OceanBase:
A 707 Million tpmC Distributed Relational Database System

Shuaijie Li

Infrastructure

Layer Level

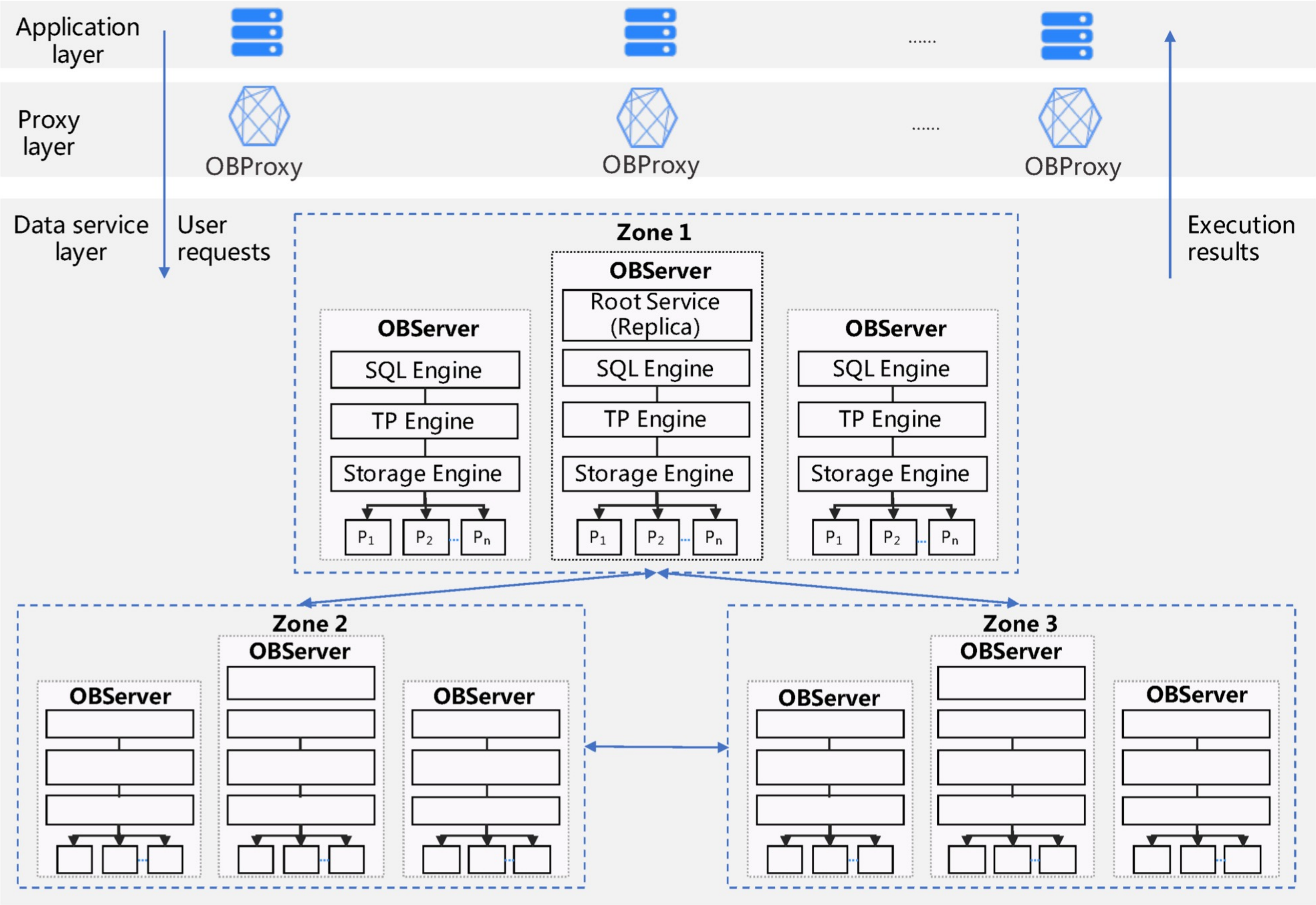
- Three Layers

Zone Level

- Zones in a Cluster
- Transaction Replication
- Cross-Region Disaster Tolerance

Node Level

- Shared-Nothing Architecture
- Table Partitioning
- Replicas and Paxos Group
- SQL Execution
- Transaction Processing
- Cluster Management



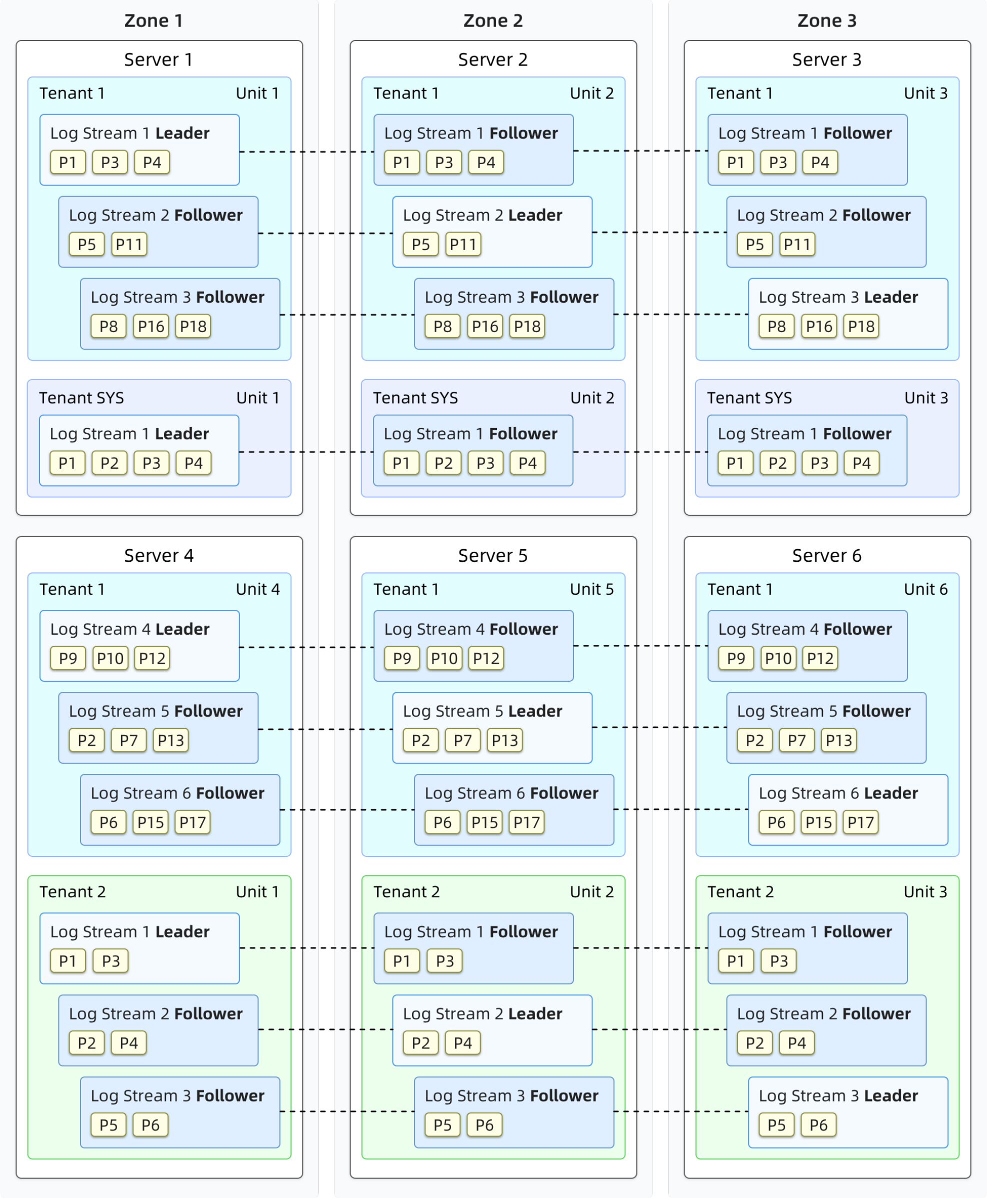
Infrastructure

Multi-Tenancy

- System Tenant
  - Container of the system table
  - Container for users with cluster management functions
  - Provides resources for maintenance and management

- Ordinary Tenant (similar to MySQL instances)
  - Can create its own users
  - All objects can be created
  - Independent information
  - Independent system variables

- Resource Isolation
  - Memory is completely isolated
  - CPUs are isolated through user-mode scheduling
  - Data structures are separated
  - Transaction-related data structures are separated



SQL Layer

Components at the SQL layer

1. The parser performs lexical and syntactic parsing.
2. The resolver performs semantic parsing.
3. The transformer rewrites the SQL statements in equivalent but different formats based on internal rules or cost models, and then sends the equivalent statements to the optimizer.
4. The optimizer generates the best execution plan for the SQL query.
5. The code generator converts the execution plan into executable code but does not optimize the plan.
6. The executor initiates the SQL execution.

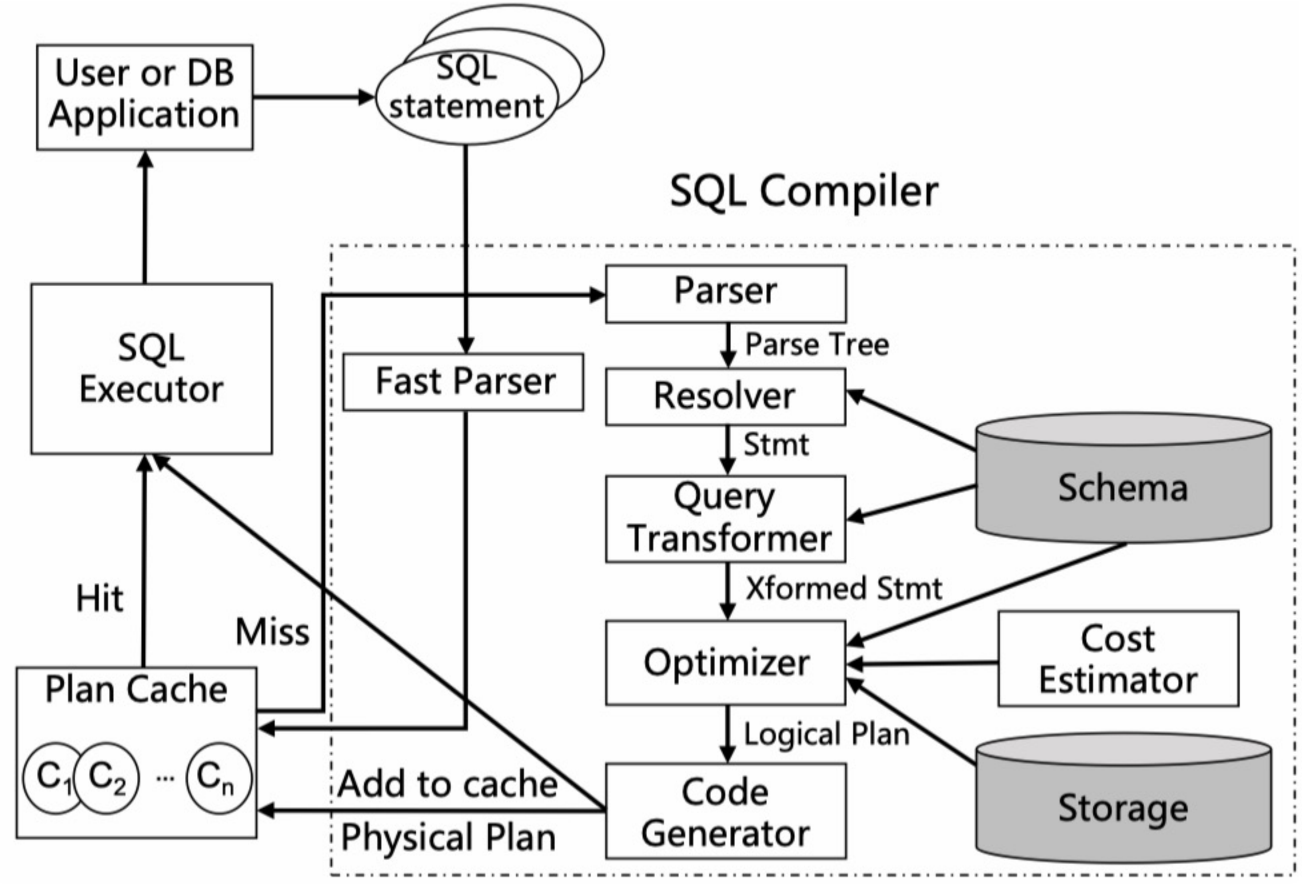
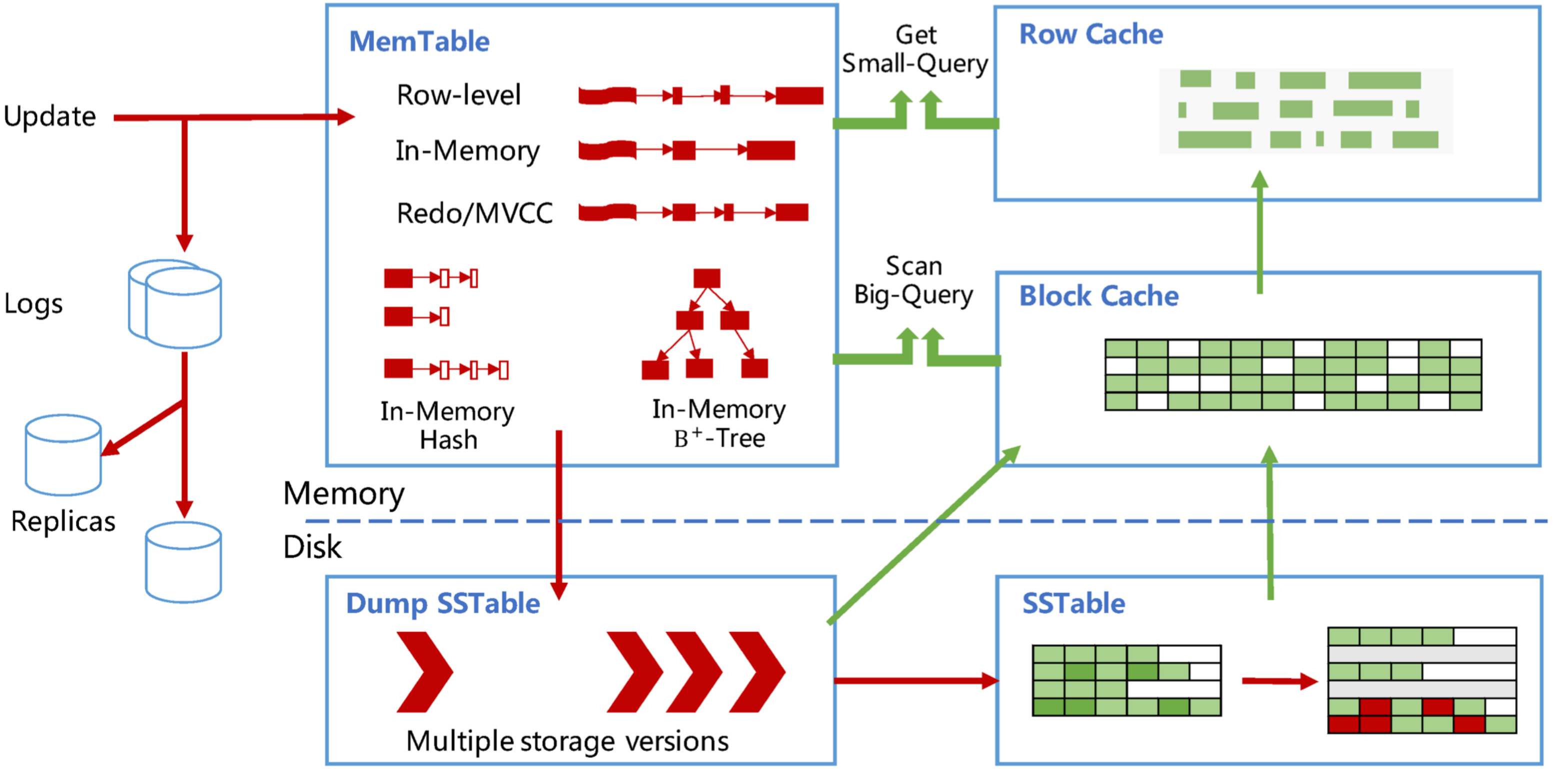


Figure 2: SQL Engine.

Storage Layer

LSM Tree-Based Architecture

- SSTable
  - Store static baseline data
  - Read-only
- MemTable
  - Store dynamic incremental data
  - Stored in memory
  - Consists of B-tree and hashtable
  - When reaches a certain size, minor compaction will be performed



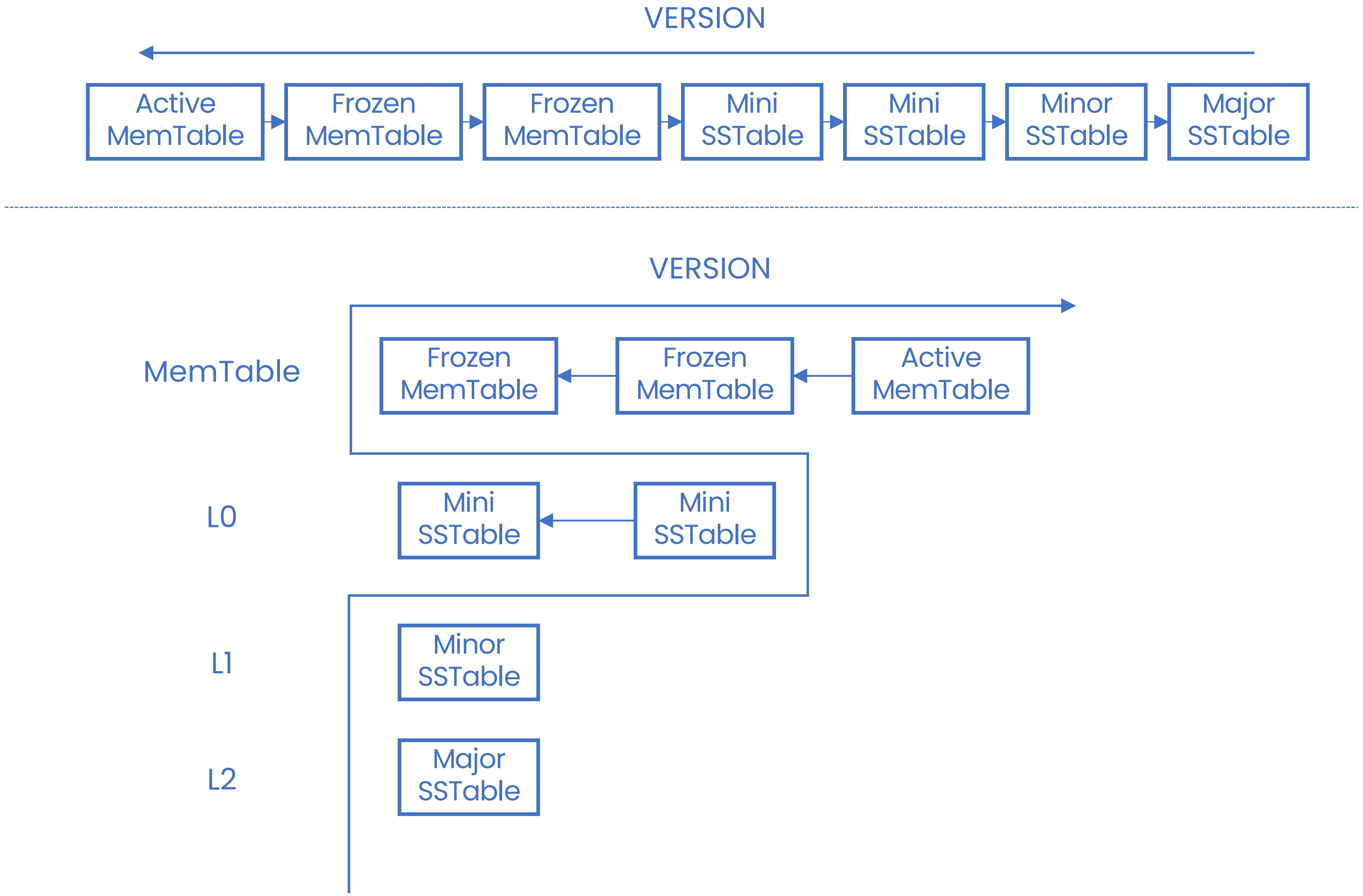
Storage Layer

Storage Structure

- Microblock (read unit): 4KB ~ 512KB
- Macroblock (write unit): 2MB
    - basic unit of allocation and garbage collection of the storage system

Major Compaction

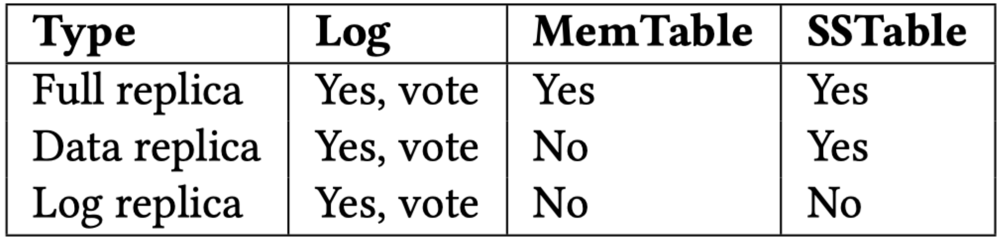
- If there is certain data modification (insert, update, delete) within a macroblock, the macroblock will be rewritten.
- Otherwise, the macroblock will be reused in the new baseline data without any IO cost.
- OceanBase staggers the normal service and the merge time through a round-robin compaction mechanism, thus isolating the normal user requests from the interference of the compaction operation.



Storage Layer

Replica Type

- Full replica
  - Baseline + Mutation increment + Redo log
- Data replica
  - Baseline + Redo log
  - Copies the minor compactions
  - Can be updated to a full replica
  - Can reduce both the CPU and memory cost
- Log replica
  - Redo log only
  - A member of the corresponding Paxos group
  - Can significantly reduce the storage and memory cost



Type	Log	MemTable	SSTable
Full replica	Yes, vote	Yes	Yes
Data replica	Yes, vote	No	Yes
Log replica	Yes, vote	No	No

Transaction Process Layer

Partition and Paxos Group

- A table partition is the basic unit for the data distribution, load balance, and Paxos synchronization.

- One Paxos group for each partition.

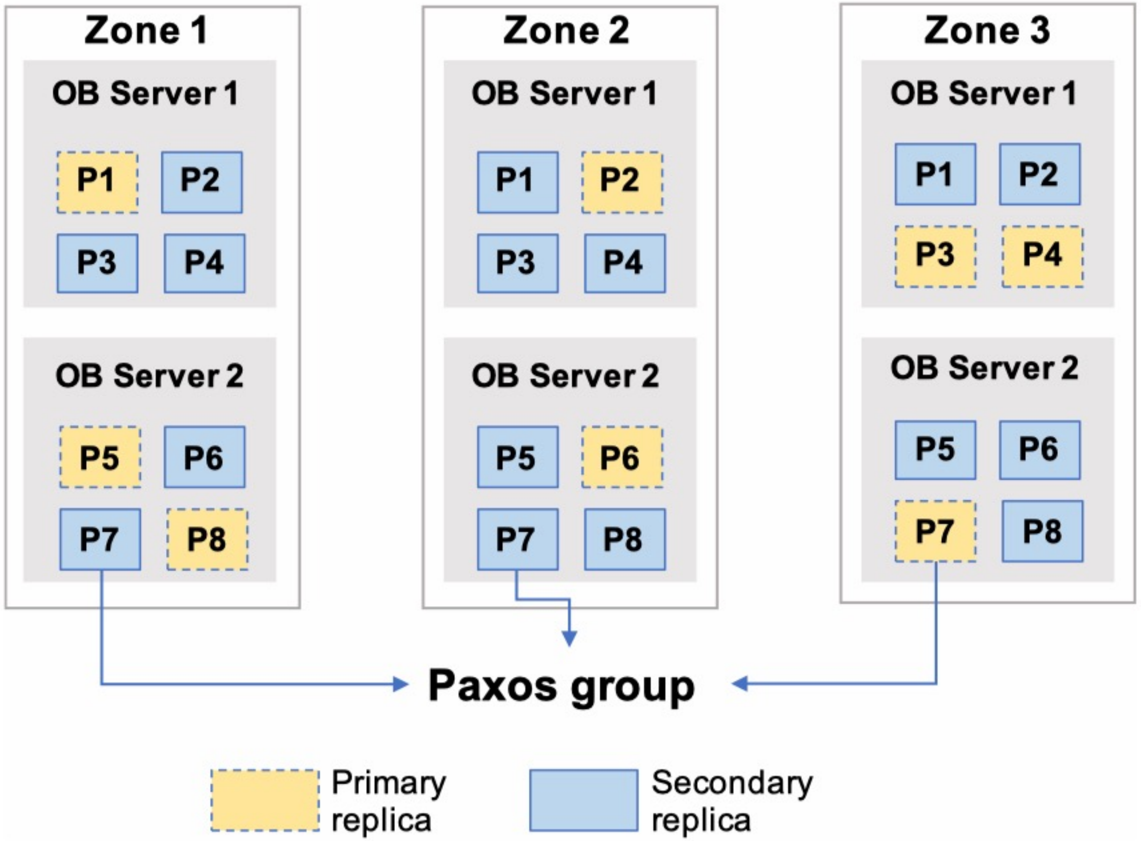


Figure 4: Paxos group.

Transaction Process Layer

Timestamp Service

- Paxos leader of the timestamp Paxos group is often in the same region as Paxos leaders of the table partitions.
- Each OceanBase node retrieves the timestamp from the timestamp Paxos leader periodically.

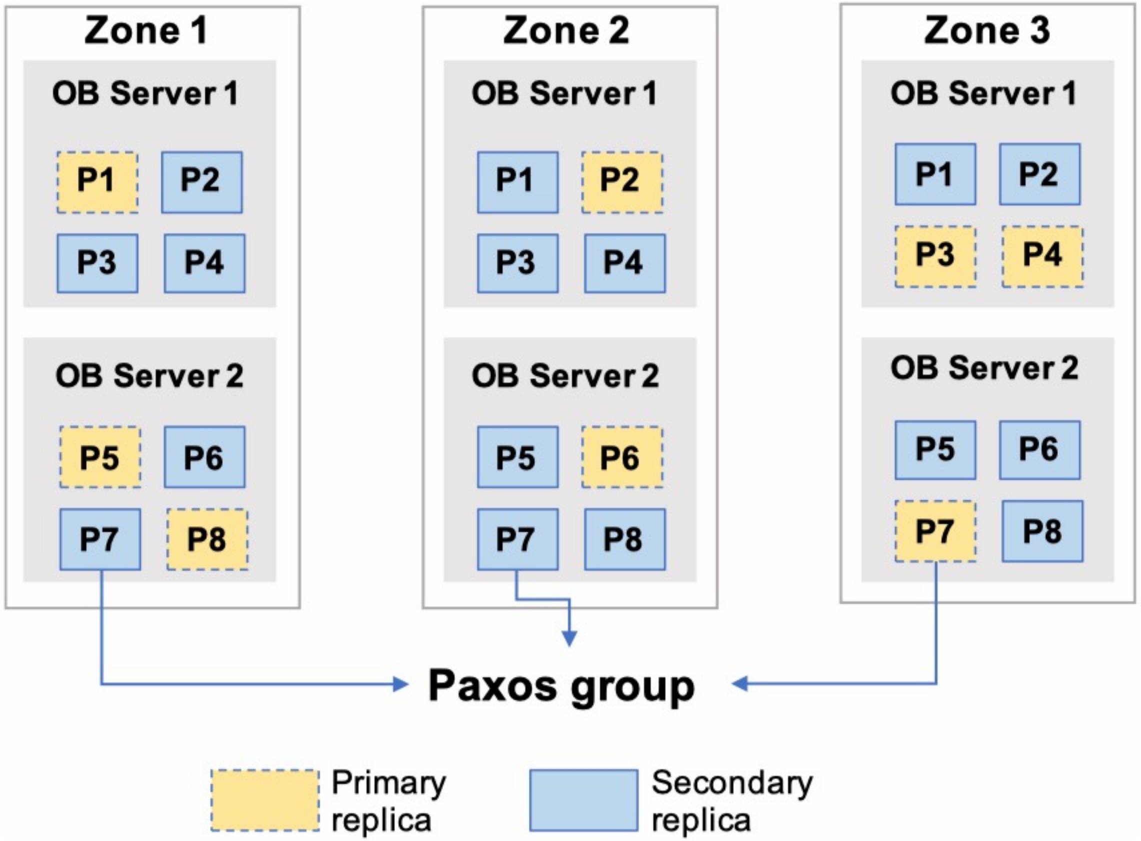


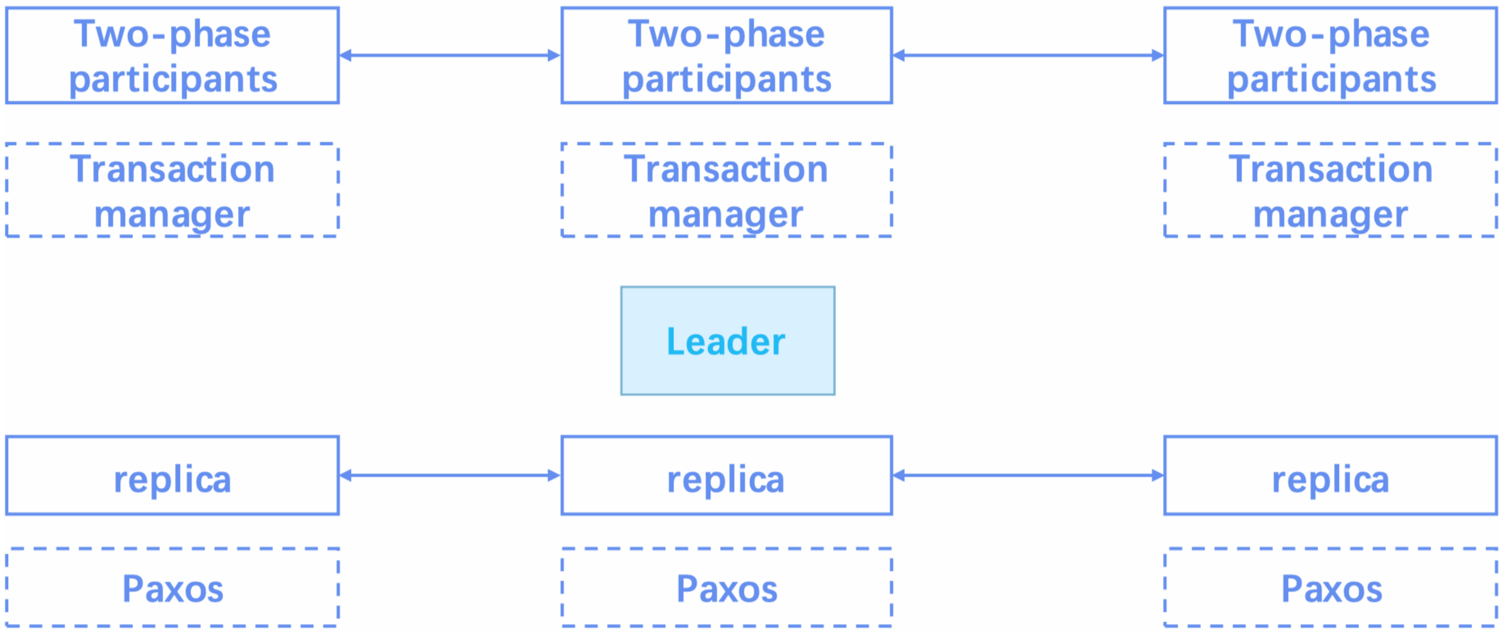
Figure 4: Paxos group.

Transaction Process Engine

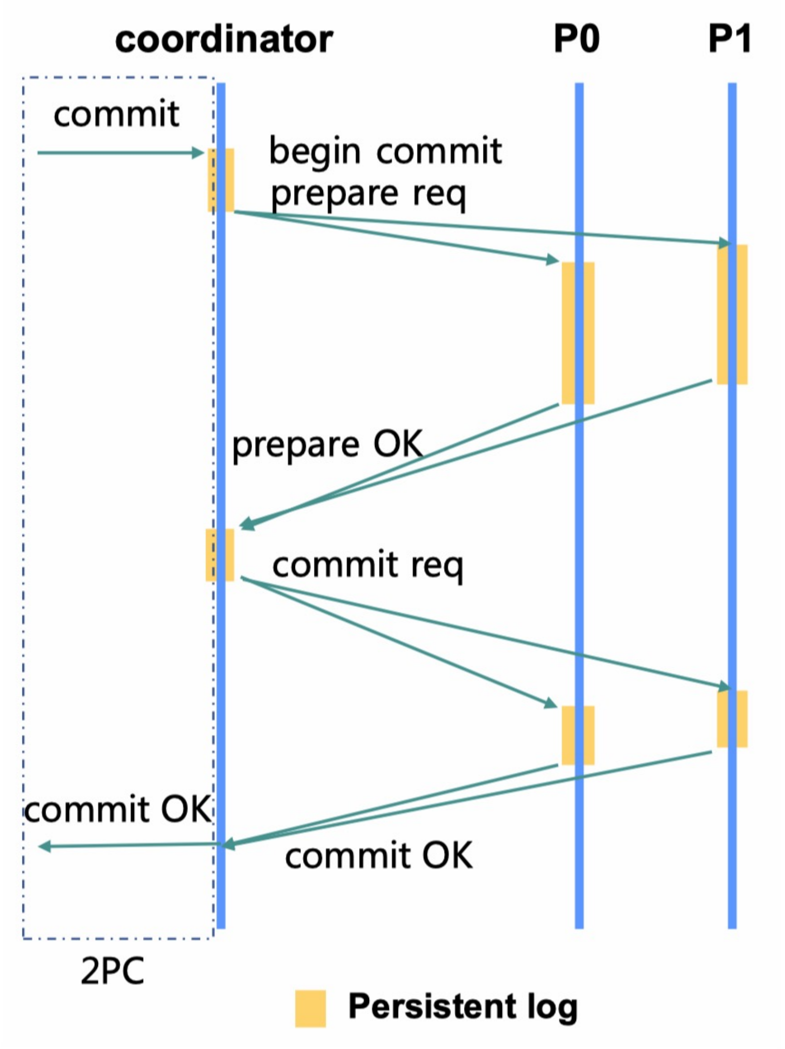
Paxos-based 2PC

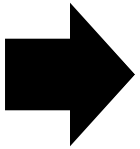
Each participant in the two-phase commit contains multiple copies, and the copies are readily available through the Paxos protocol.

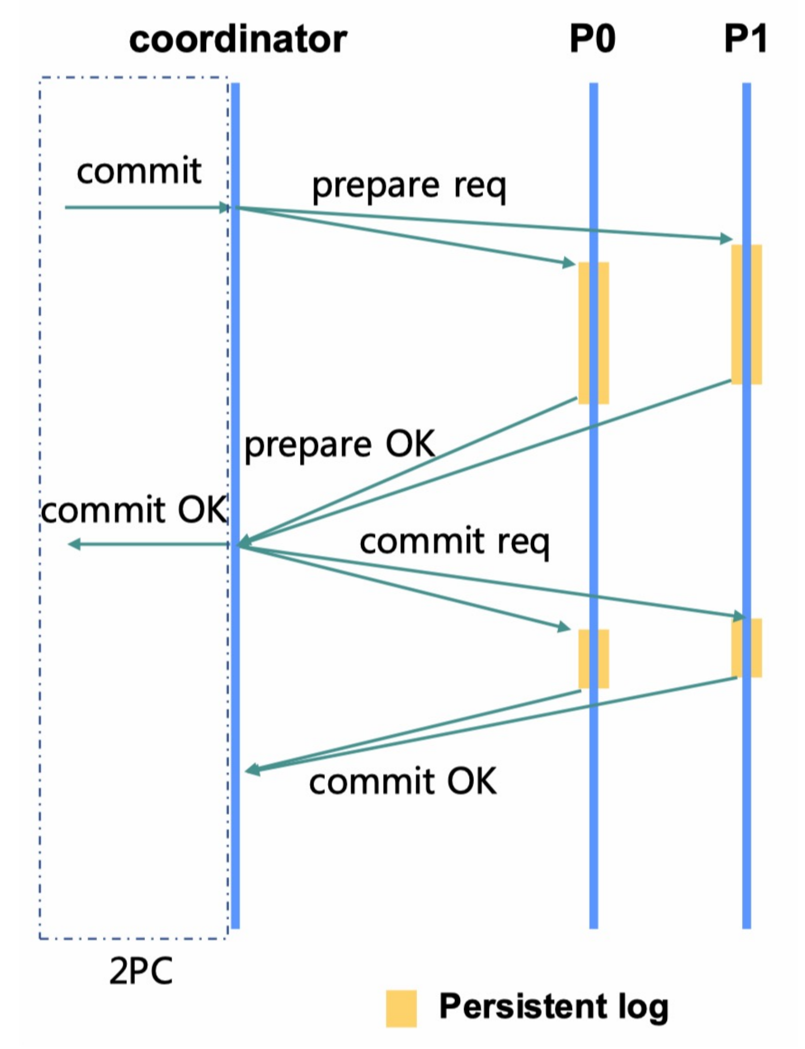
When a participant node fails, the Paxos protocol can quickly elect another replica to replace the original participant to continue providing services, and restore the state of the original participant.



Traditional 2PC vs OceanBase 2PC







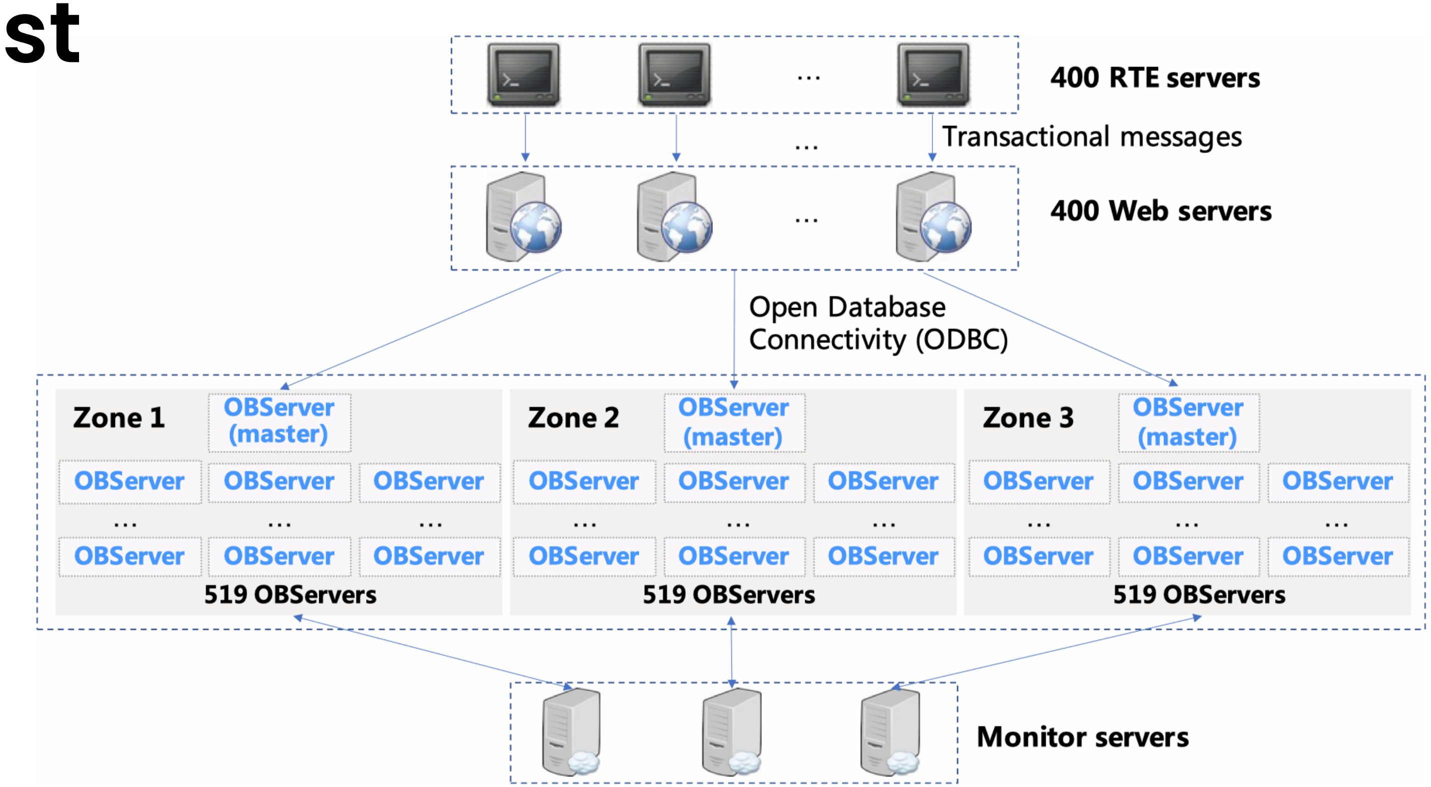
TPC-C Benchmark Test

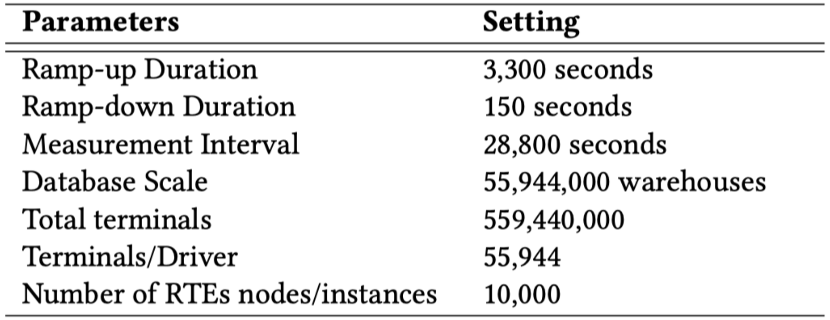
Benchmark Configuration

- 400 remote terminal emulator (RTE) servers to emulate the total 559,440,000 users

- 400 web servers

- The OceanBase cluster in this benchmark test consists of 1,557 servers in a shared-nothing architecture





Parameters	Setting
Ramp-up Duration	3,300 seconds
Ramp-down Duration	150 seconds
Measurement Interval	28,800 seconds
Database Scale	55,944,000 warehouses
Total terminals	559,440,000
Terminals/Driver	55,944
Number of RTEs nodes/instances	10,000

TPC-C Benchmark Test

Transaction per minute, Class C (tpmC)

- tpmC rises linearly as the number of data nodes increases.
- OceanBase is highly scalable.
- OceanBase has an online transaction processing performance of 707 million tpmC in 2020

- The cumulative tpmC variations during these tests are quite small

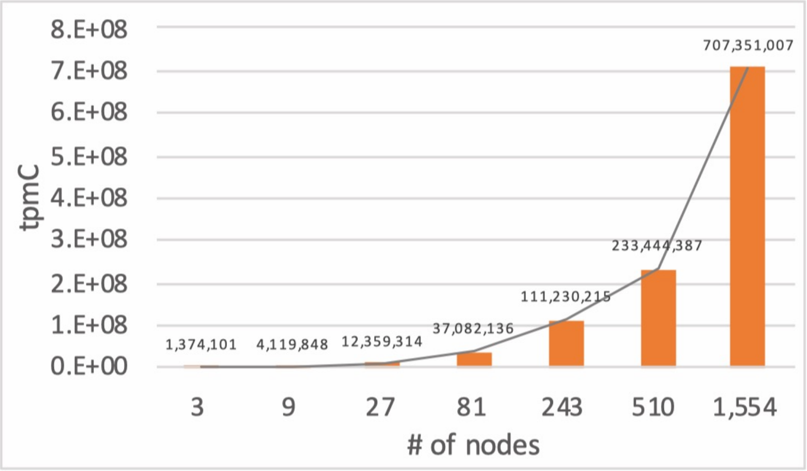
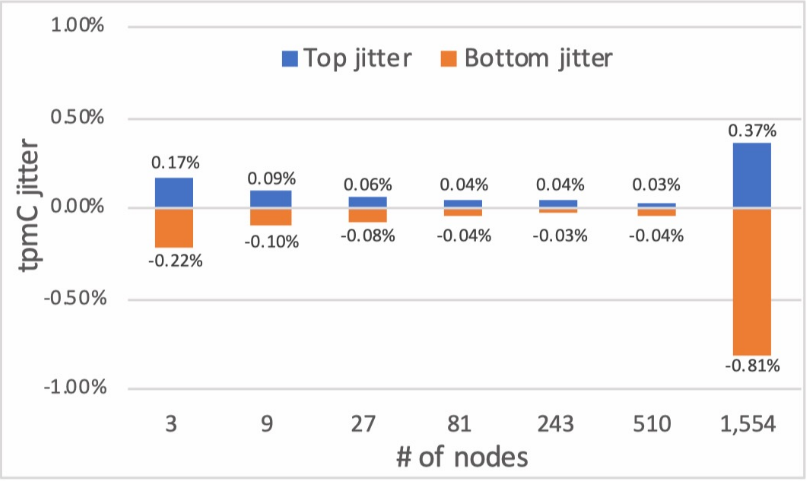
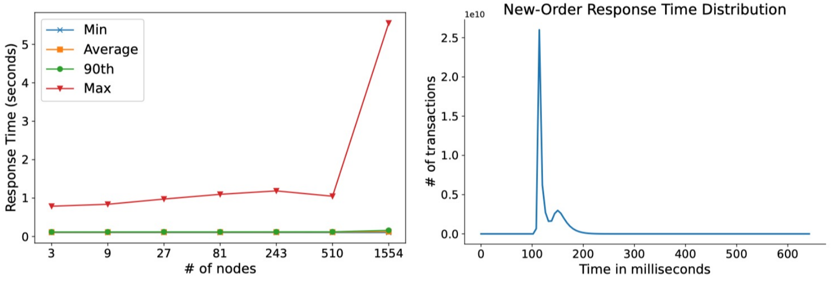


Figure 8: tpmC.



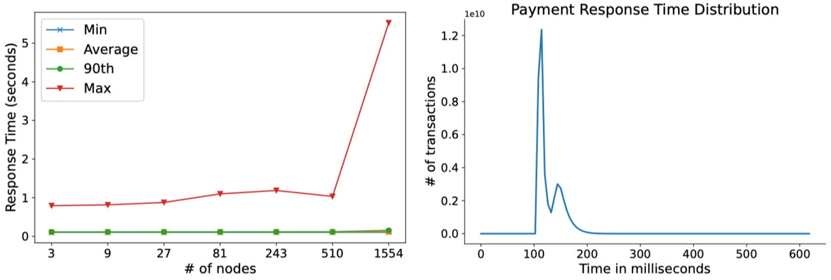
TPC-C Benchmark Test

Response Time (RT)



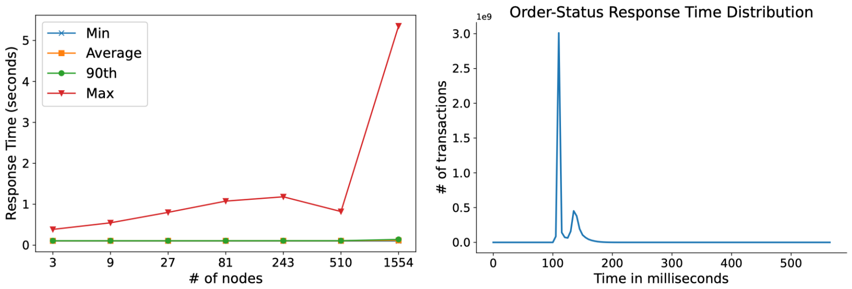
(a) Scalability (b) RT Distribution with 1,554 nodes

Figure 10: New-Order Response Time (RT).



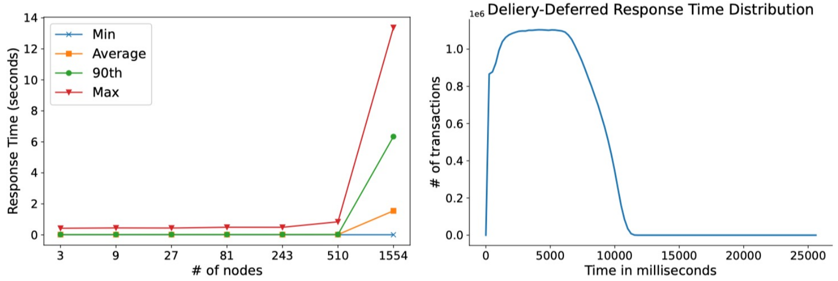
(a) Scalability (b) RT Distribution with 1,554 nodes

Figure 11: Payment Response Time (RT).



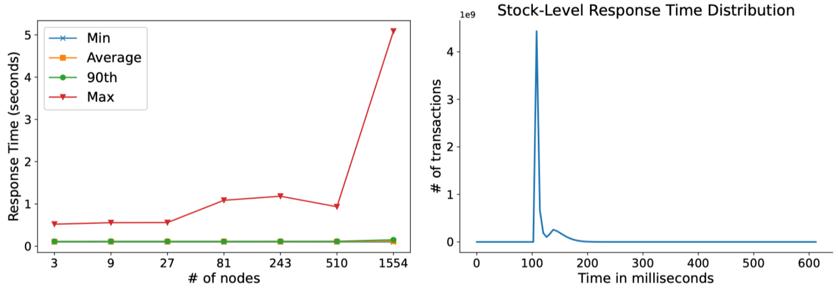
(a) Scalability (b) RT Distribution with 1,554 nodes

Figure 12: Order-Status Response Time (RT).



(a) Scalability (b) Delivery RT with 1,554 nodes

Figure 13: Delivery Response Time (RT).



(a) Scalability (b) RT Distribution with 1,554 nodes

Figure 14: Stock-Level Response Time (RT).

Questions?