

PCC POSTGRESCONF
CN 2021 | **PGConf.Asia** 12.14-17

A Middleware to Synchronize Multiple Instances in a PostgreSQL Cluster

SRA OSS, Inc. Japan
Takeshi MISHIMA

<https://2021.postgresconf.cn>

About me

- I had been researching database systems at NTT and Keio University in Japan.
- My representative papers were accepted by VLDB2009 and SIGMOD 2015.
- I received Ph.D. degree from The University of Tokyo.
- Now, I work for SRA OSS, Inc. Japan.

CONTENT



1. What is Pangea?



2. The disadvantage of streaming replication



3. The algorithms of Pangea



4. Summary



PART 1

What is Pangea?

What is Pangea?

- Proposed Pangea and accepted by VLDB2009.

A synchronization middleware for multiple Postgres instances without modifying the source codes of Postgres.

- Pangea is a just prototype but the algorithms have been implemented into the latest version of Pgpool-II with the snapshot isolation mode.
- In today's my presentation, I will explain Pangea, and enjoy the fruit by using Pgpool-II if you are interested in Pangea !

The advantages of Pangea

- The advantages

(1) not only it increases availability but also,

(2) any instance returns the latest consistent data.

cf. Streaming replication provides (1) but does not (2).

Next, I will explain the detail of it.



PART 2

The disadvantage of streaming replication

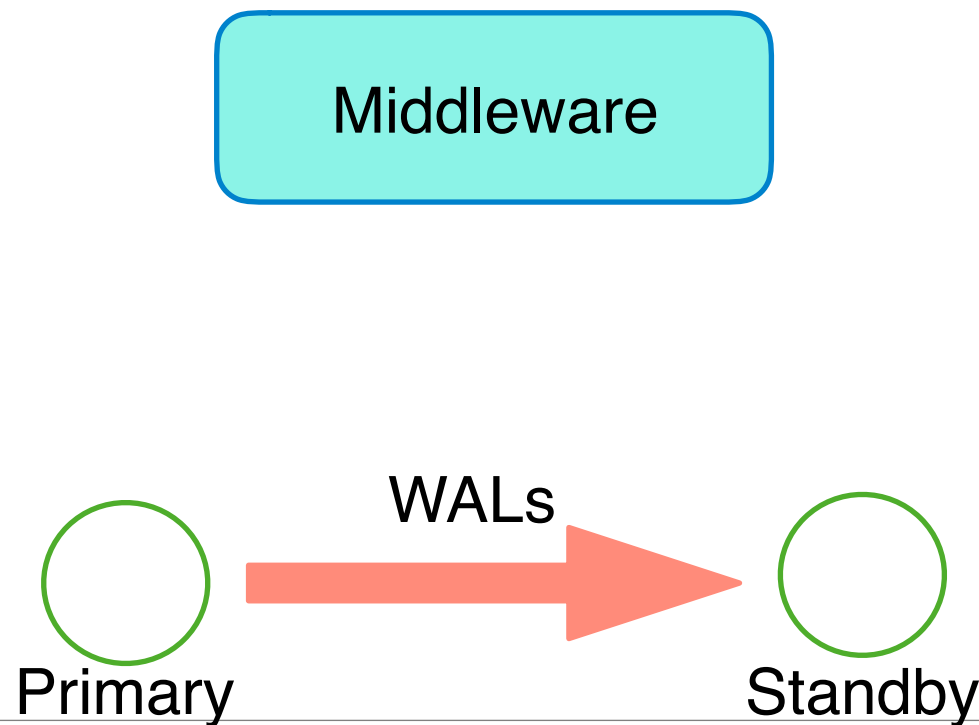
The disadvantage of streaming replication

- Standby instance cannot return the latest data.

ex.

T0:

```
BEGIN;  
UPDATE test_table;  
SELECT * FROM test_table;  
COMMIT;
```



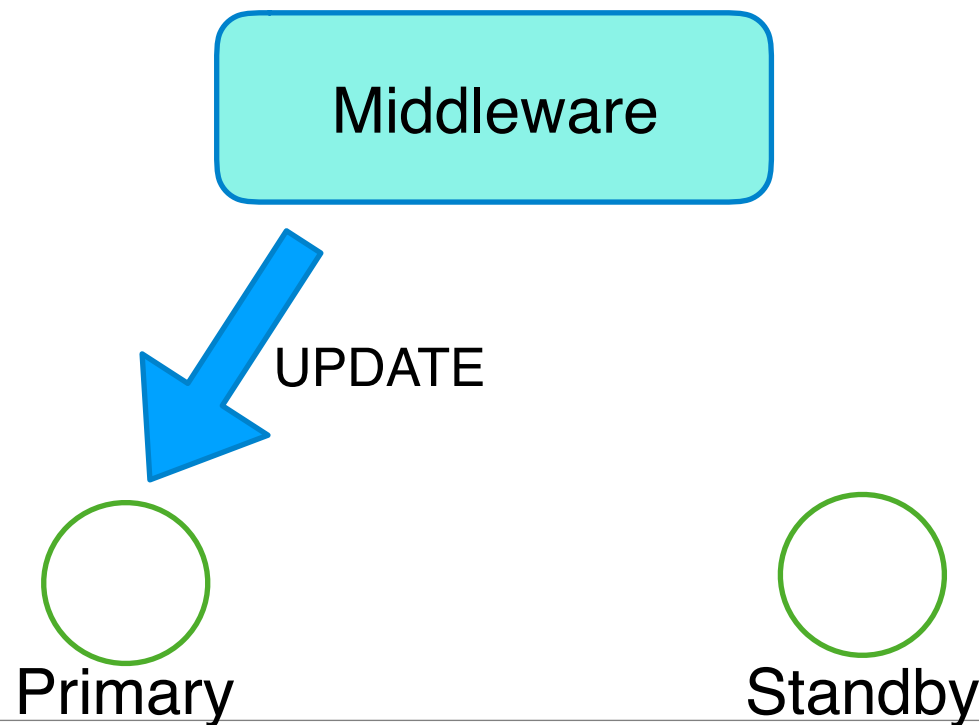
The disadvantage of streaming replication

- Standby instance cannot return the latest data.

ex.

T0:

```
BEGIN;  
UPDATE test_table;  
SELECT * FROM test_table;  
COMMIT;
```



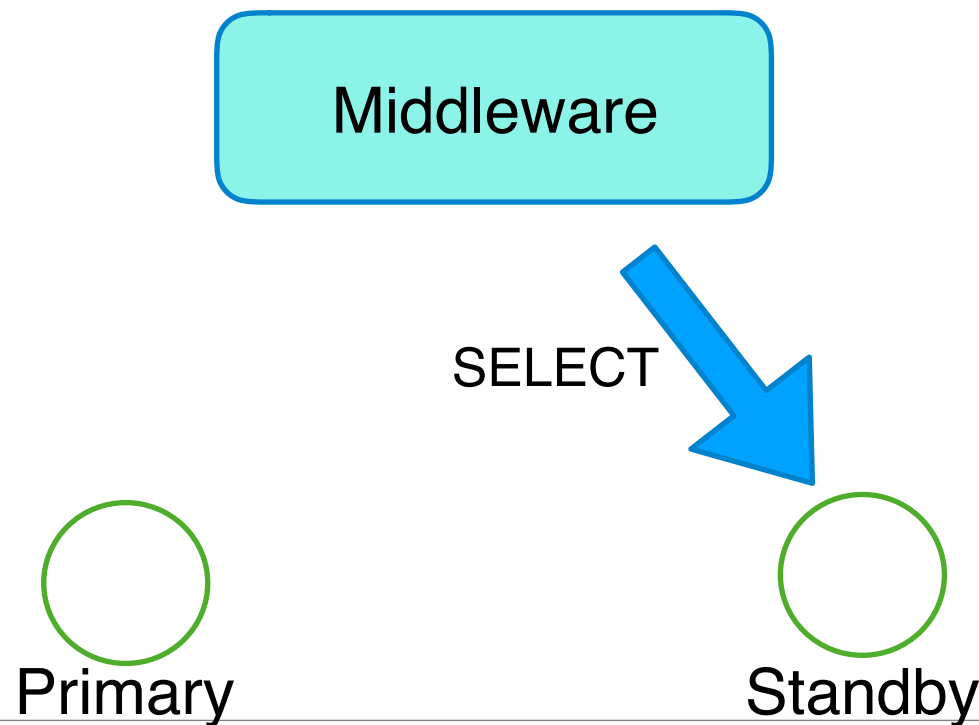
The disadvantage of streaming replication

- Standby instance cannot return the latest data.

ex.

T0:

```
BEGIN;  
UPDATE test_table;  
SELECT * FROM test_table;  
COMMIT;
```



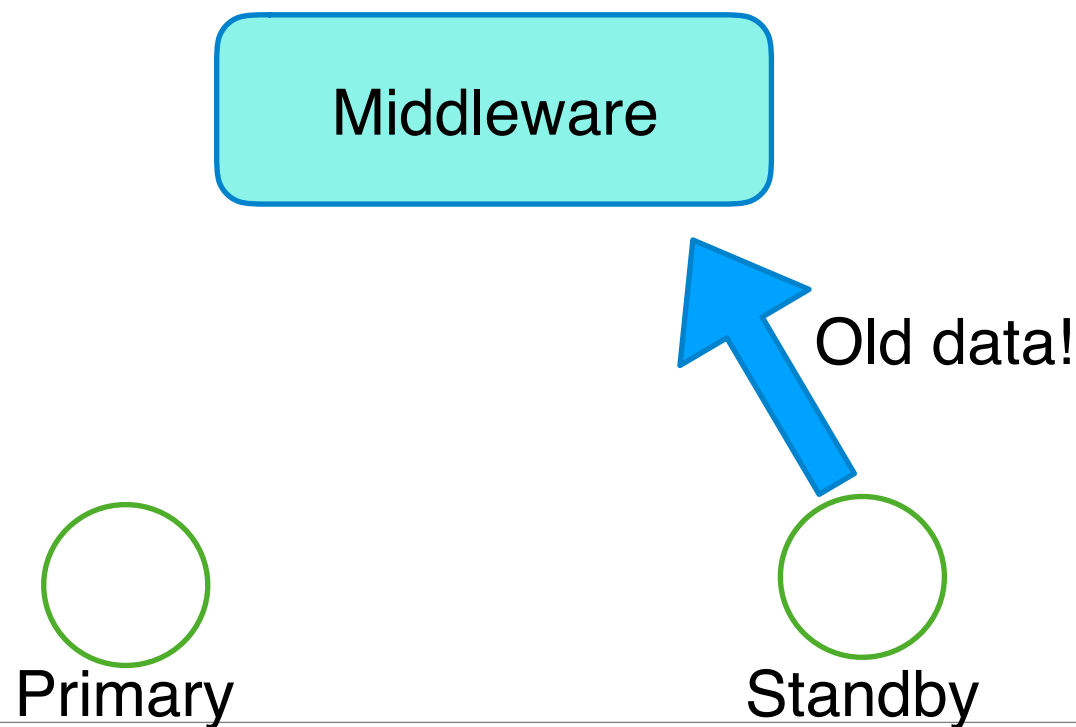
The disadvantage of streaming replication

- Standby instance cannot return the latest data.

ex.

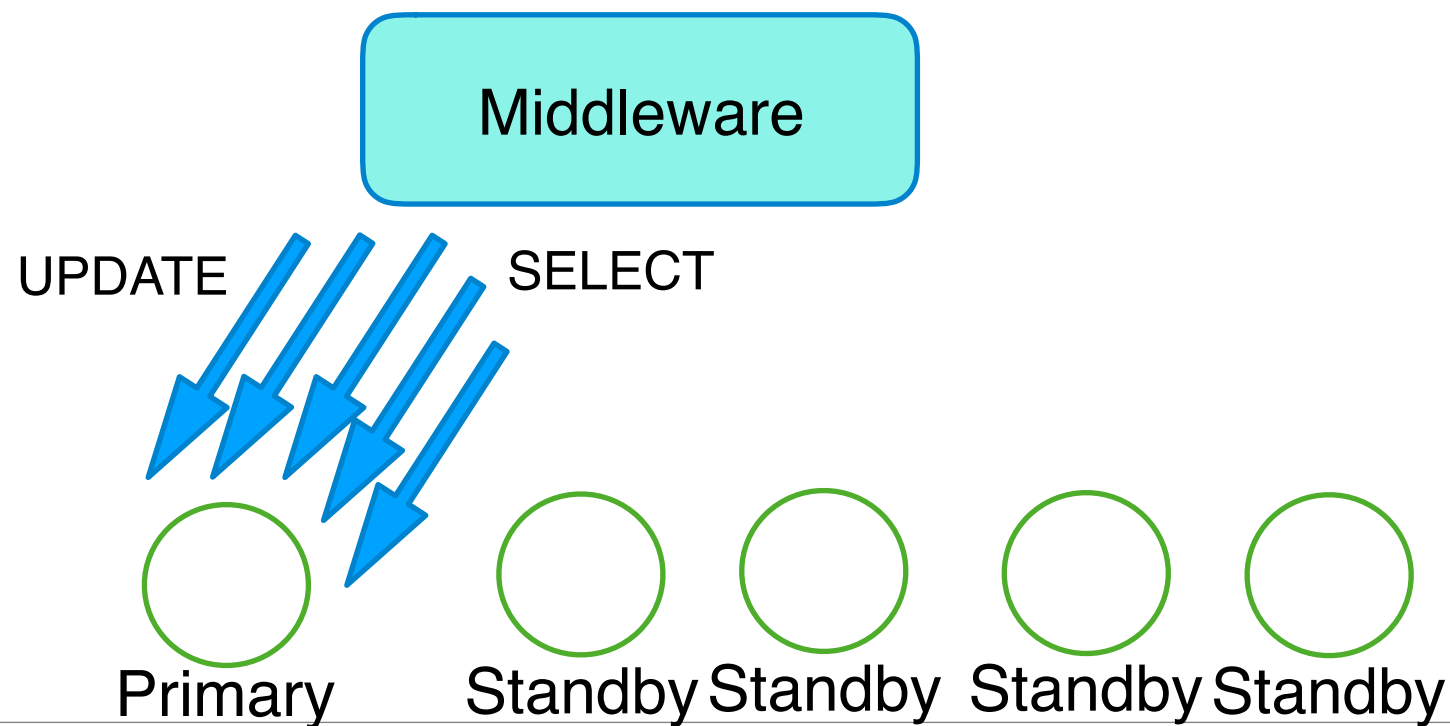
T0:

```
BEGIN;  
UPDATE test_table;  
SELECT * FROM test_table;  
COMMIT;
```



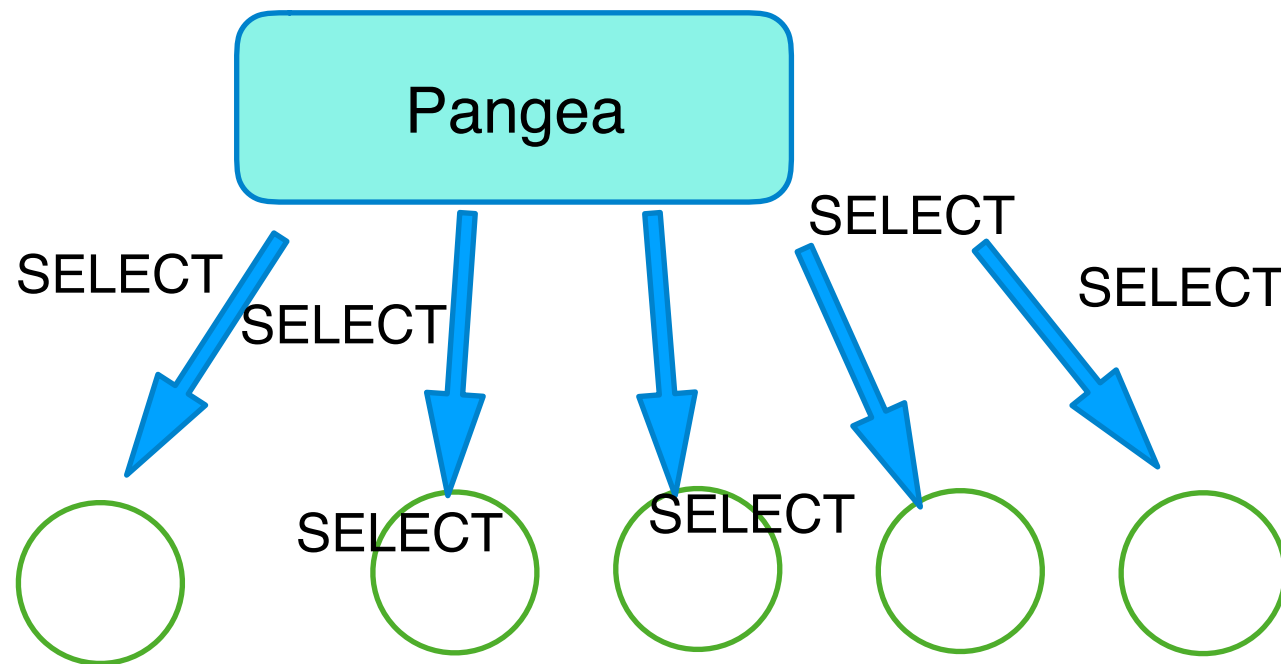
The disadvantage of streaming replication

- Even if we have a lot of standby instances, all queries must be sent to only the primary to ensure consistency.



The goal of Pangea

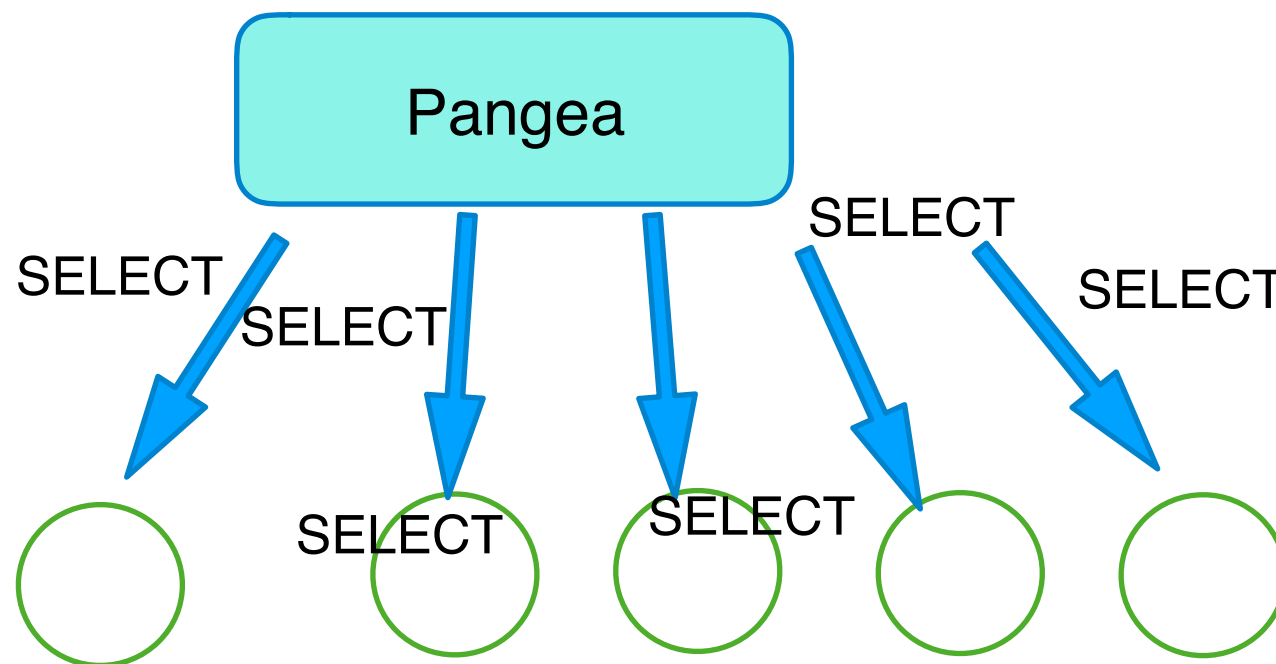
- Even if Pangea sends a SELECT query to any instance, it must return the latest consistent data.



The goal of Pangea

- Even if Pangea sends a SELECT query to any instance, it must return the latest consistent data.

Load-balancing without violating consistency!



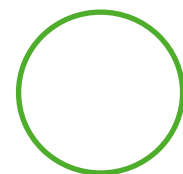


PART 3

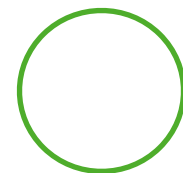
Pangea's algorithms to address the disadvantage

Assumption for Pangea

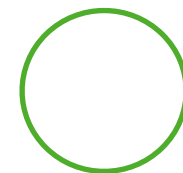
- Snapshot isolation (REPEATABLE READ isolation level)
- A query does not have any non-deterministic functions such as RANDOM and NOW.
- We call one replica *the leader* and the others *the followers*.
- We call the first DML query *a snapshot query*.



Leader



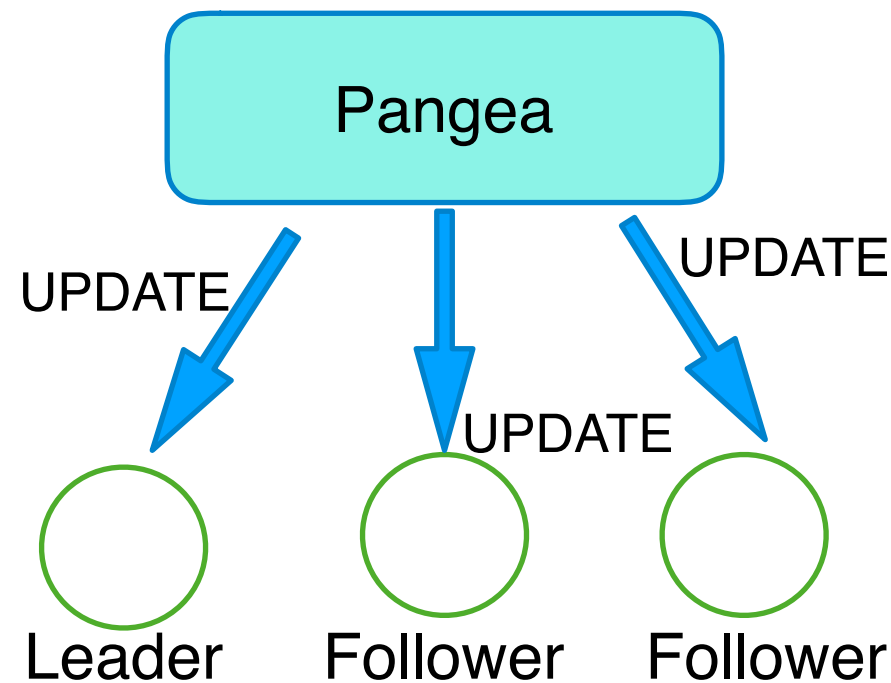
Follower



Follower

The replication method of Pangea

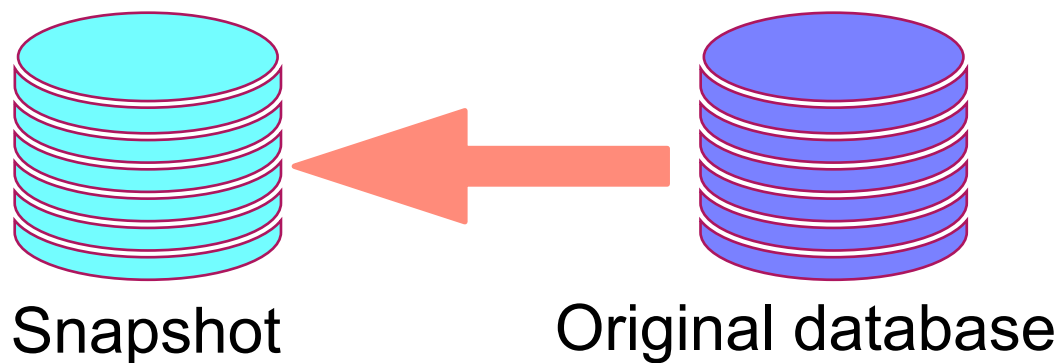
- Instead of using streaming replication, Pangea sends an UPDATE query to all instances.



Recap: snapshot isolation

Recap1

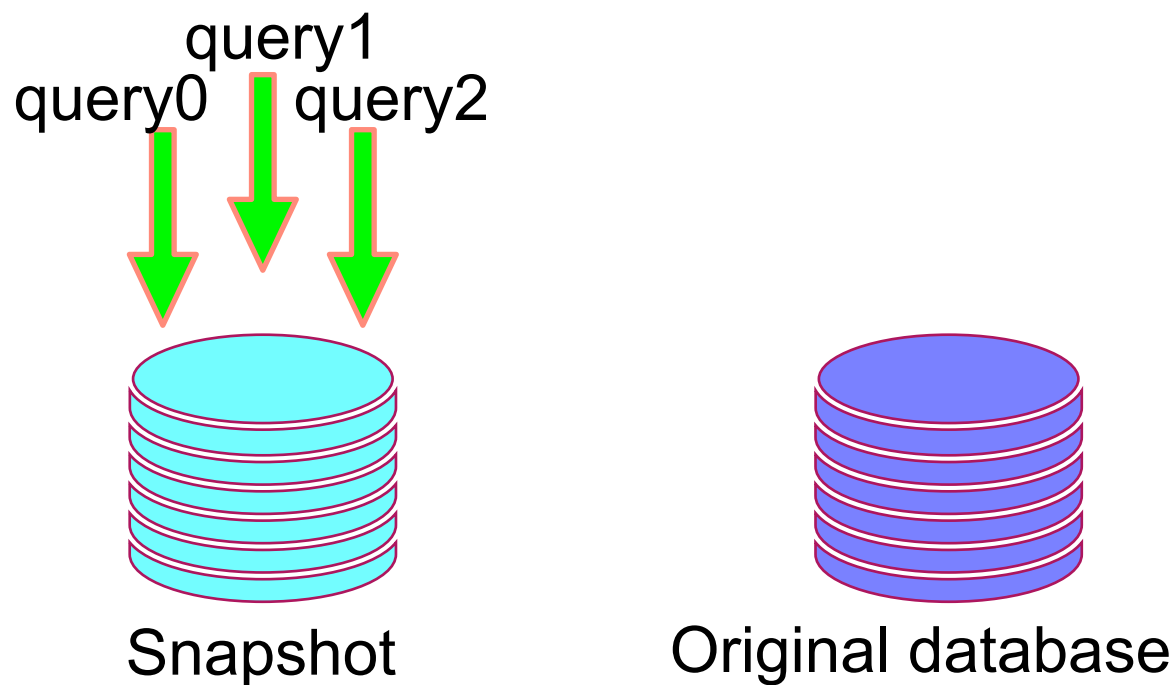
A snapshot is taken just before a snapshot query (the first DML query) is executed.



Recap: snapshot isolation

Recap2

All queries are executed on the snapshot.

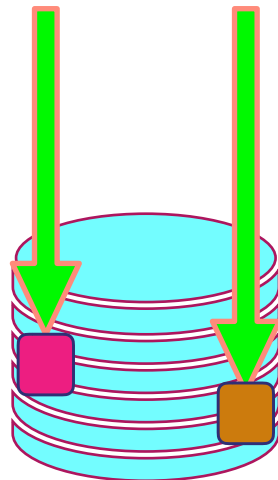


Recap: snapshot isolation

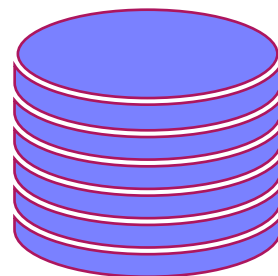
Recap2

Update queries are also executed on the snapshot.

UPDATE DELETE



Snapshot

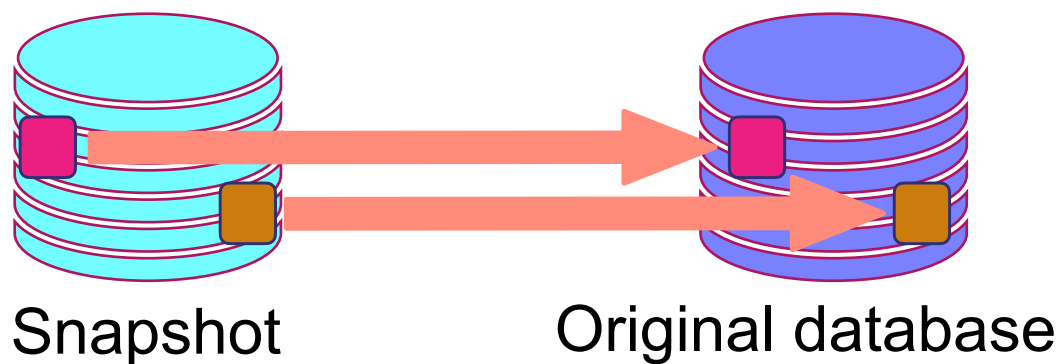


Original database

Recap: snapshot isolation

Recap3

The updates are applied to the original database when the transaction commits.

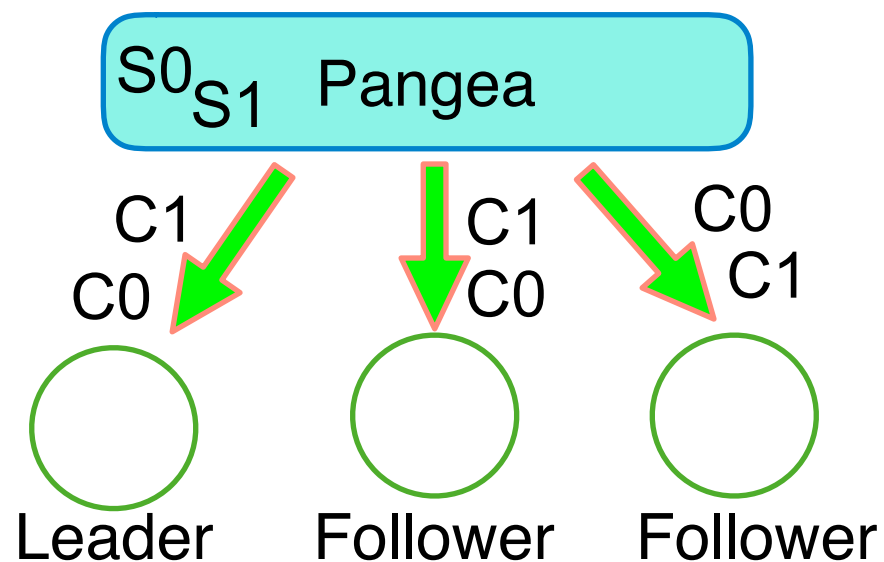


Pangea's algorithm1

- Pangea propagates **snapshot** and **commit** queries **exclusively**.
- That is, Pangea propagates **commit** queries only when all instances do not execute **snapshot** queries.
- Conversely, Pangea propagates **snapshot** queries only when all instances do not execute **commit** queries.

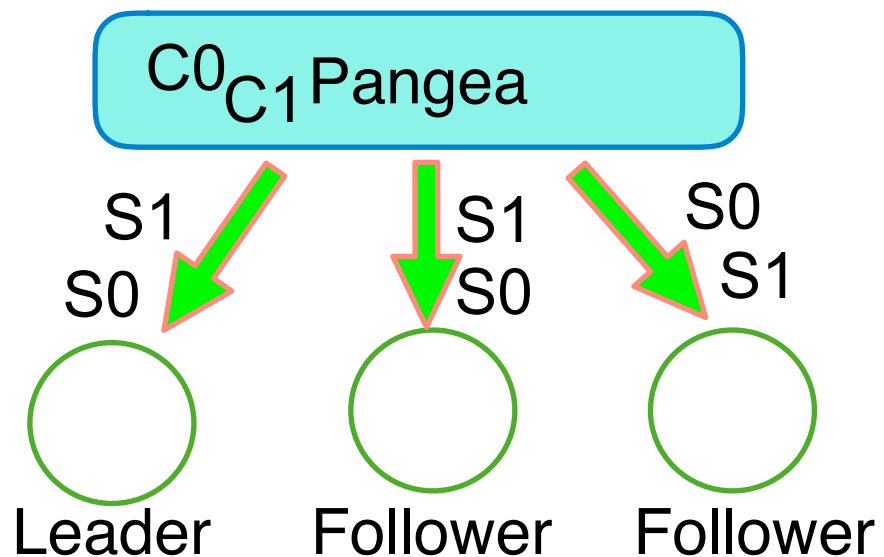
Pangea's algorithm1

- When Pangea sends **commit** queries to all instances, it stops sending **snapshot** ones.



Pangea's algorithm1

- When Pangea sends **snapshot** queries to all instances, it stops sending **commit** ones.

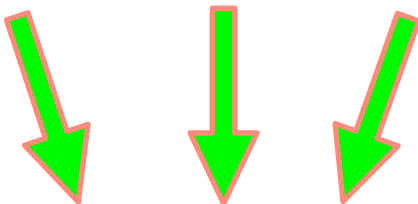


Algorithm1

This algorithm introduces that all instances must provide the same snapshot data.



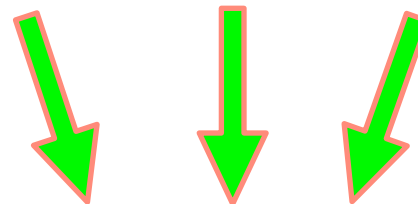
S0 S1 S2



Instance0



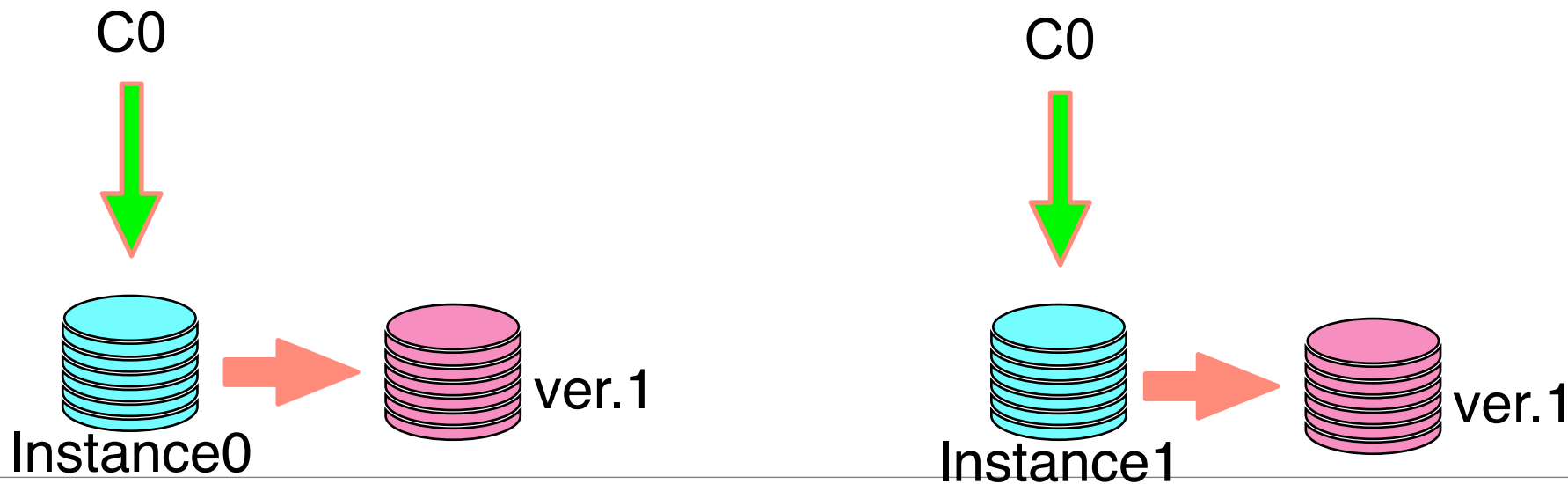
S0 S1 S2



Instance1

Algorithm1

This algorithm introduces that all instances must provide the same snapshot data.



Algorithm1

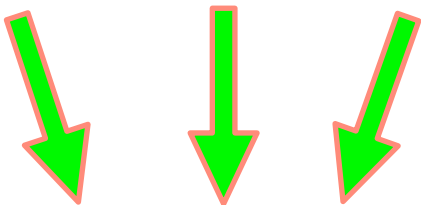
This algorithm introduces that all instances must provide the same snapshot data.



S3

S4

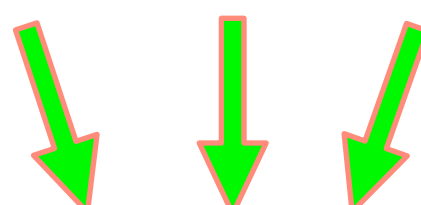
S5



S3

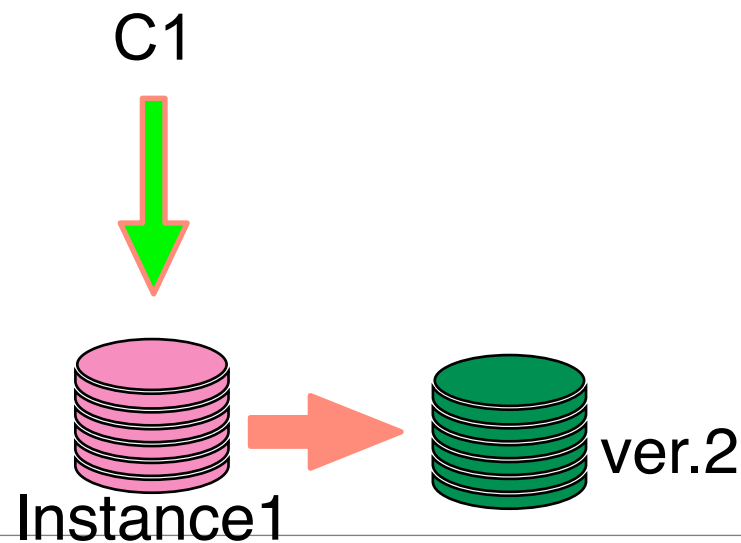
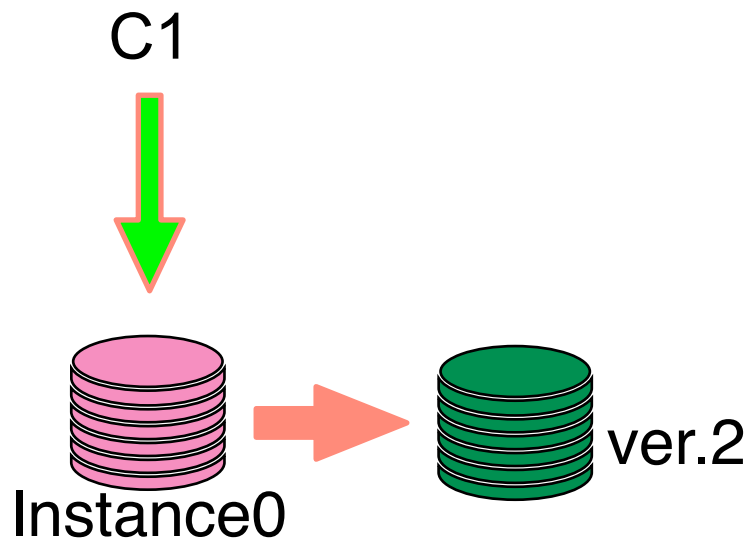
S4

S5



Algorithm1

This algorithm introduces that all instances must provide the same snapshot data.

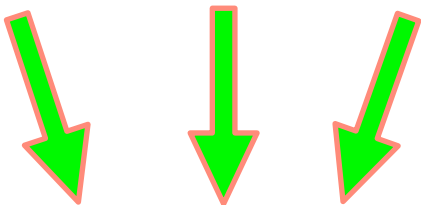


Algorithm1

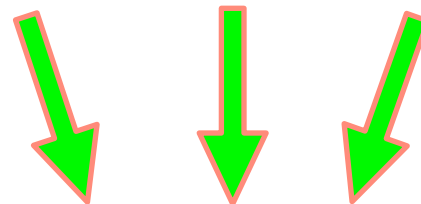
This algorithm introduces that all instances must provide the same snapshot data.



S6 S7 S8



S6 S7 S8

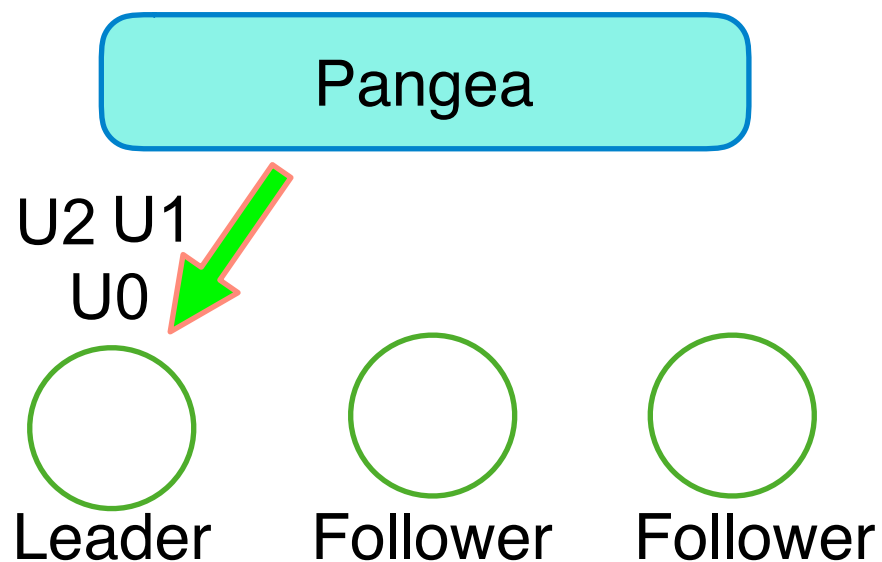


Algorithm1

The exclusive execution of **snapshot** and **commit** queries makes all instances create the same snapshot!

Pangea's algorithm2

- First, Pangea propagates all UPDATE queries to only the leader.

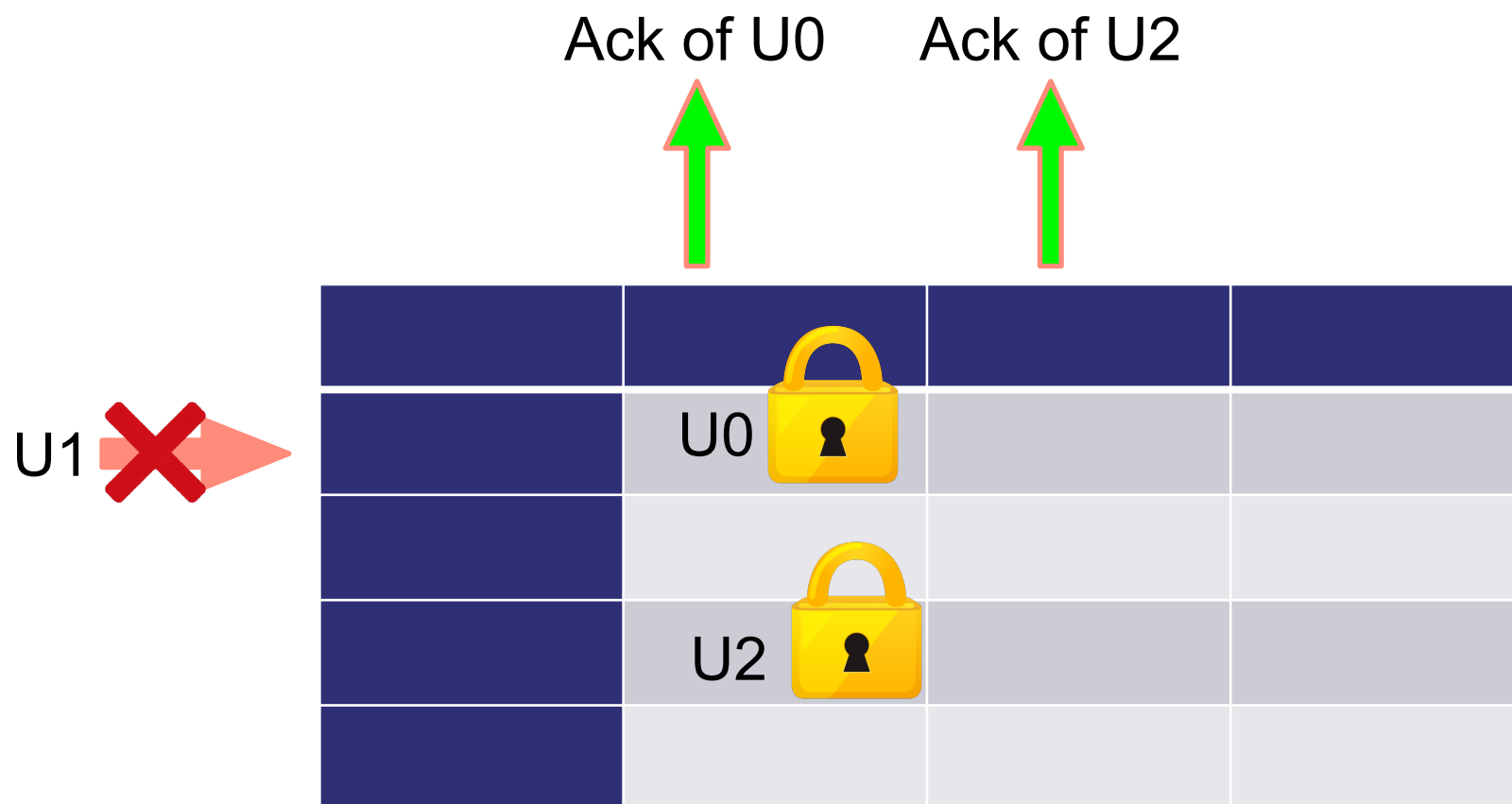


U0, U1



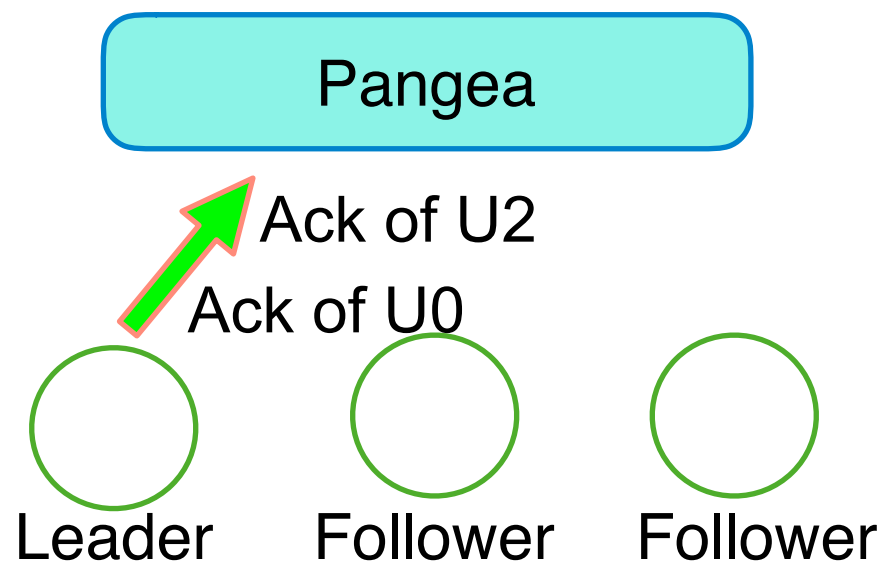
U2





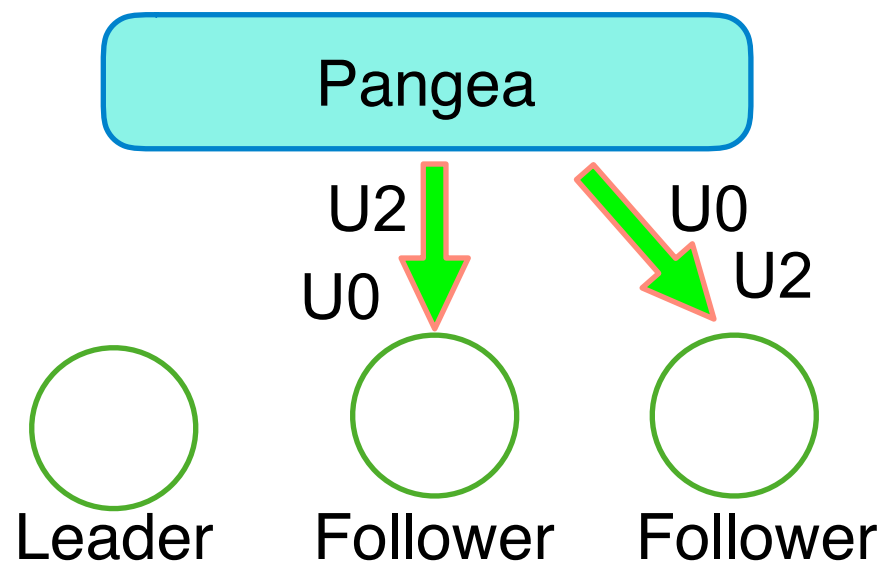
Pangea's algorithm2

- First, Pangea propagates all UPDATE queries only to the leader.



Pangea's algorithm2

- Second, if Pangea receives the answer, it sends the query to all the followers.

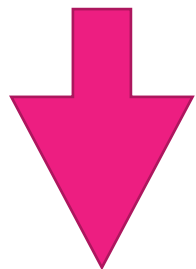


Pangea's algorithm2

- This protocol introduces that write/write conflicting queries are serialized by the leader,
- and non-conflicting queries are executed on all instances simultaneously.

In a nutshell

- Algorithm1 makes all instances create the same snapshot, and
- Algorithm2 makes the order of only conflicting queries serialized by the leader.



- Pangea's algorithms keep consistency between the leader and followers loosely.
- Any instance can return the latest consistent data.
- Surprisingly, Pangea's algorithms need not modify the source codes of Postgres!



PART 4

Summary

Summary

- We proposed the algorithms to address the disadvantage of streaming replication.
- Formally, we gave the full proof of the correctness in our VLDB2009 paper, please see it for detail.
- Our algorithms have been implemented into the latest version of Pgpool-II, and therefore, enjoy the fruit of Pangea by using Pgpool-II !

THANK YOU

mishima@sraoss.co.jp



PCC POSTGRESCONF
CN 2021

PGConf.Asia 12.14-17