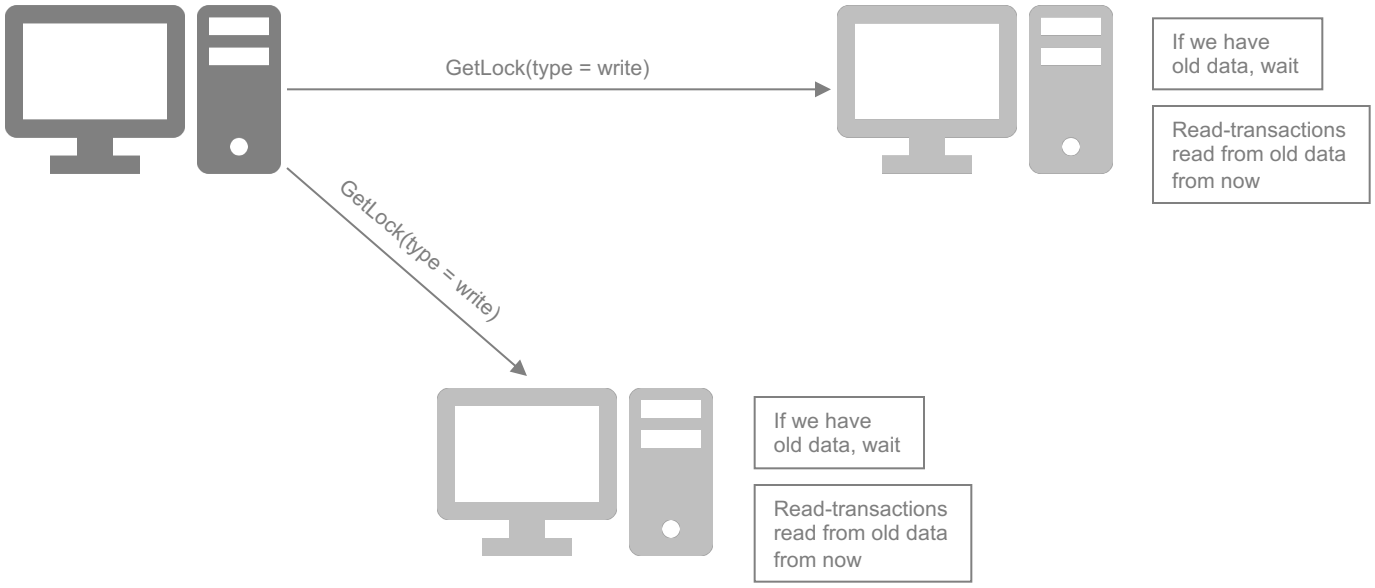
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



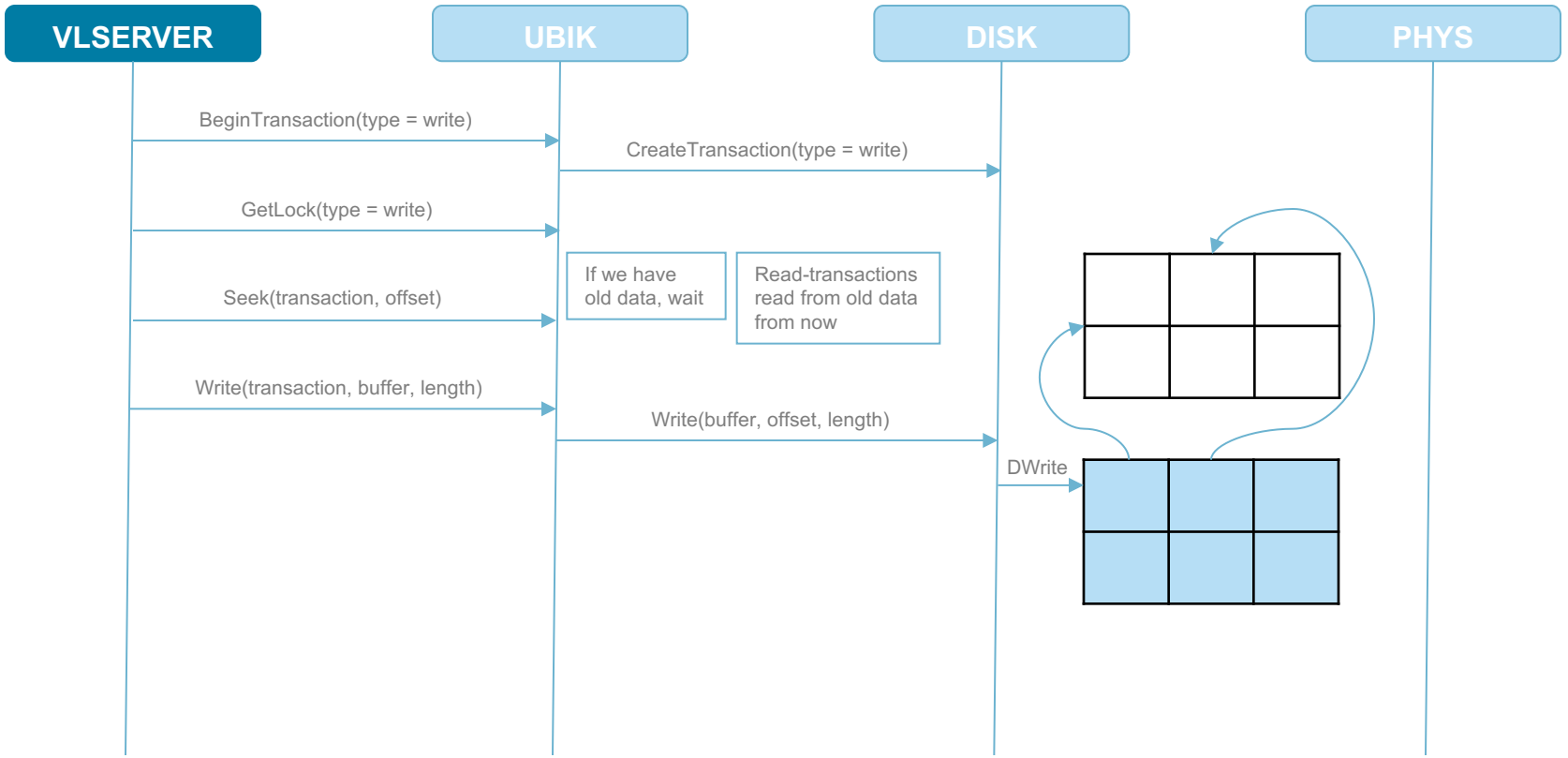
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



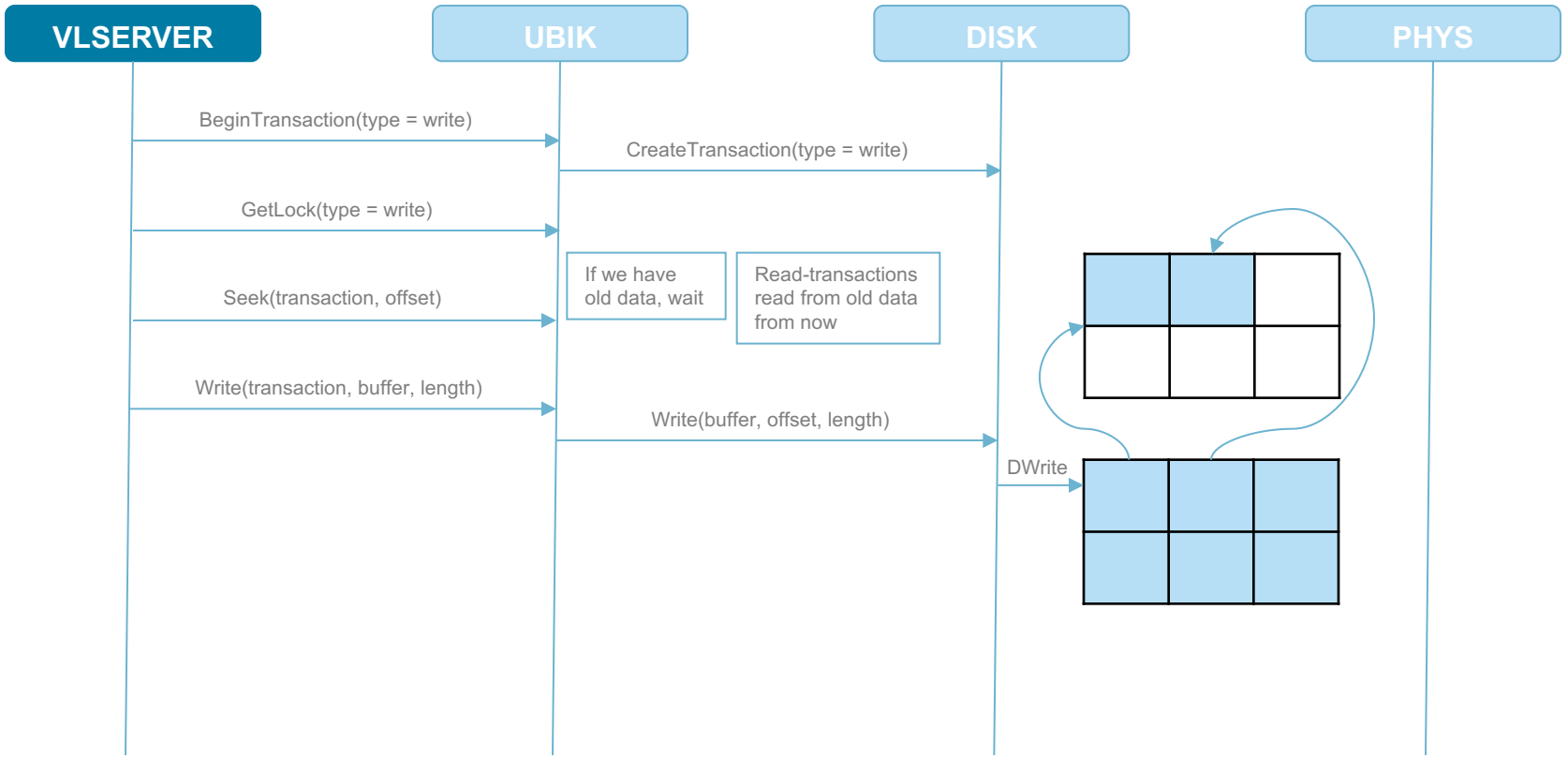
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



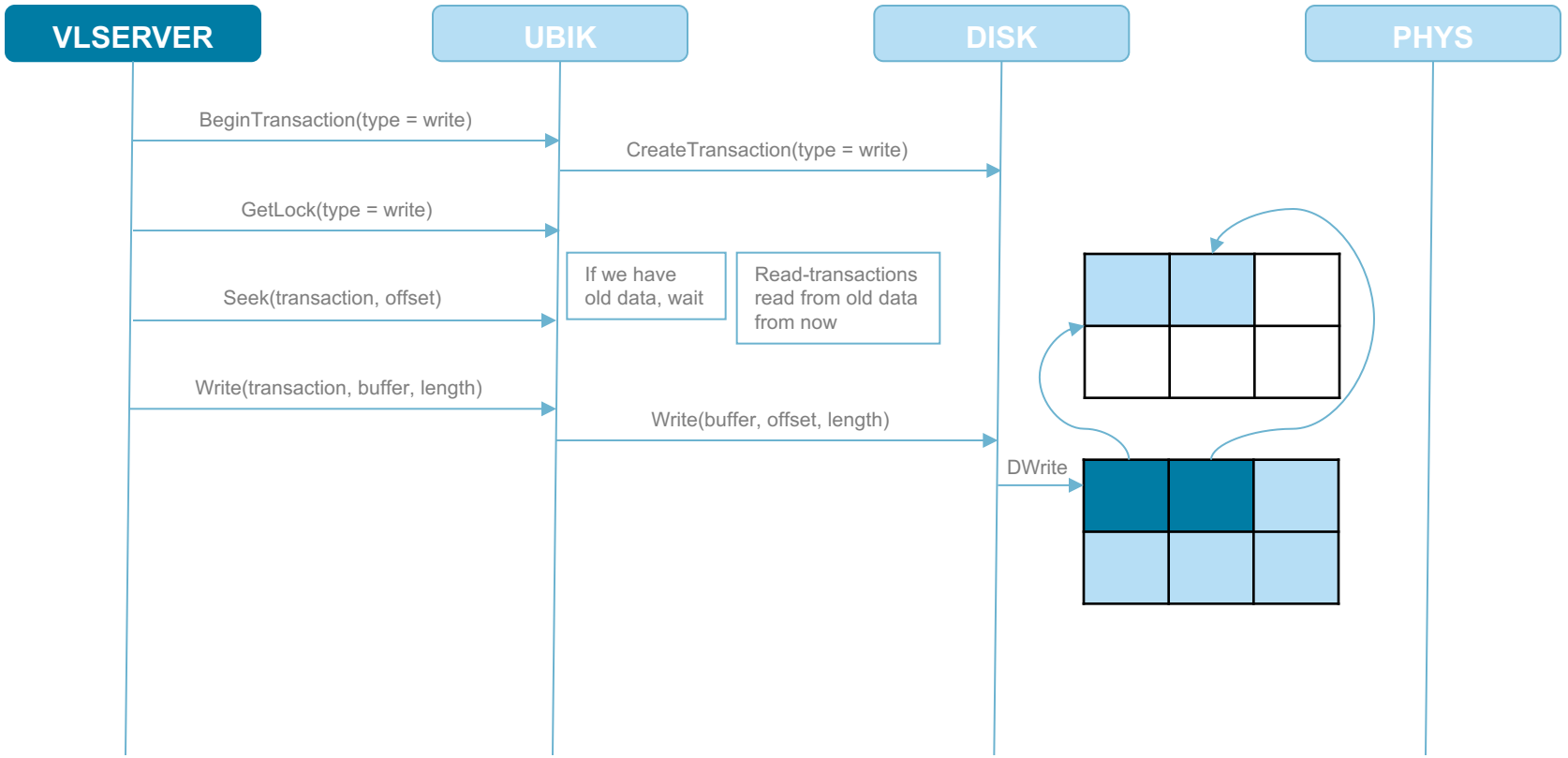
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



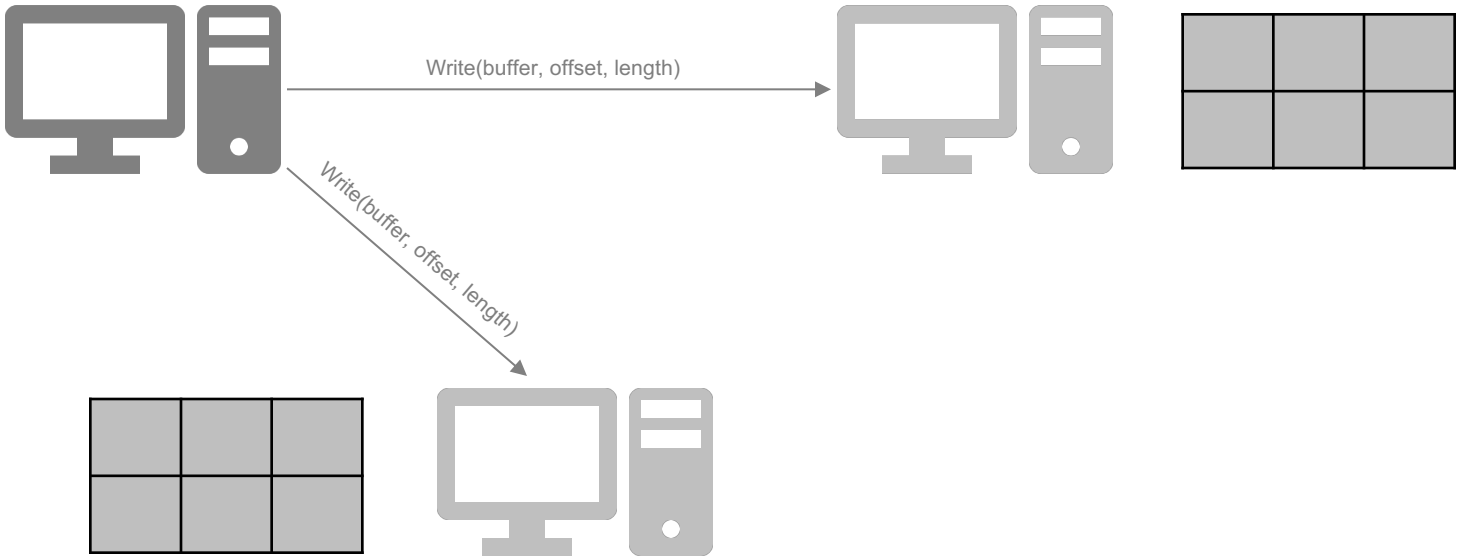
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



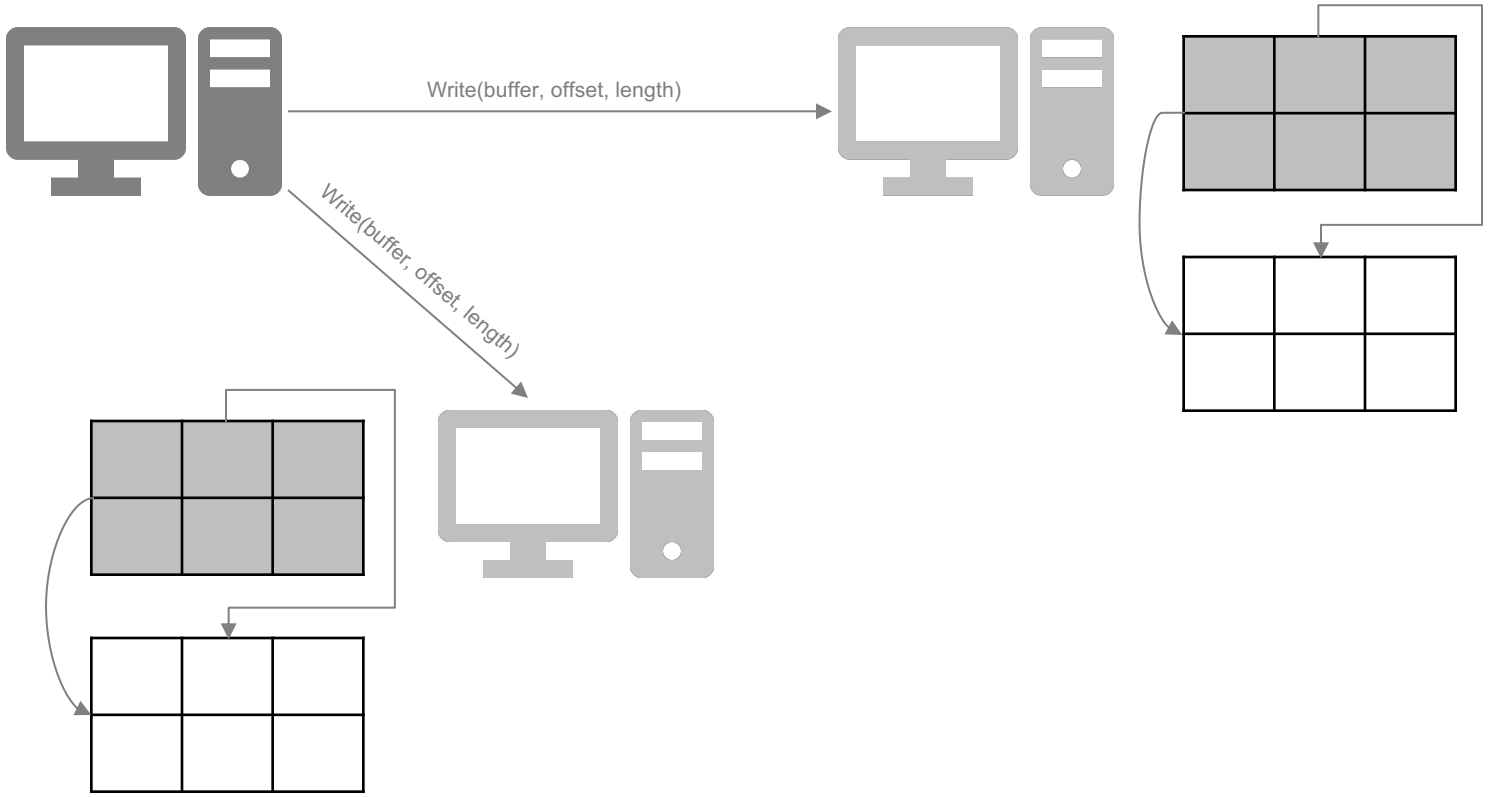
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

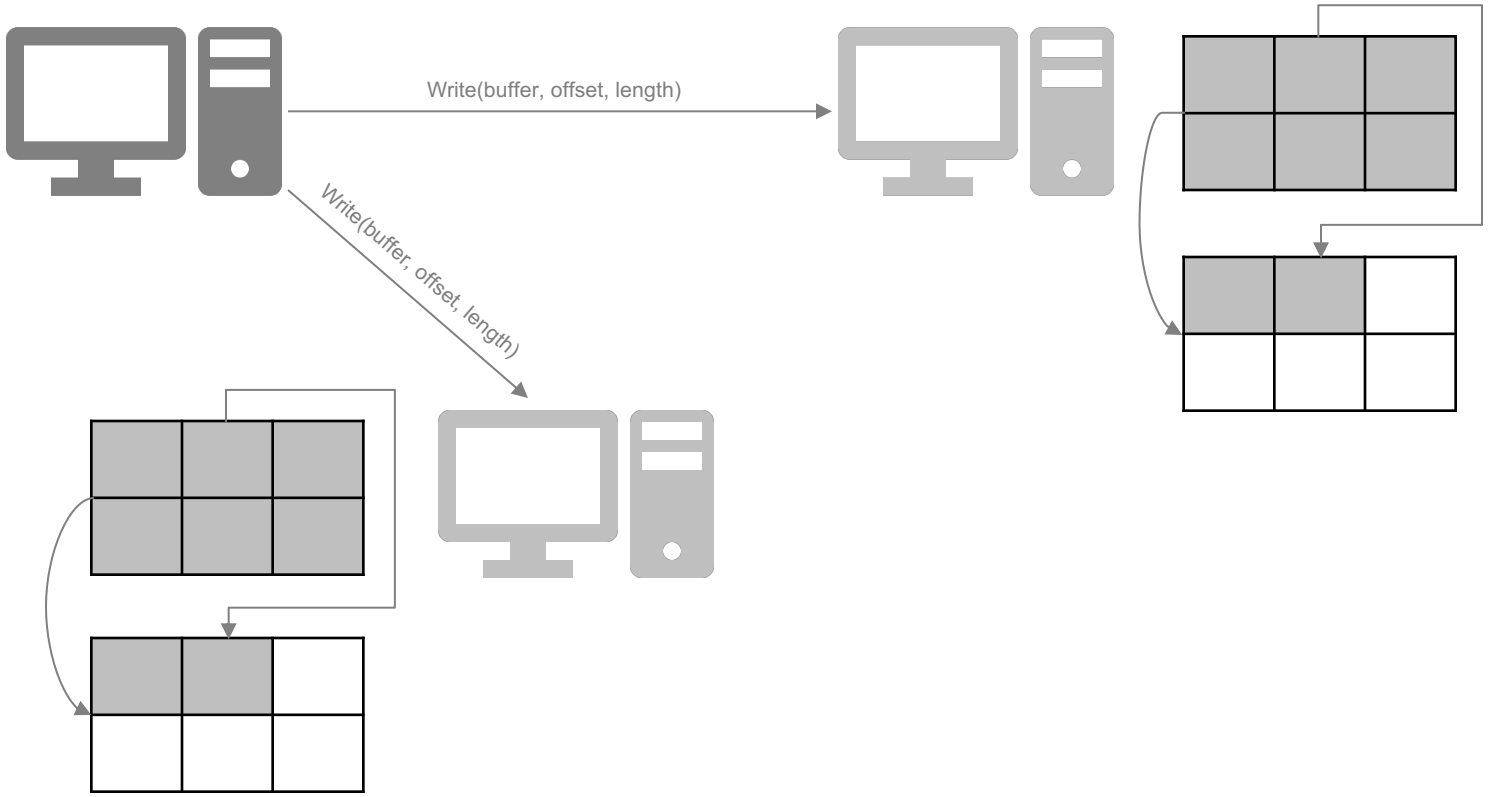


# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES

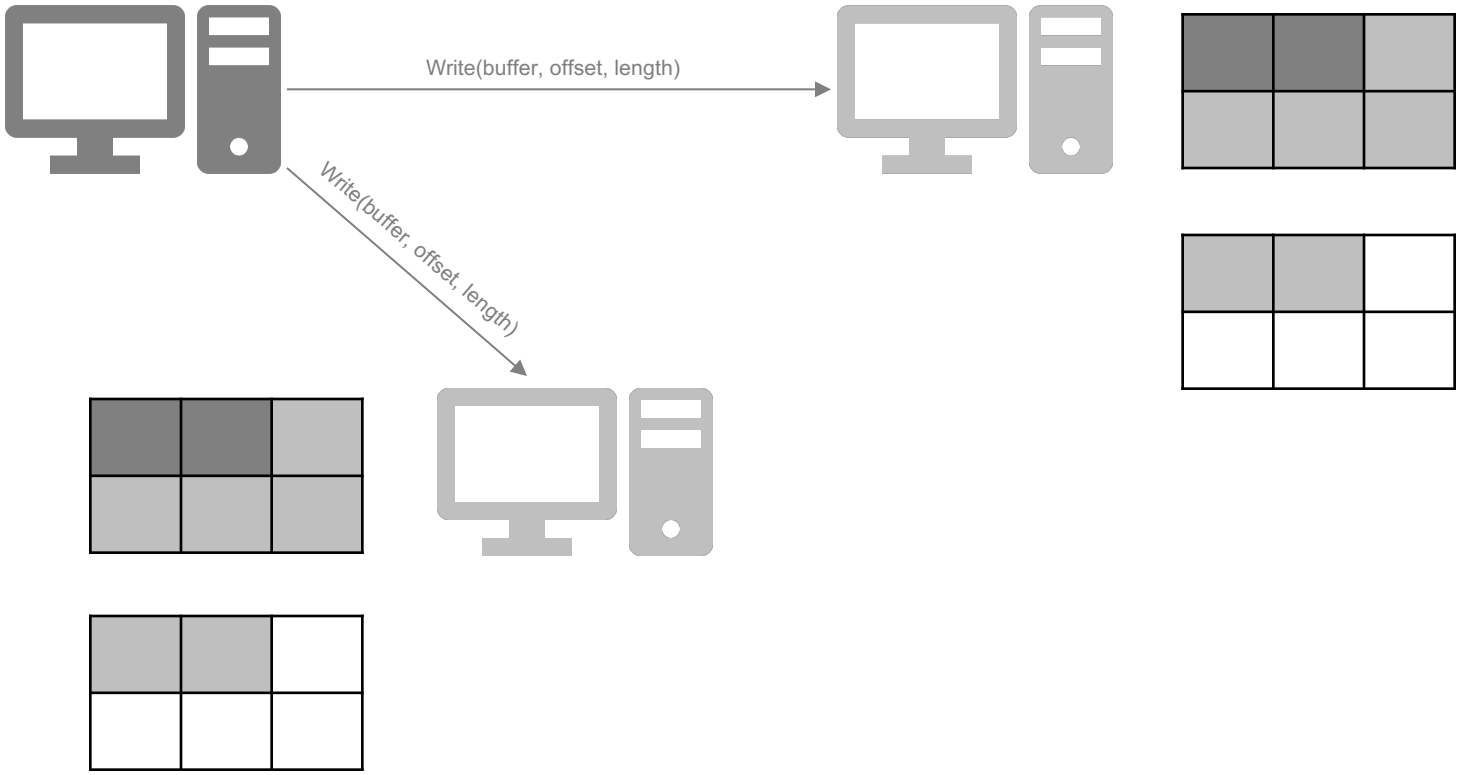


# WRITE-TRANSACTIONS (SIMPLIFICATION)



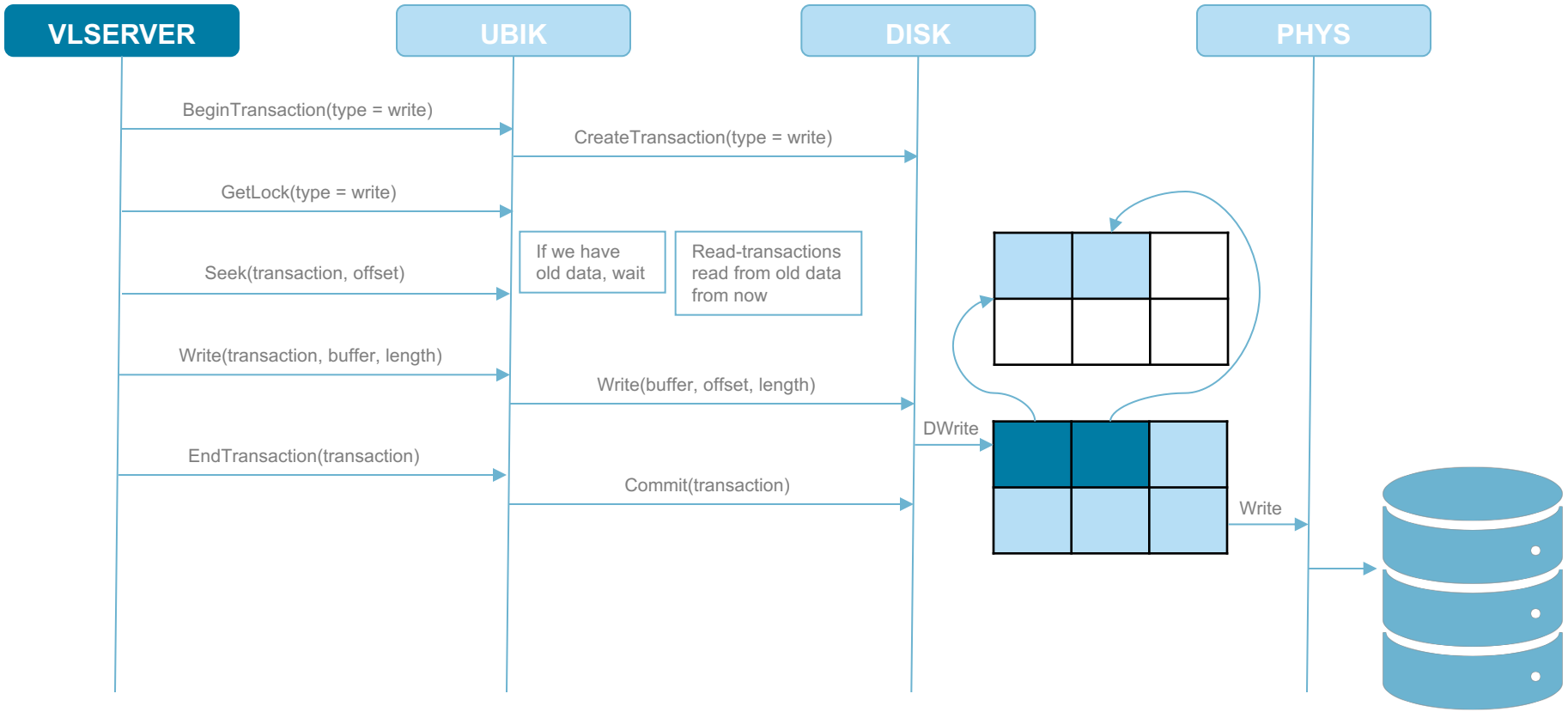
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



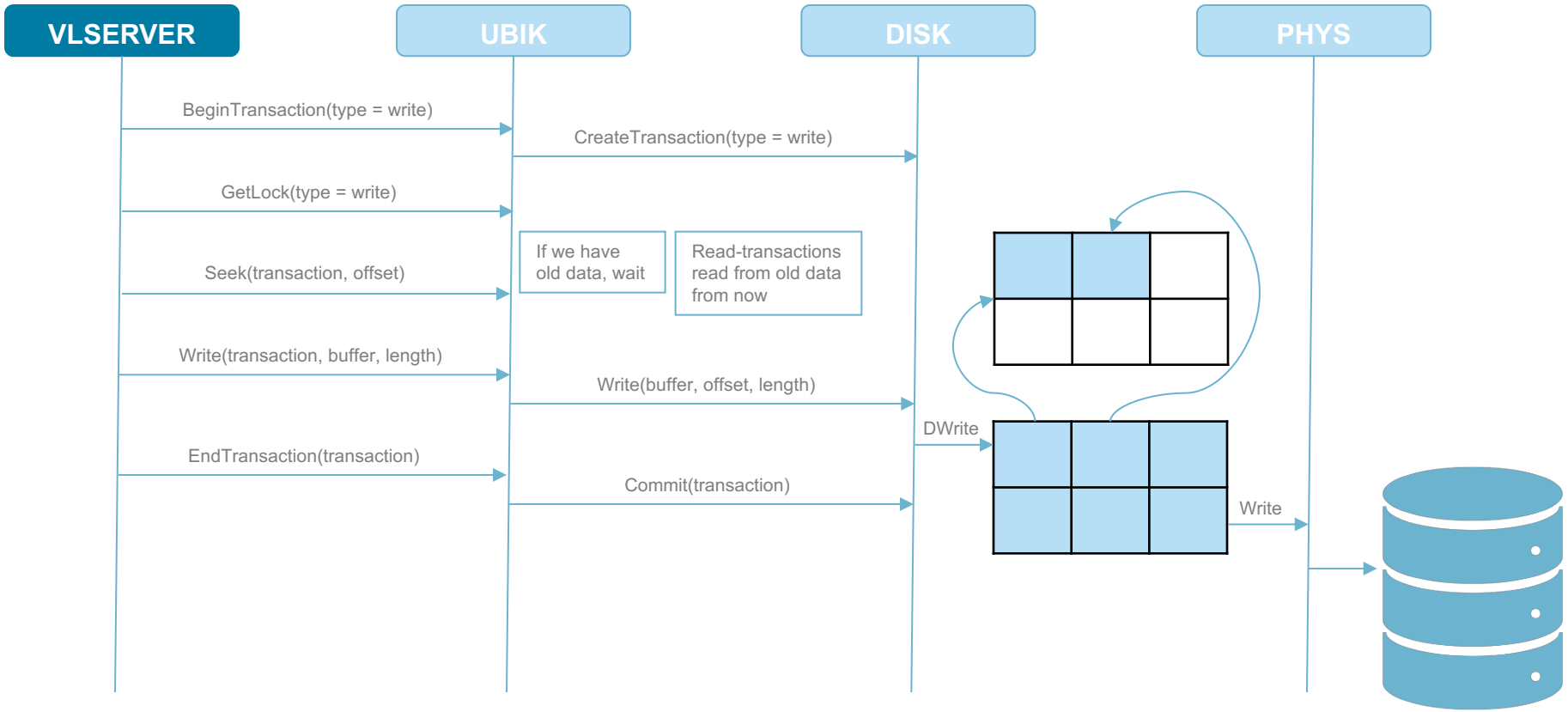
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



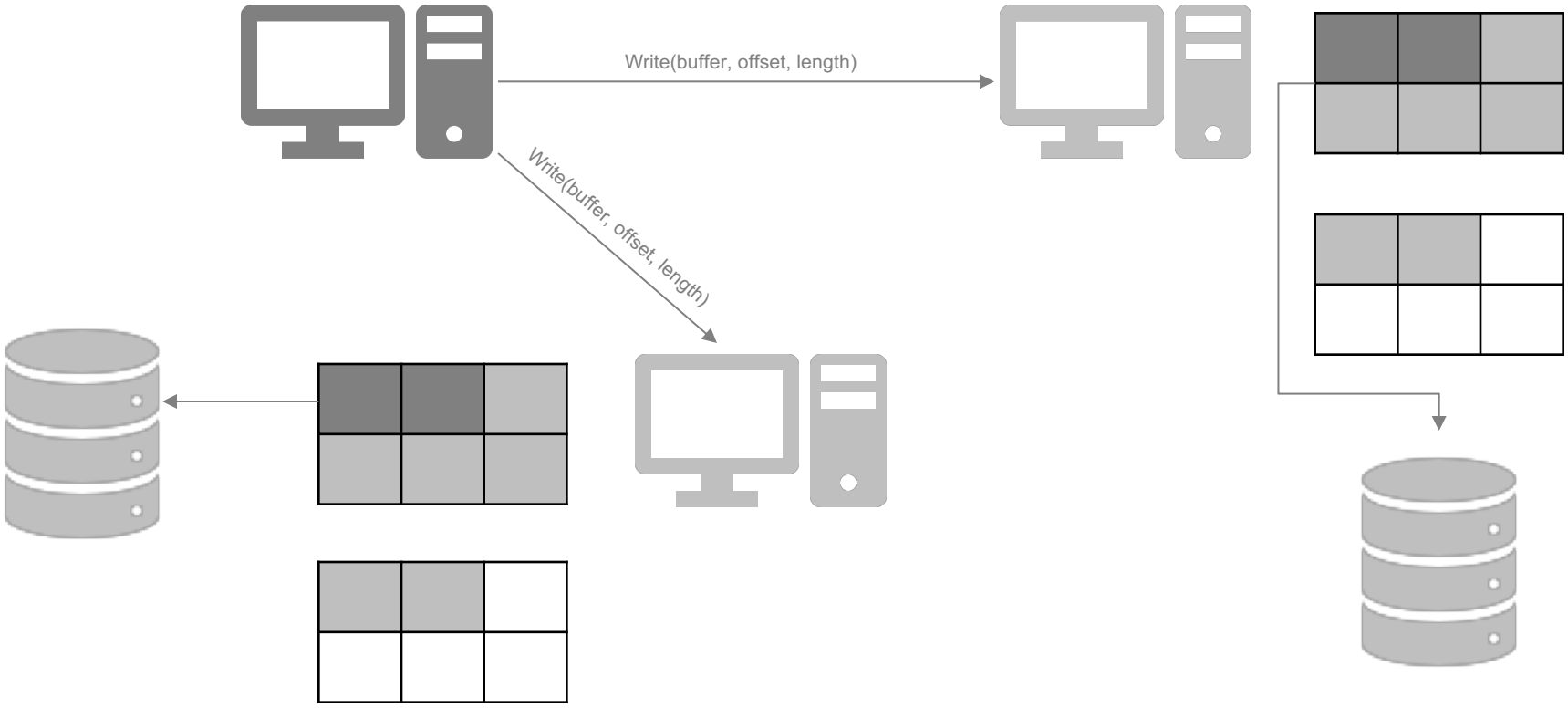
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



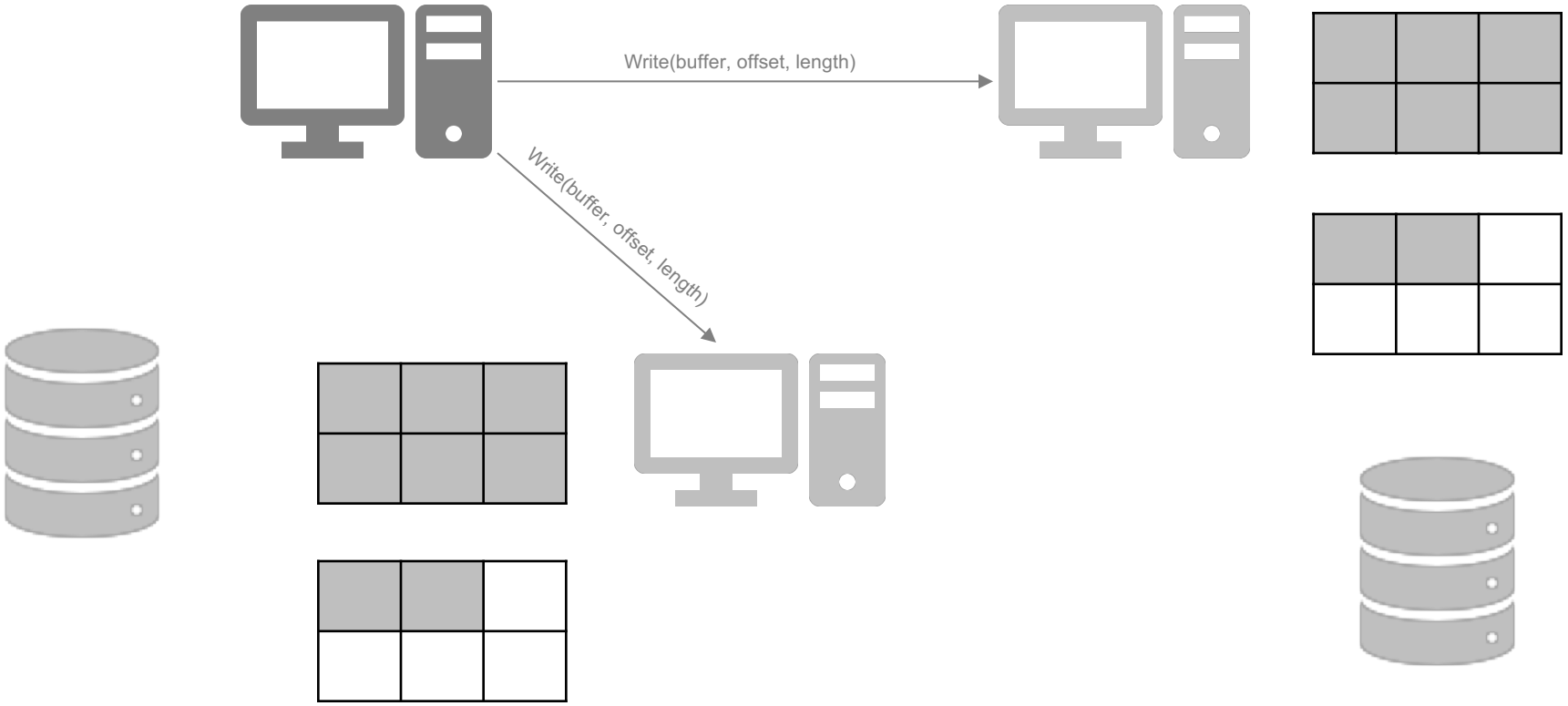
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



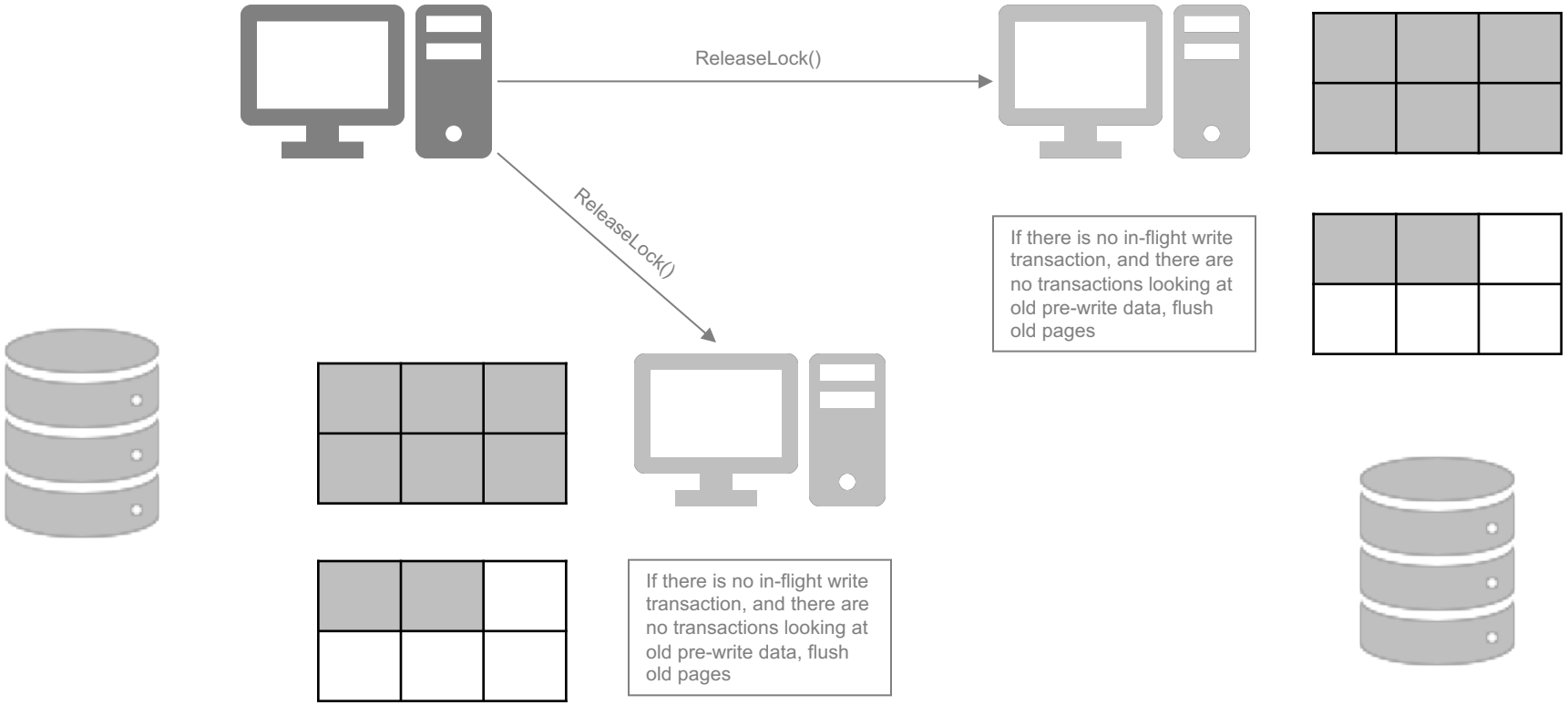
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



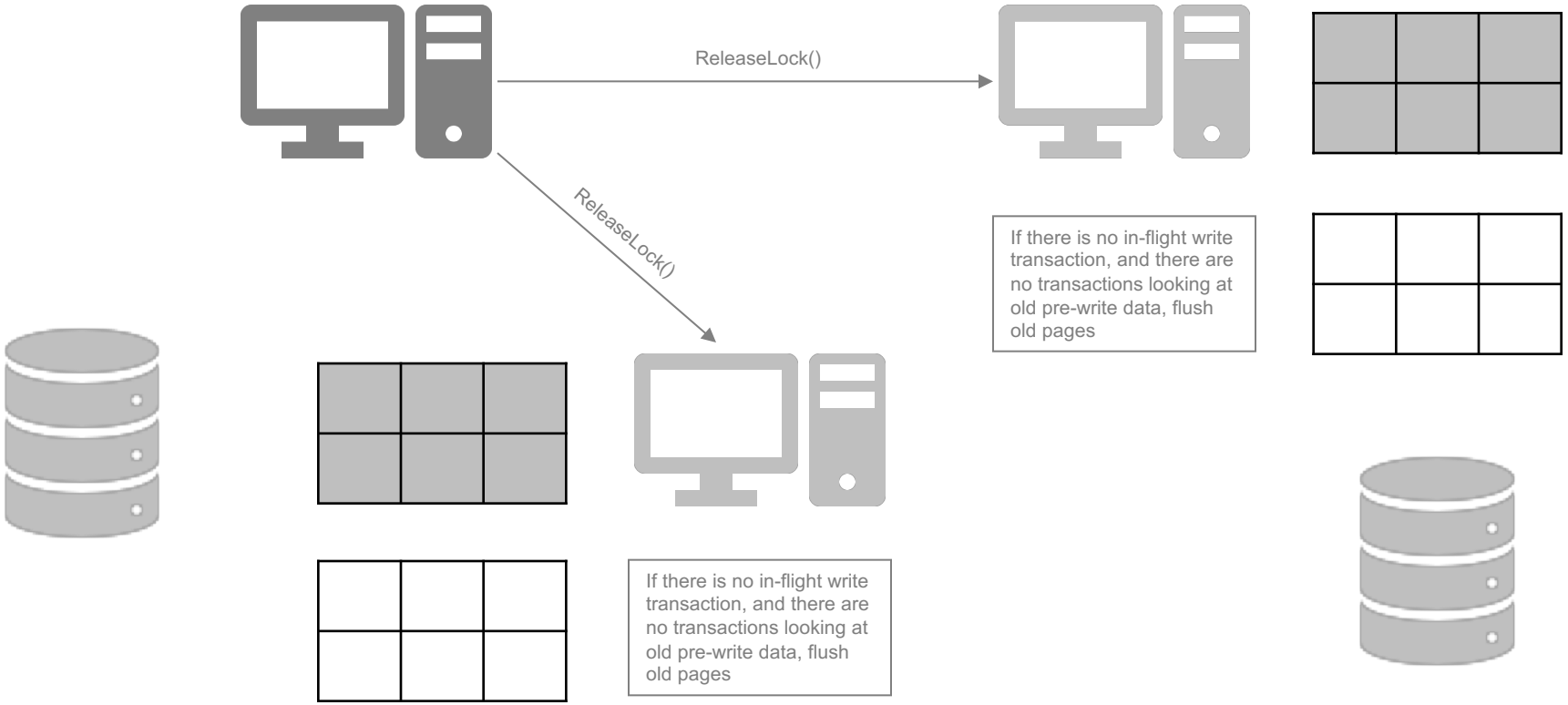
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



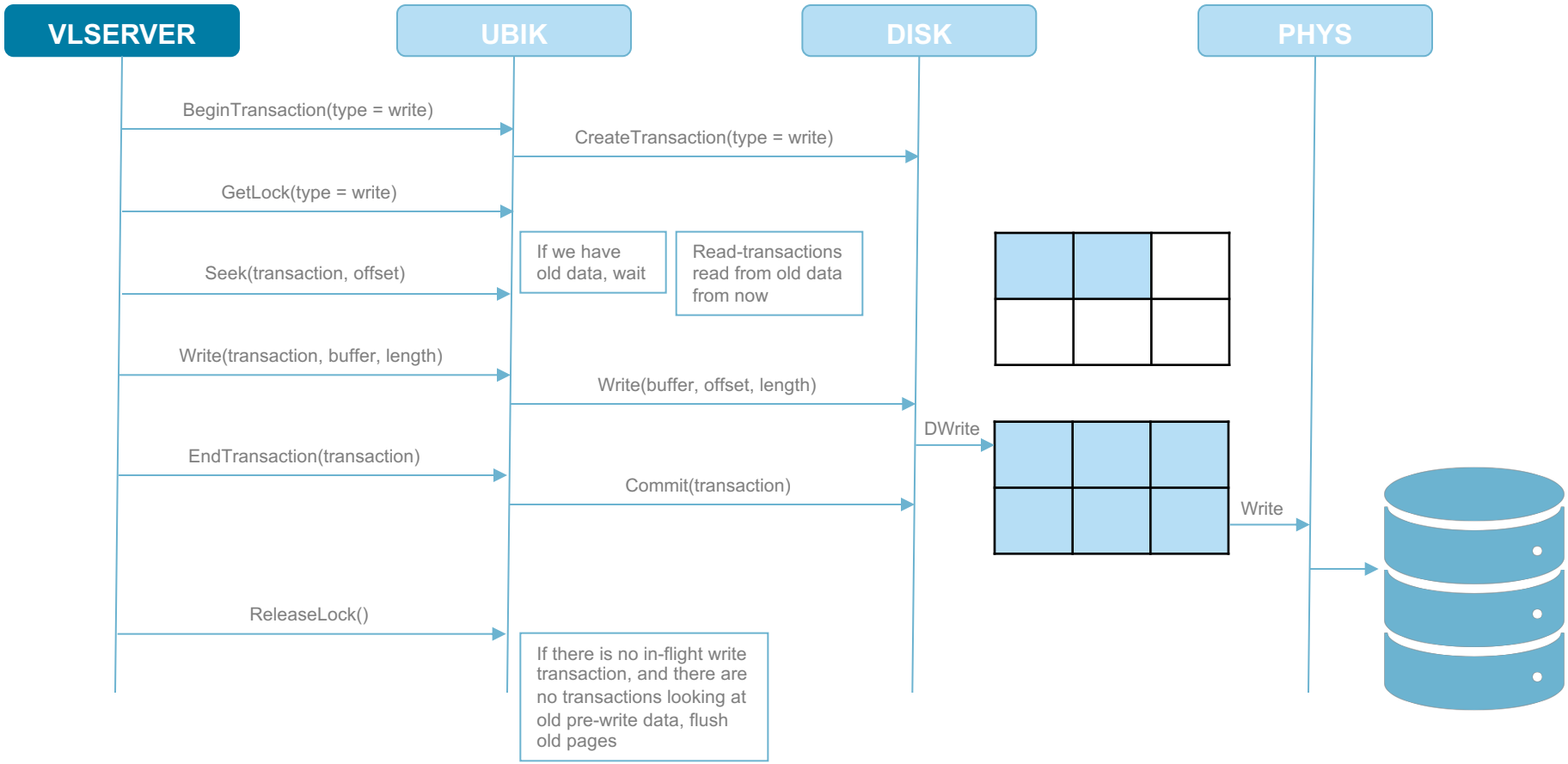
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



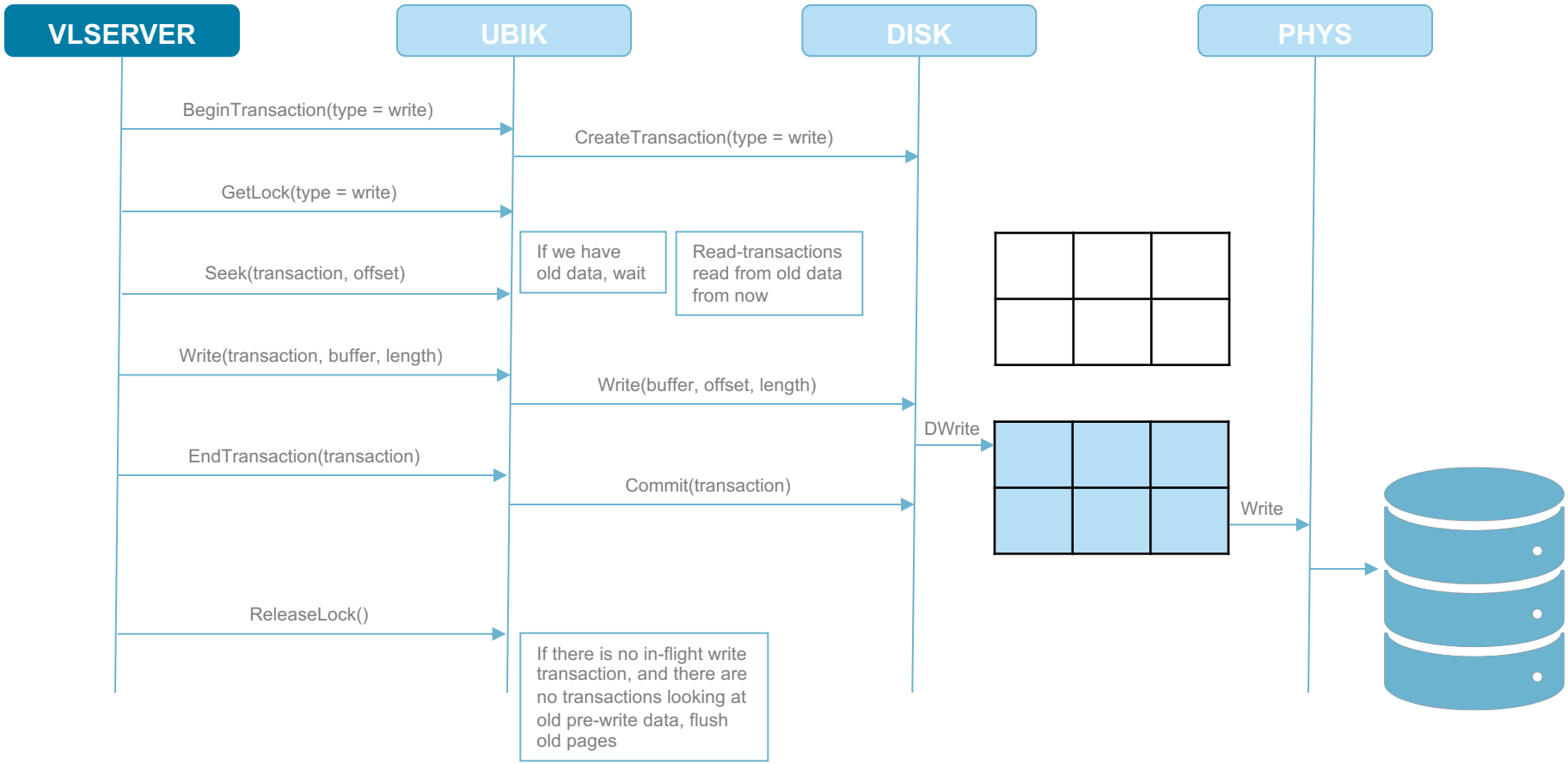
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



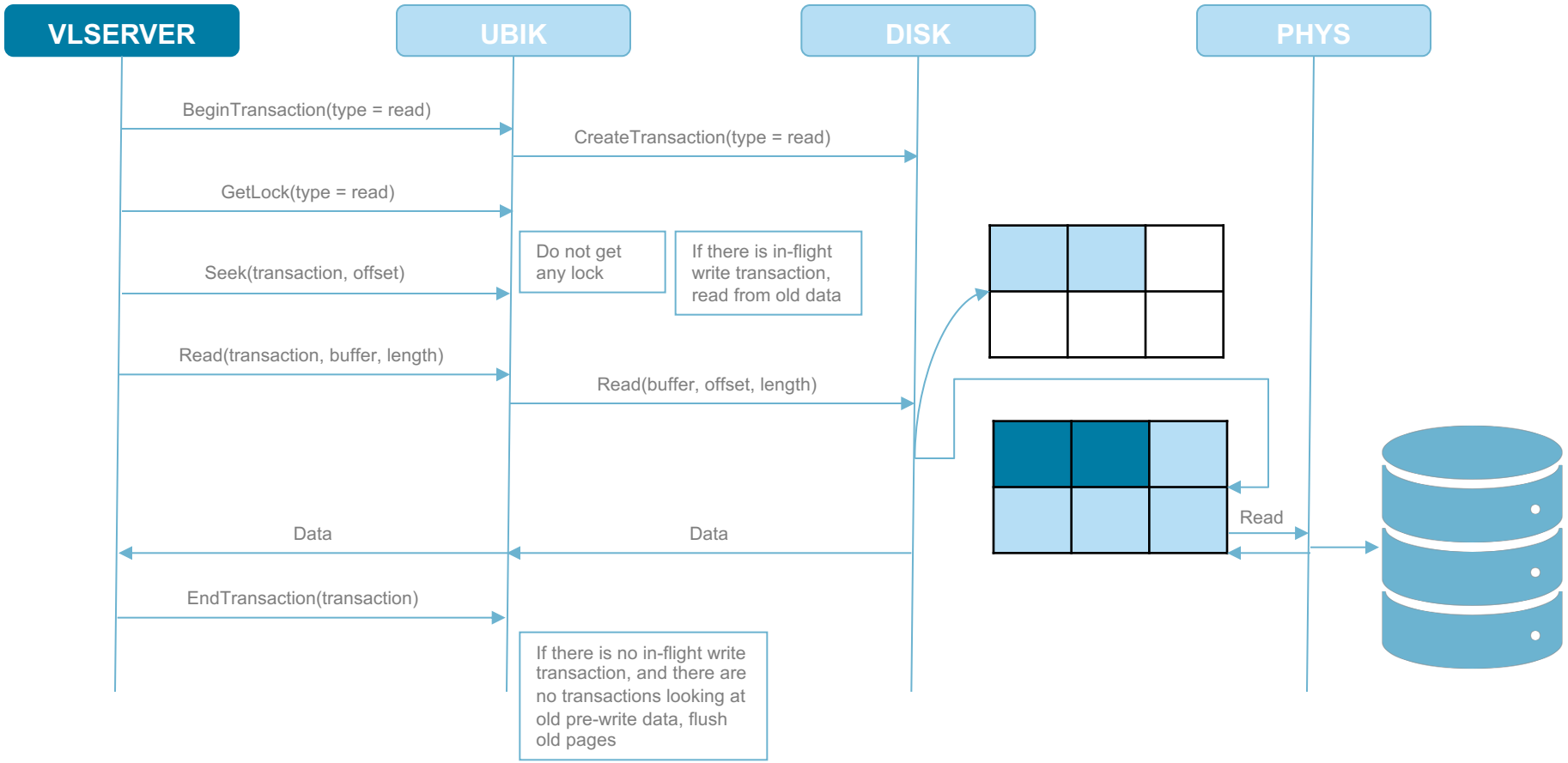
# WRITE-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



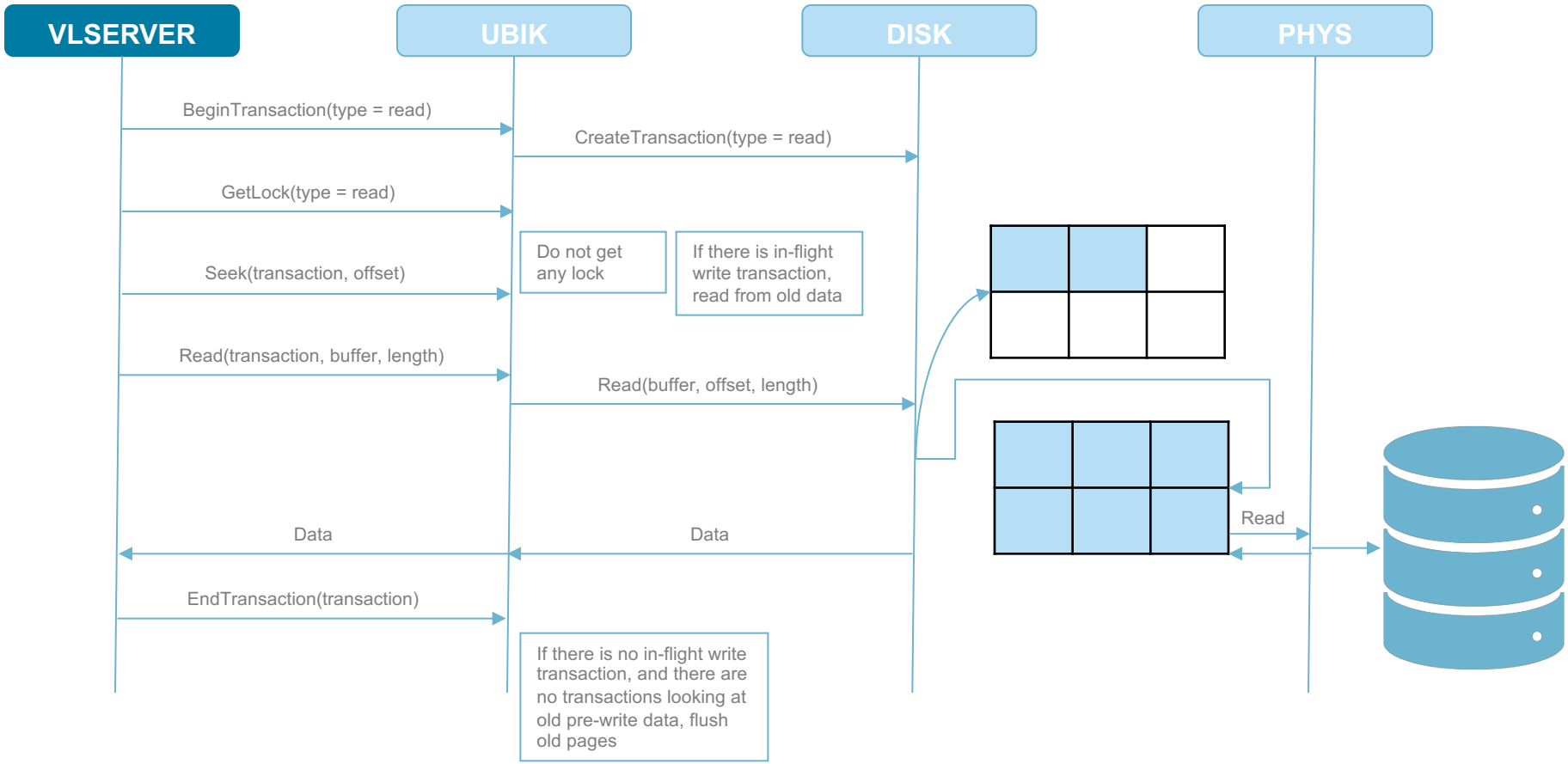
# READ-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



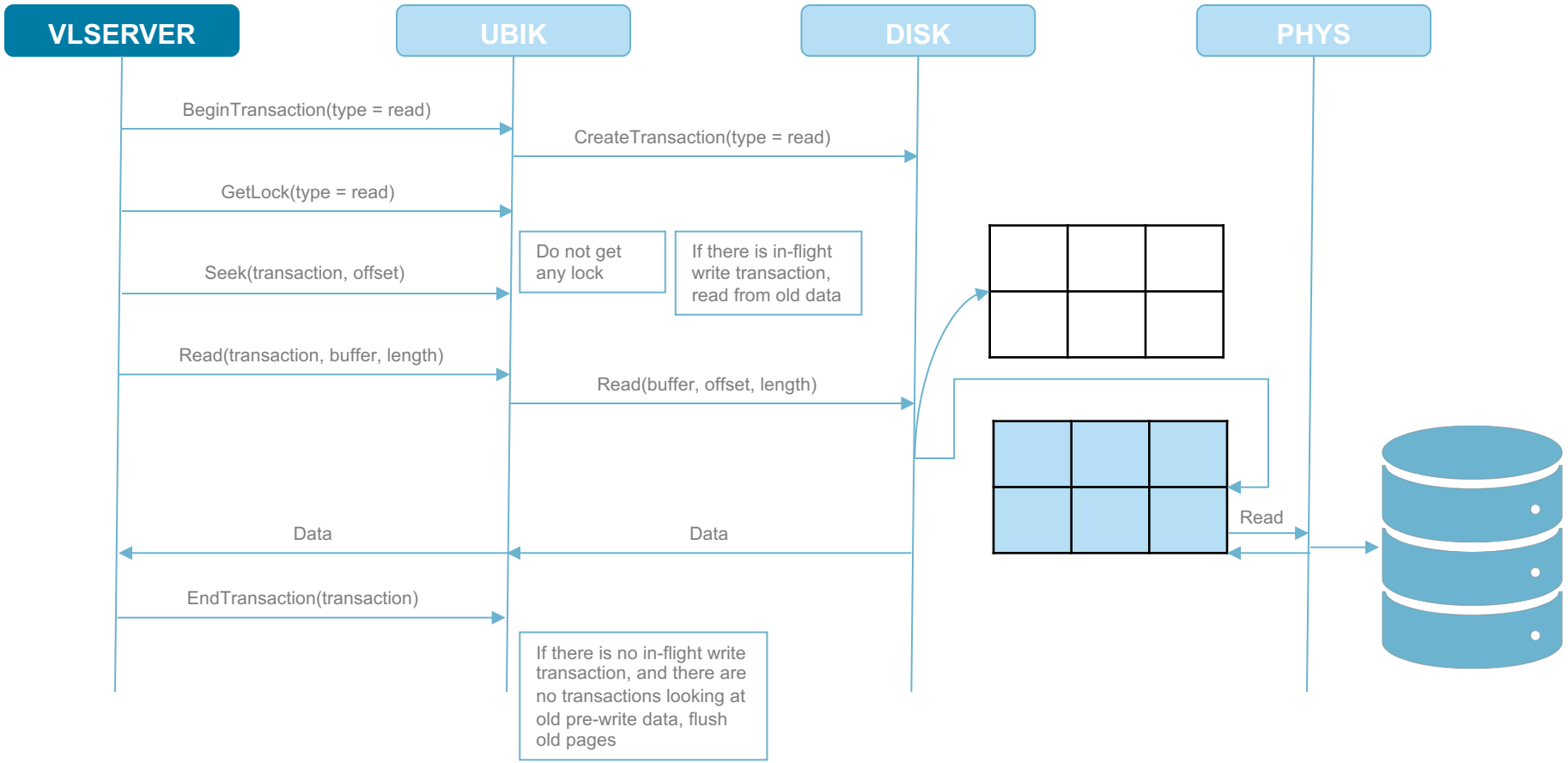
# READ-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



# READ-TRANSACTIONS (SIMPLIFICATION)

SINE NOMINE ASSOCIATES



# READS-DURING-COMMIT

SINE NOMINE ASSOCIATES

- Write-transactions do not block read-transactions;
- Read-transactions do not block write-transactions;
- Limitations
  - Can not have multiple write-transactions running at the same time;
  - New write-transactions are blocked if we still have any read-transaction looking at old data;

# PATCHES

SINE NOMINE ASSOCIATES

- Patches can be found on gerrit;
- Reads-during-recovery;
  - Topic: ubik-reads-during-recovery;
- Reads-during-commit;
  - Topic: ubik/read-during-commit;

SINE NOMINE
ASSOCIATES

# OTHER FIXES

SINE NOMINE
ASSOCIATES

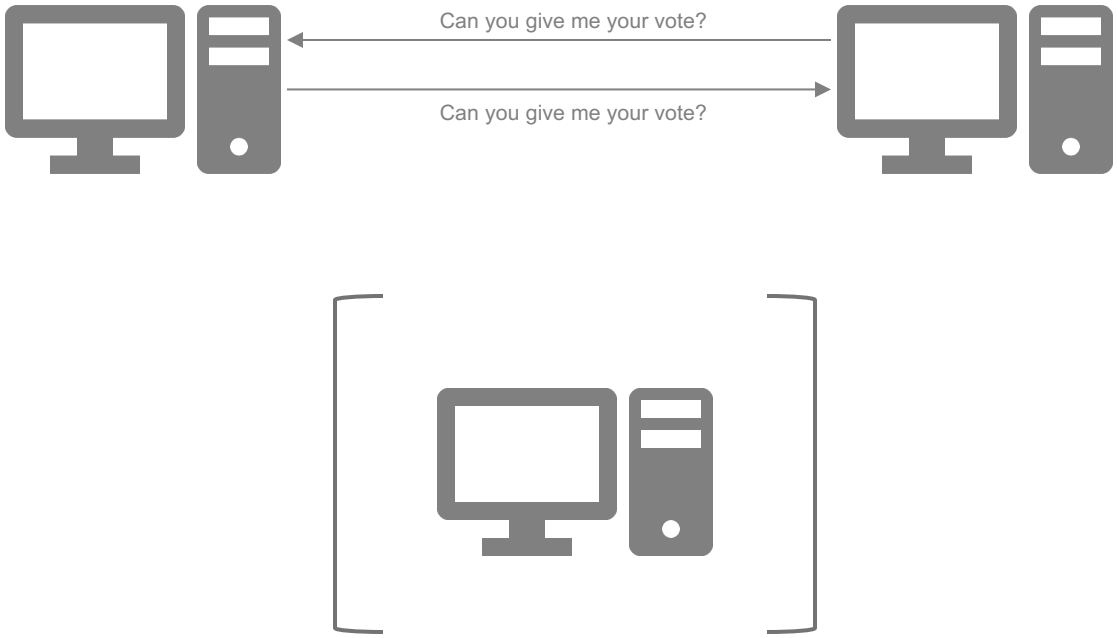
# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

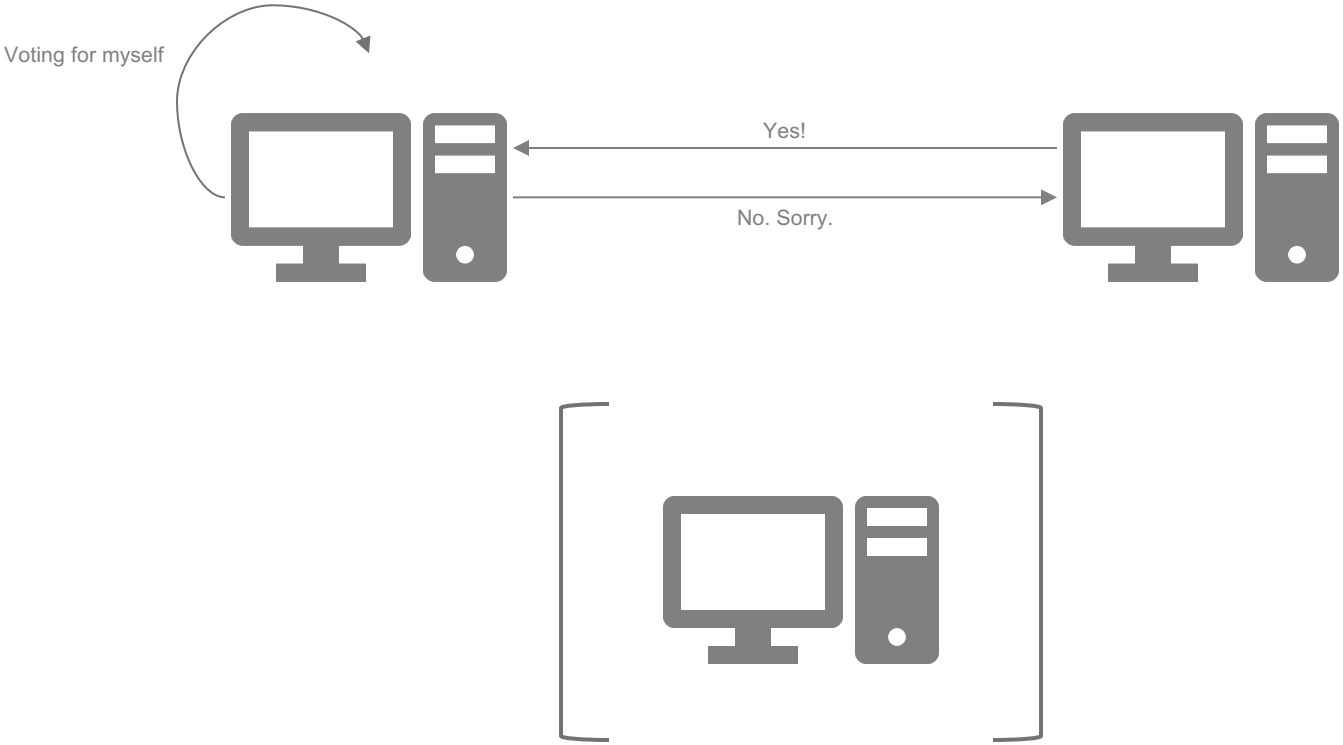
SINE NOMINE ASSOCIATES

- Remote-sites do not create write-transactions if it didn't vote for the sync-site;
- What happens when the request for a new transaction is refused?
  - Sync-site assumes that the remote-site is not "available";

# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

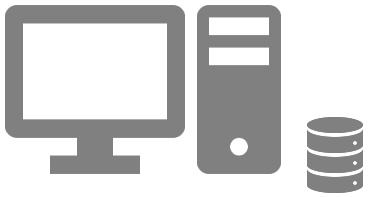


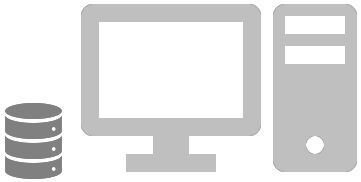
# AVOIDING SITES THAT DID NOT VOTE FOR SYNC



# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

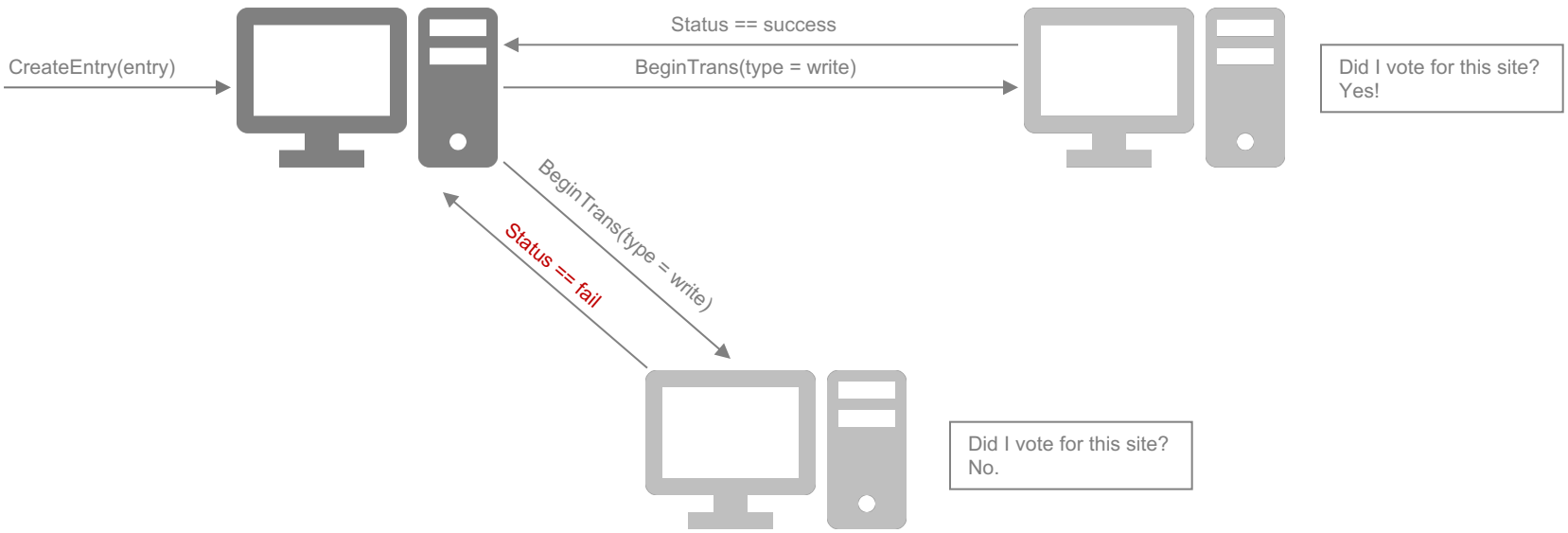
SINE NOMINE ASSOCIATES





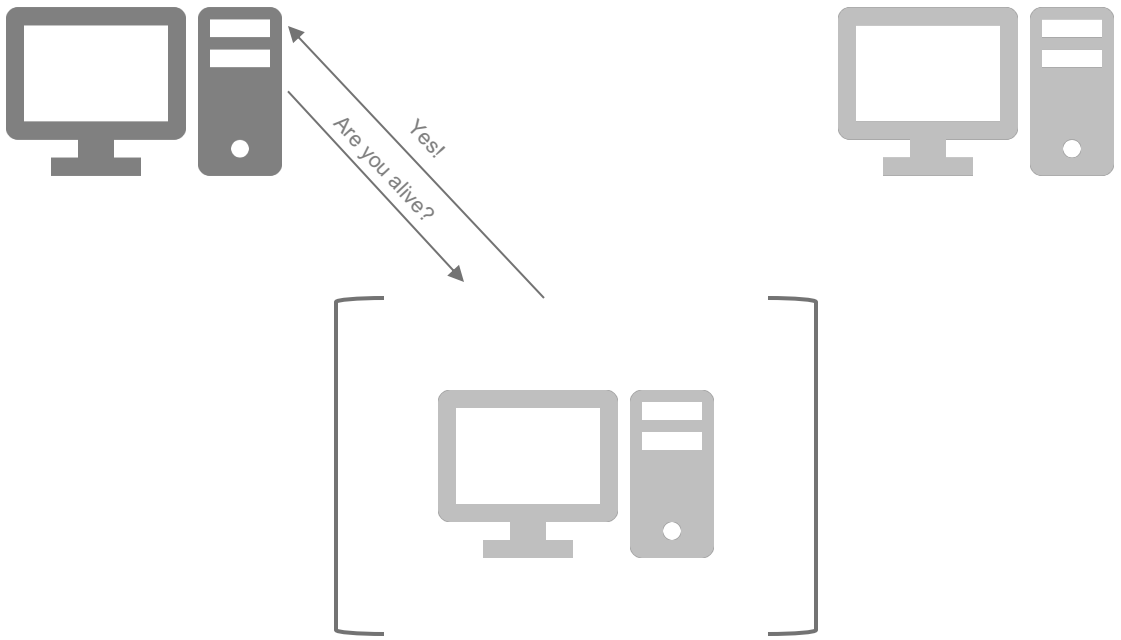
# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

SINE NOMINE ASSOCIATES



# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

SINE NOMINE ASSOCIATES

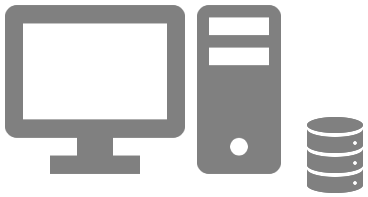


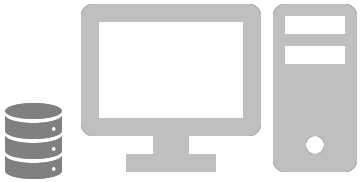
# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

SINE NOMINE ASSOCIATES

# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

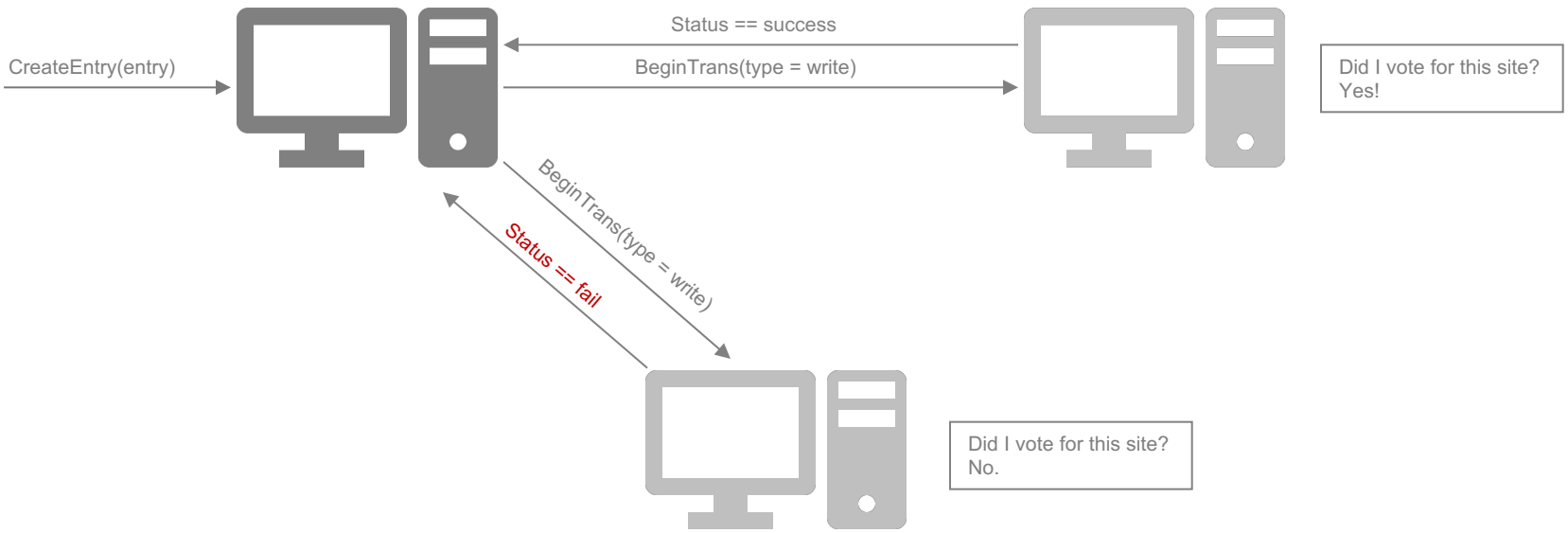
SINE NOMINE ASSOCIATES



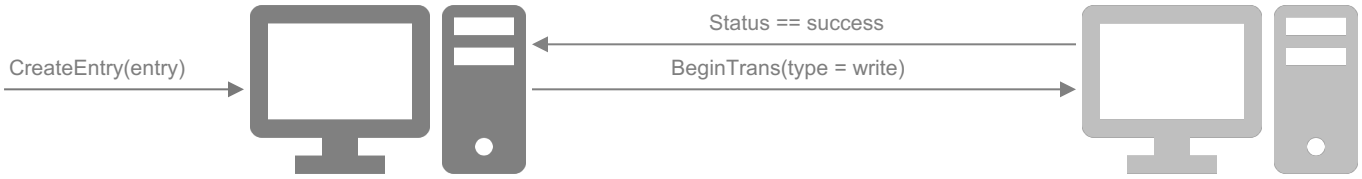


# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

SINE NOMINE ASSOCIATES

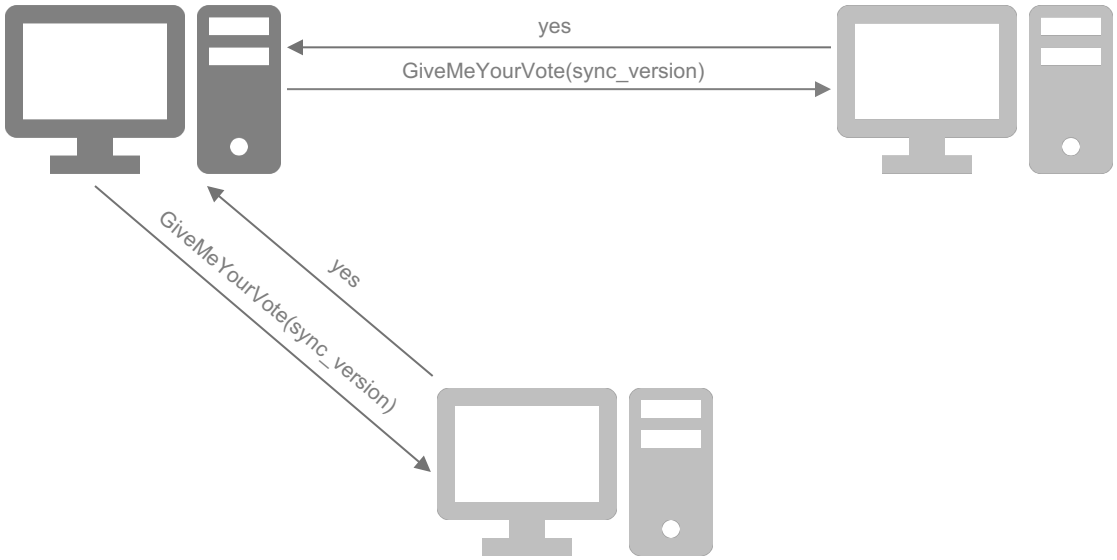


# AVOIDING SITES THAT DID NOT VOTE FOR SYNC



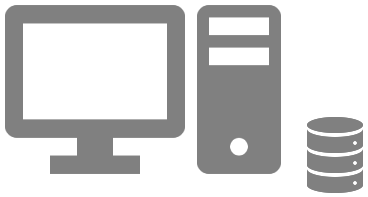

# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

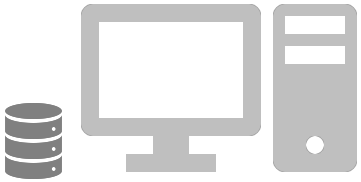
SINE NOMINE ASSOCIATES



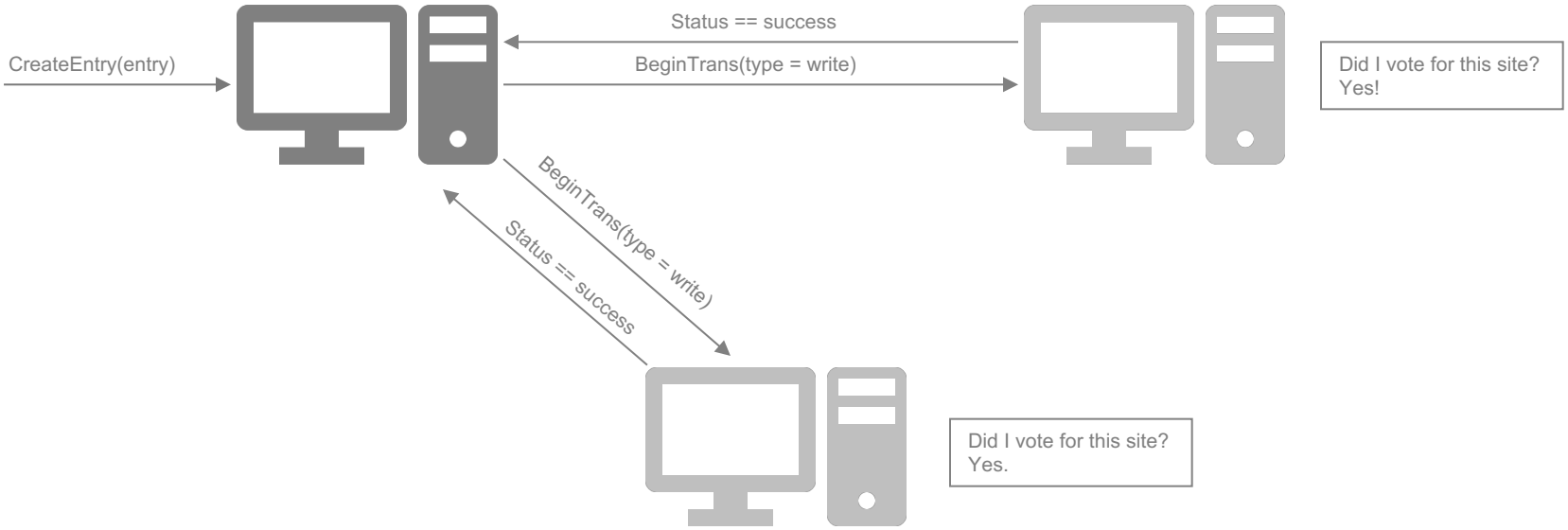
# AVOIDING SITES THAT DID NOT VOTE FOR SYNC

SINE NOMINE ASSOCIATES





# AVOIDING SITES THAT DID NOT VOTE FOR SYNC



SINE NOMINE
ASSOCIATES

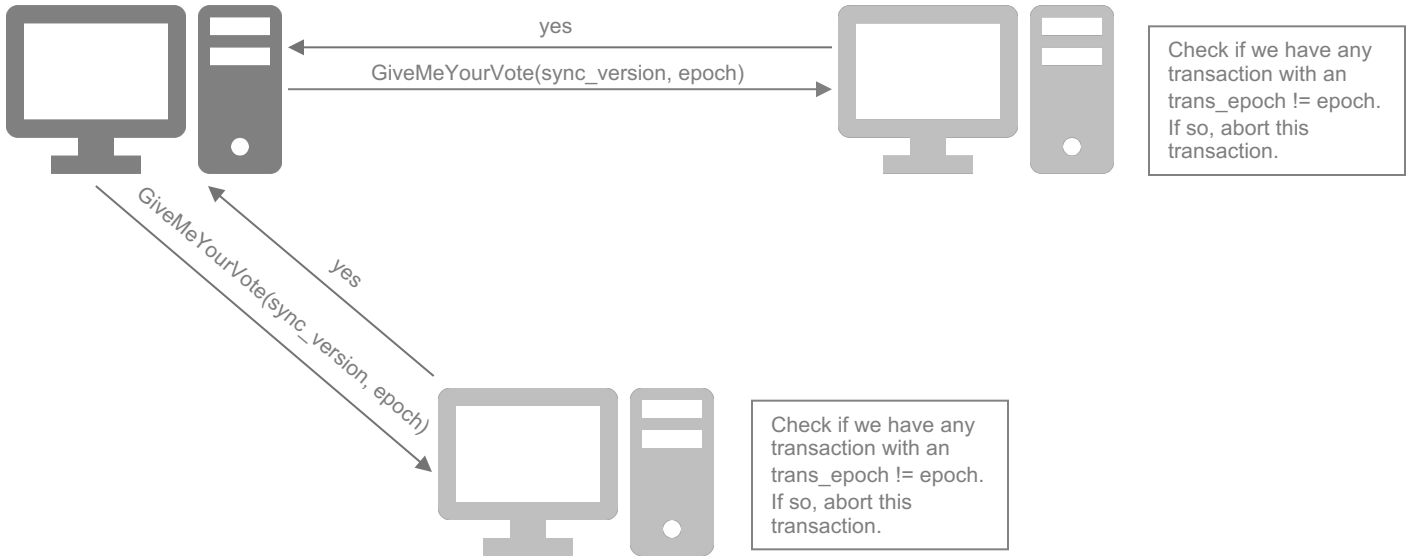
# UPDATE EPOCH AS SOON AS SYNC-SITE IS ELECTED

# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

SINE NOMINE ASSOCIATES

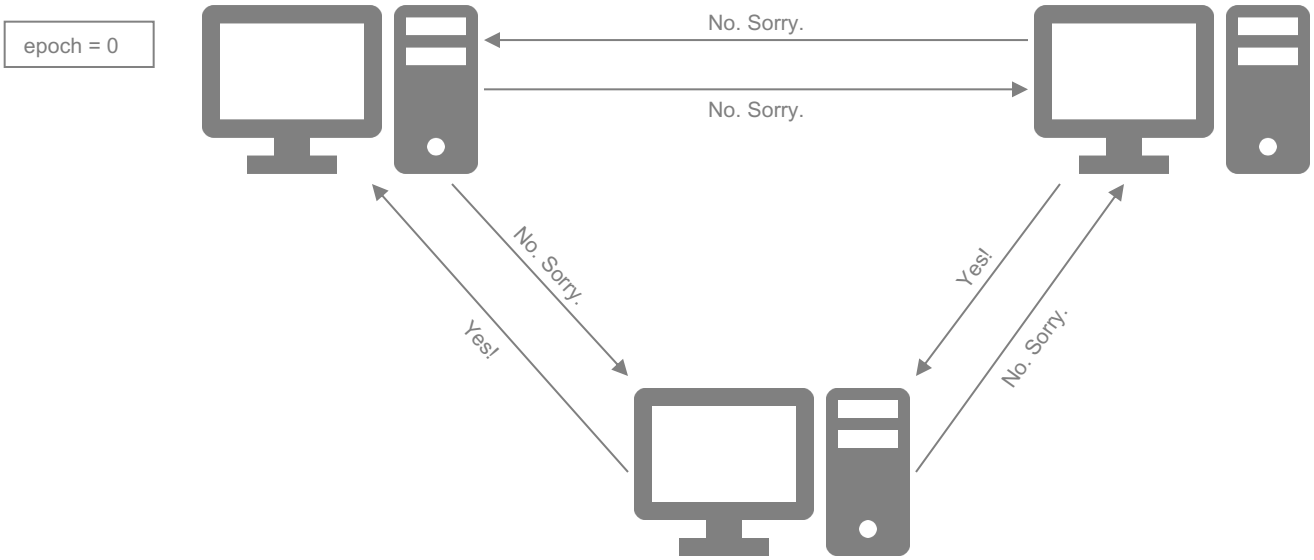
- Epoch is global that represents the time in which the sync-site was elected;
- We use this global to differentiate transactions from different mandates;
  - Every transaction has an epoch;
  - If this epoch is not equal to the epoch advertised by the current sync-site, this transaction should be aborted;

# UPDATE EPOCH AS SOON AS SYNC IS ELECTED



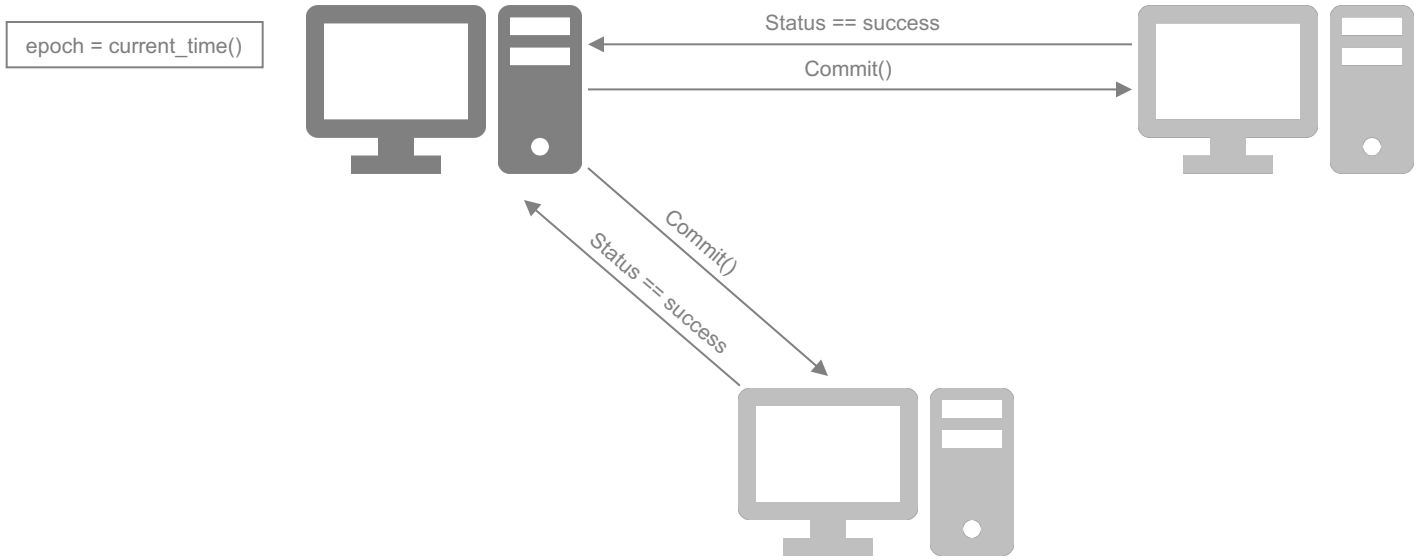
# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

SINE NOMINE ASSOCIATES

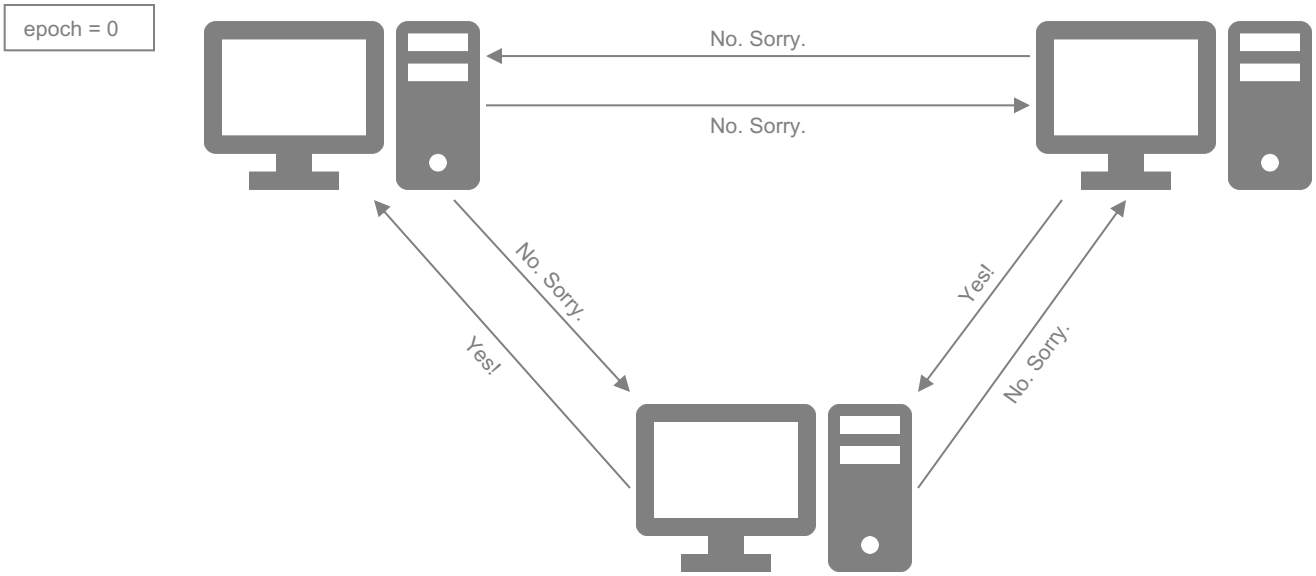


# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

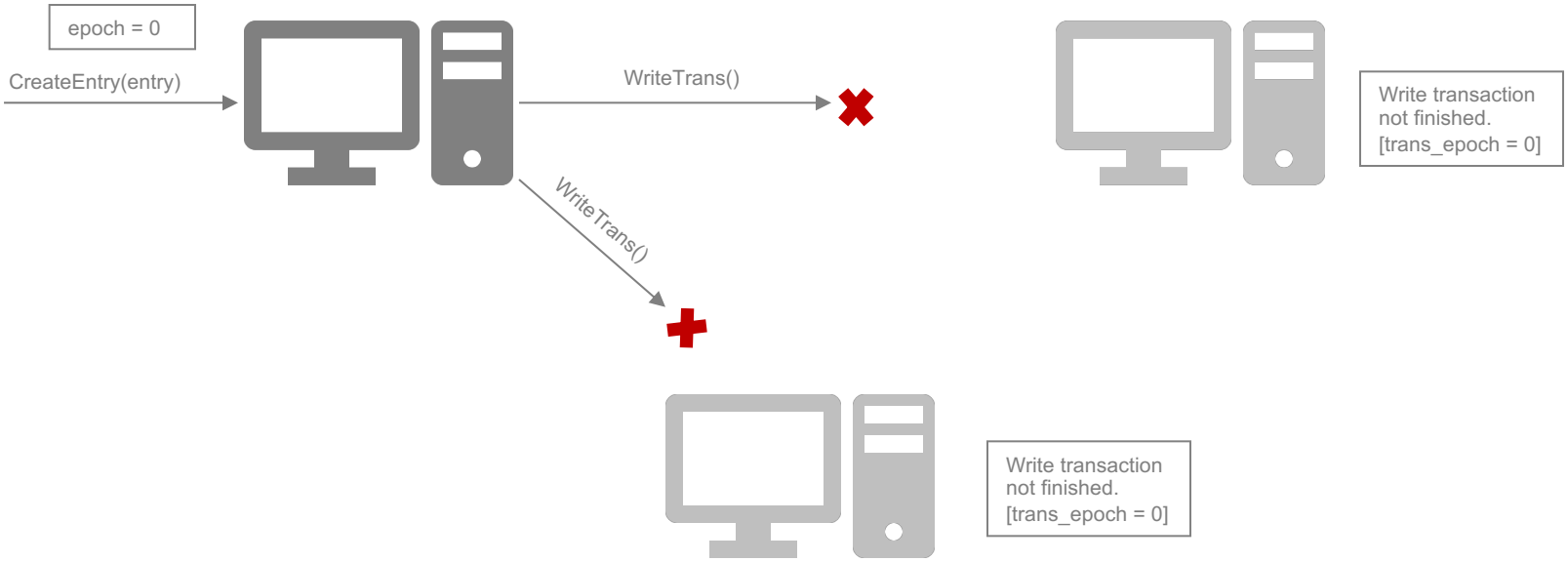
SINE NOMINE ASSOCIATES



# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

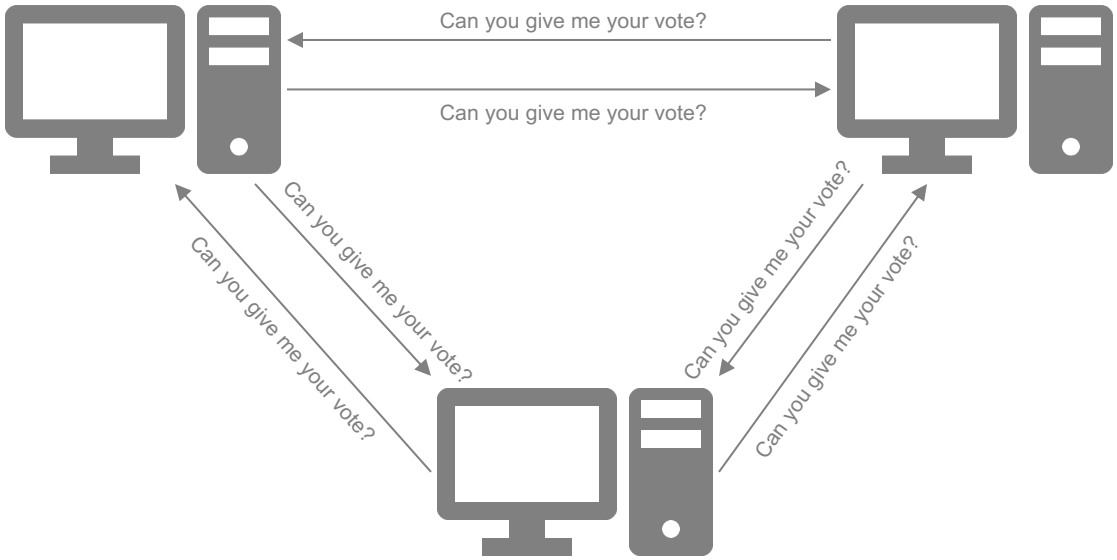


# UPDATE EPOCH AS SOON AS SYNC IS ELECTED



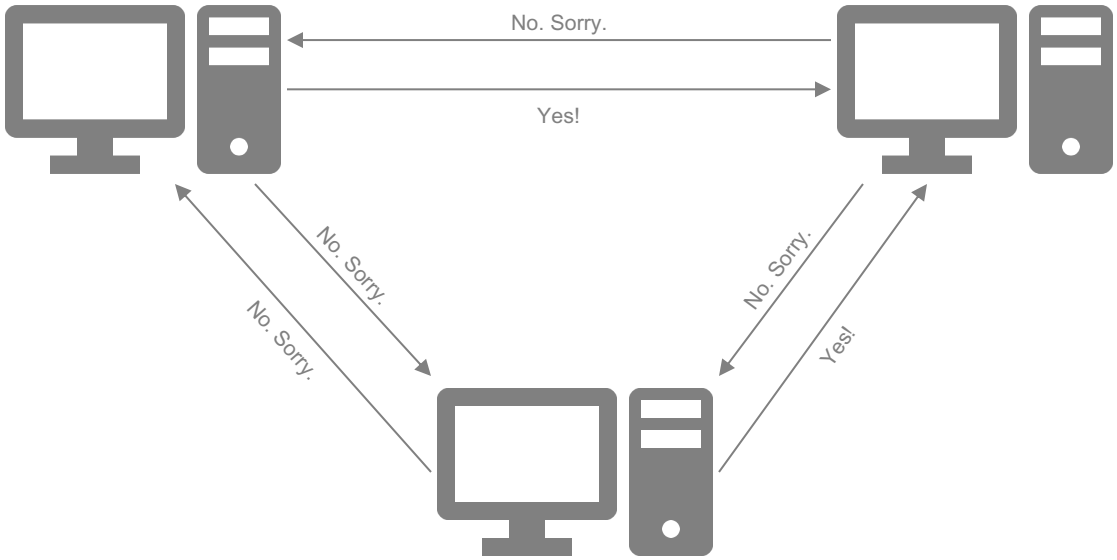
# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

SINE NOMINE ASSOCIATES

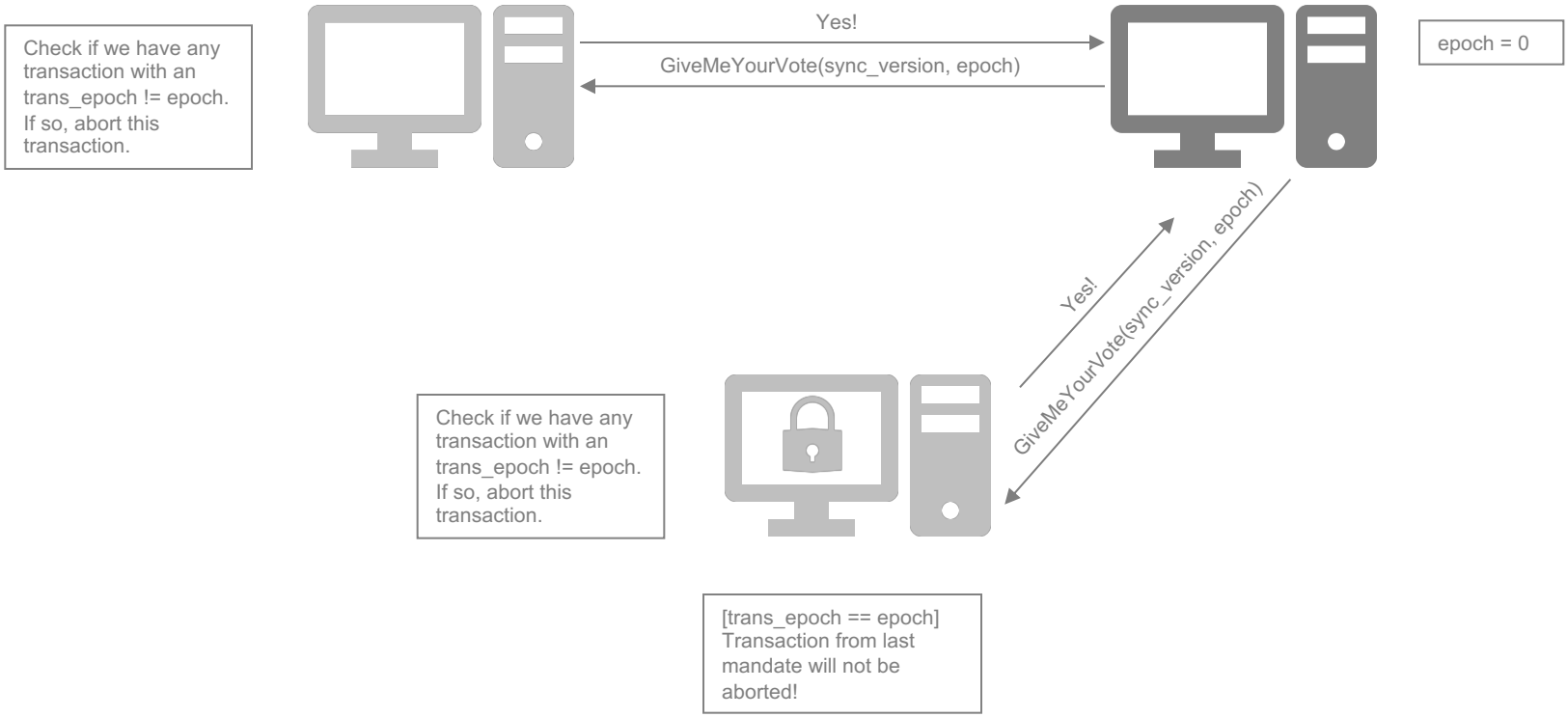


# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

SINE NOMINE ASSOCIATES

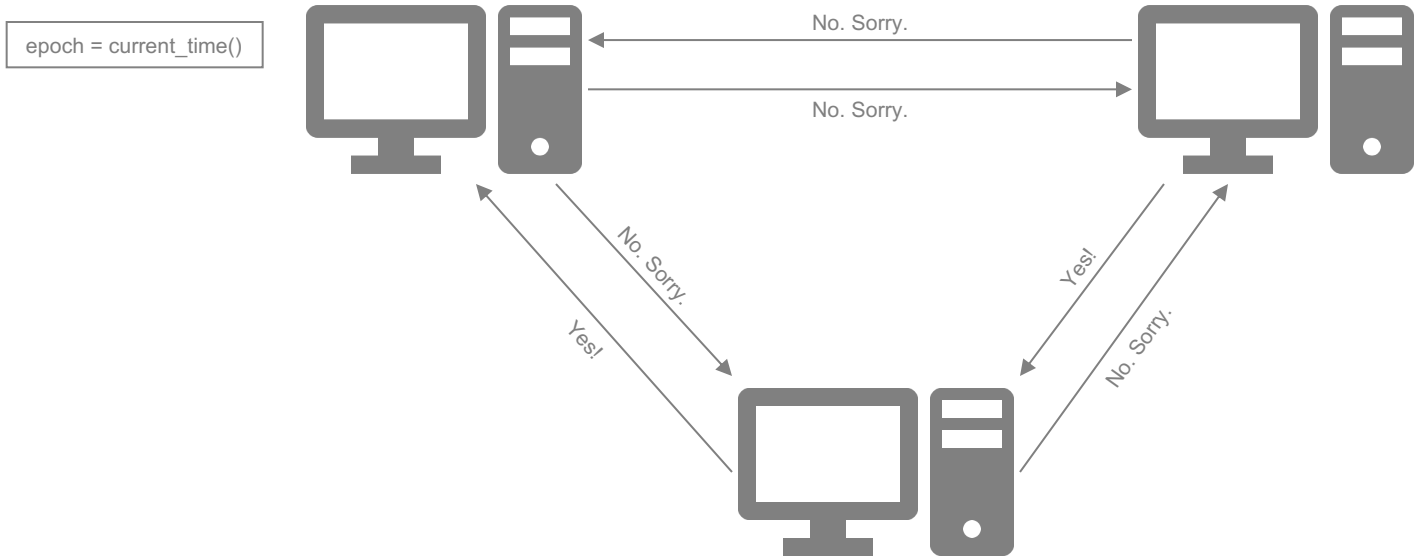


# UPDATE EPOCH AS SOON AS SYNC IS ELECTED



# UPDATE EPOCH AS SOON AS SYNC IS ELECTED

SINE NOMINE ASSOCIATES




Thank you!