

Solar: Towards a Shared-Everything Database on Distributed Log-Structured Storage

Tao Zhu¹, Zhuoyue Zhao², Feifei Li², Weining Qian¹, Aoying Zhou¹, Dong Xie²,
Ryan Stutsman², Haining Li³, Huiqi Hu^{1,3}

East China Normal University¹

University of Utah²

Bank of Communications³

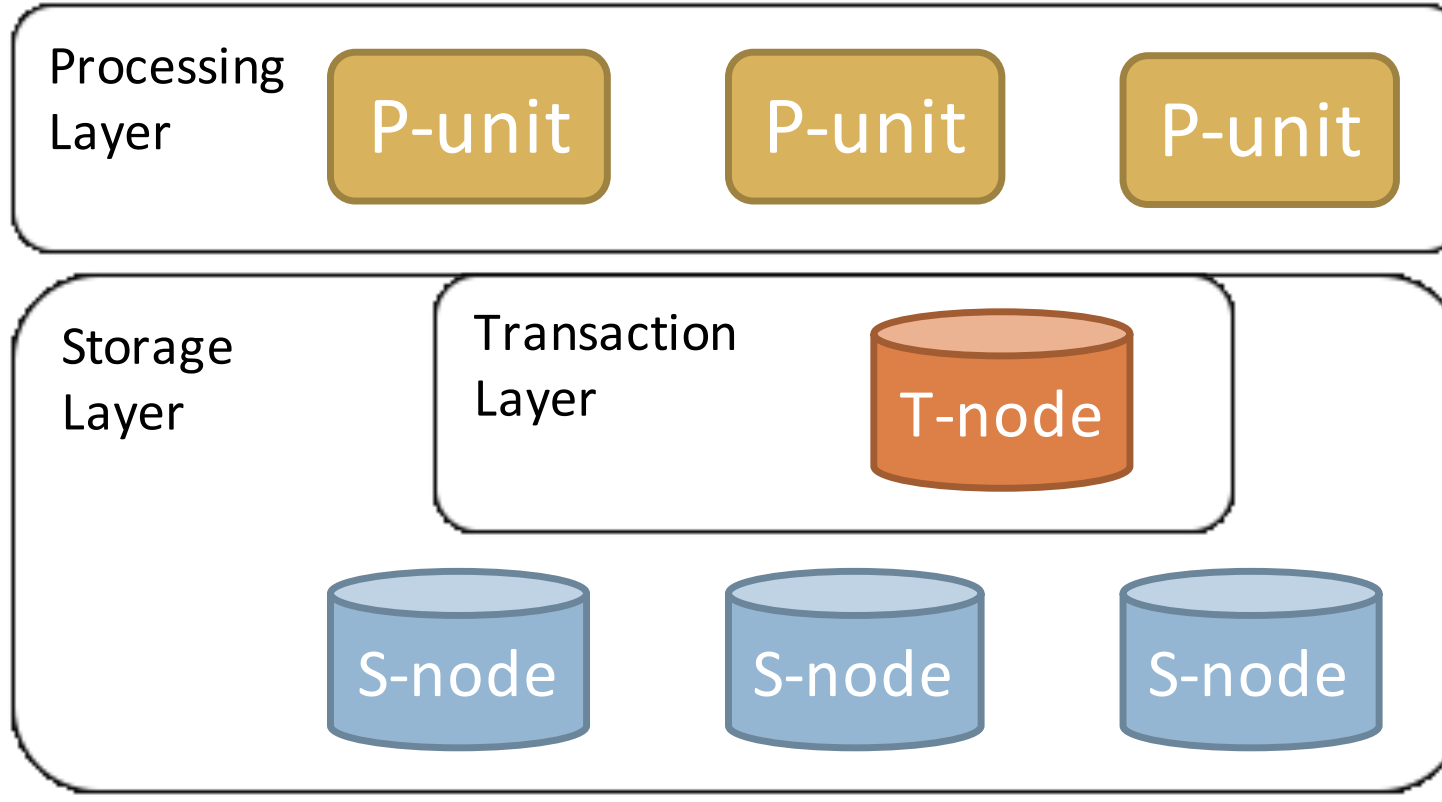
Architecture

Design Considerations

- (1) A shared-everything architecture
- (2) High performance in-memory transaction
- (3) Fine-grained remote data access

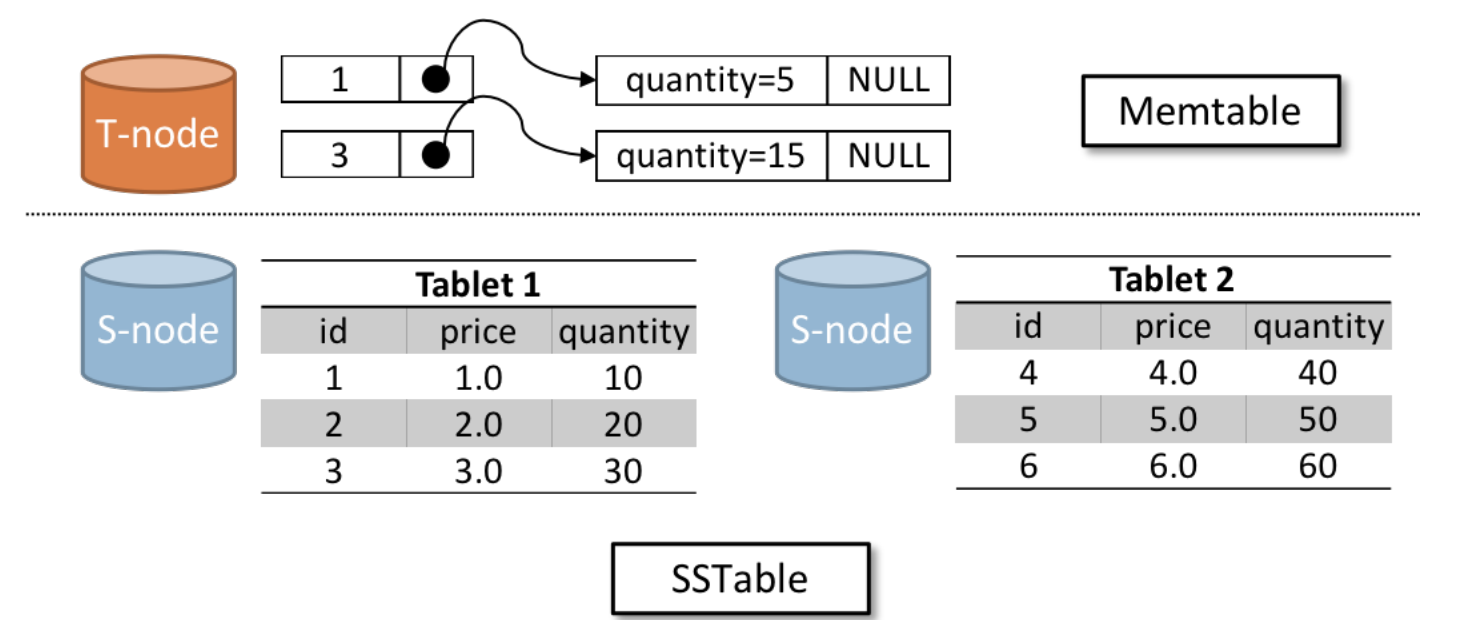
- *No assumption on workloads*
- *Do not rely on advanced hardwares*

Overview



LSM-Tree style storage

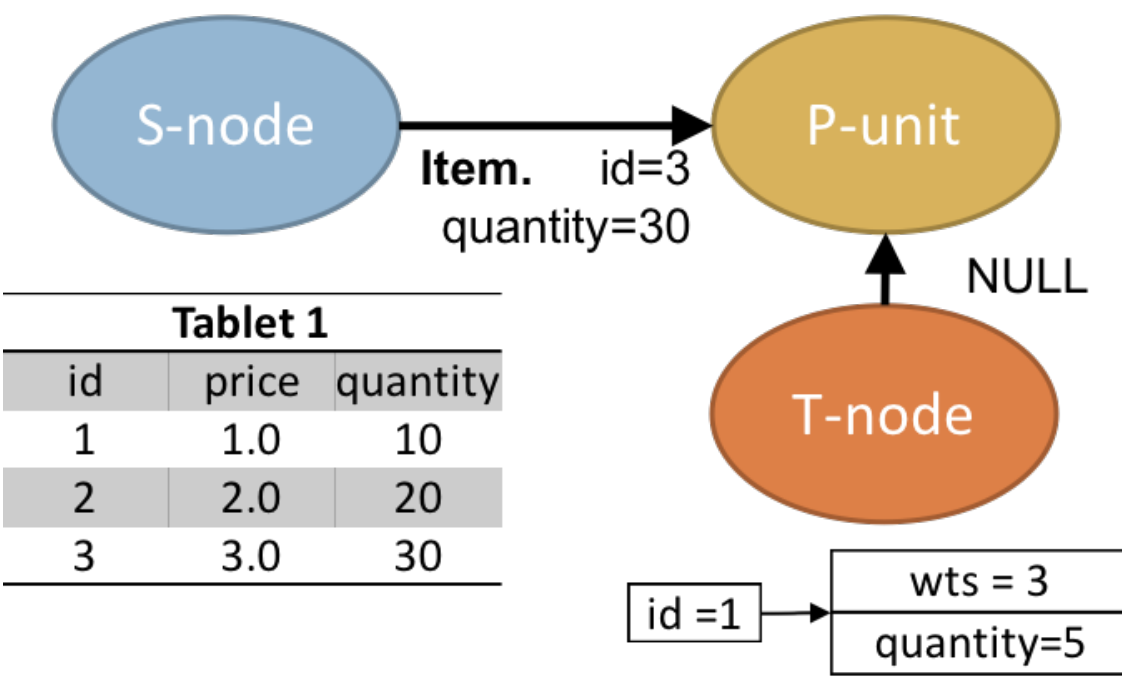
- (1) *SSTable*: a consistent snapshot
- (2) *Memtable*: newly committed data



Transaction Process

Execution

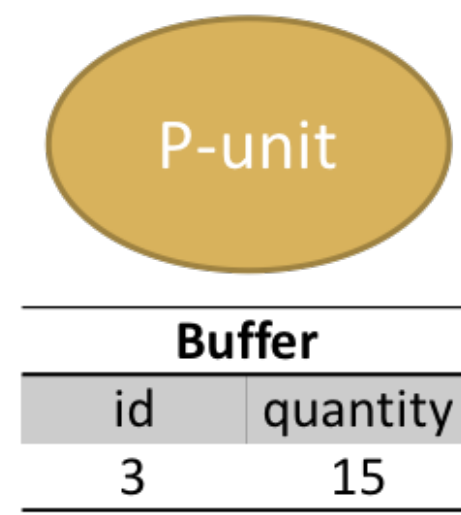
- (1) Acquire a read-timestamp (rts)
- (2) Read & execute & buffer writes locally



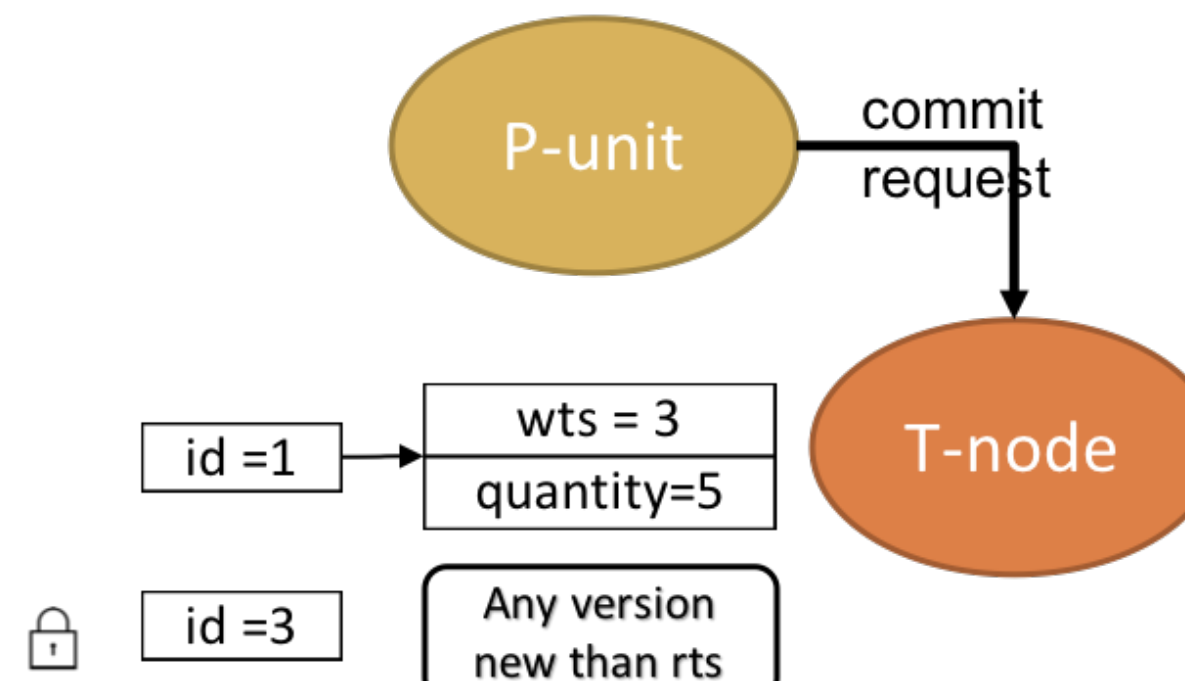
(a) Read

Validation

- (1) Acquire latches and check there is no new versions for records in write set



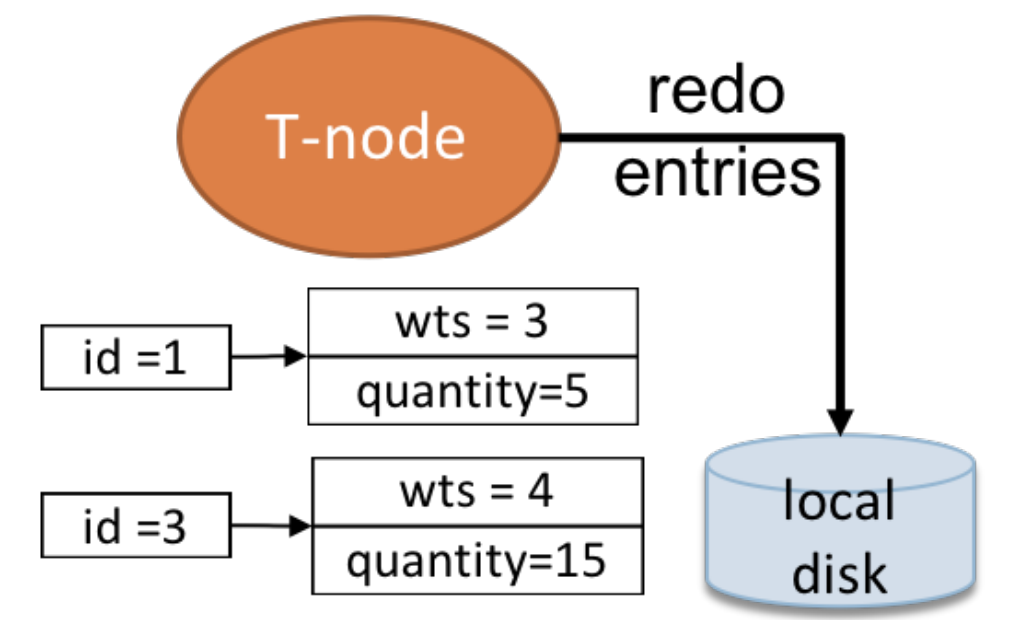
(b) Write



(c) Validate

Commit

- (1) Create new versions
- (2) Flush redo log entries

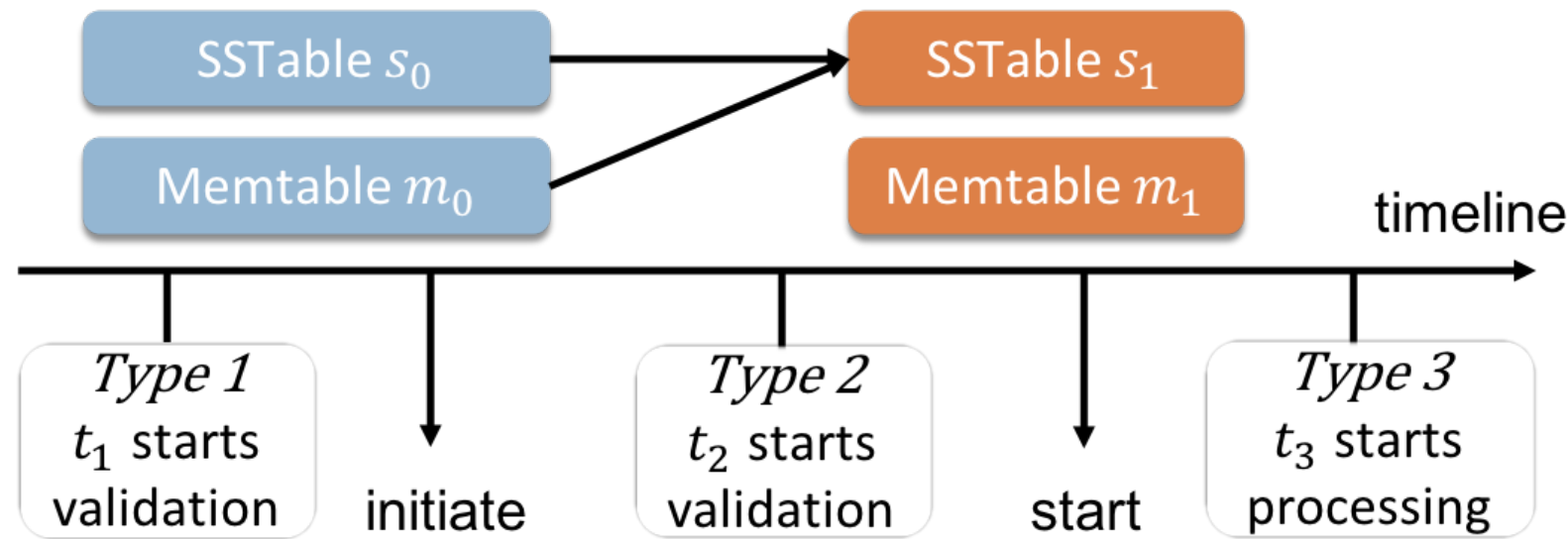


(c) Commit

Data Compaction

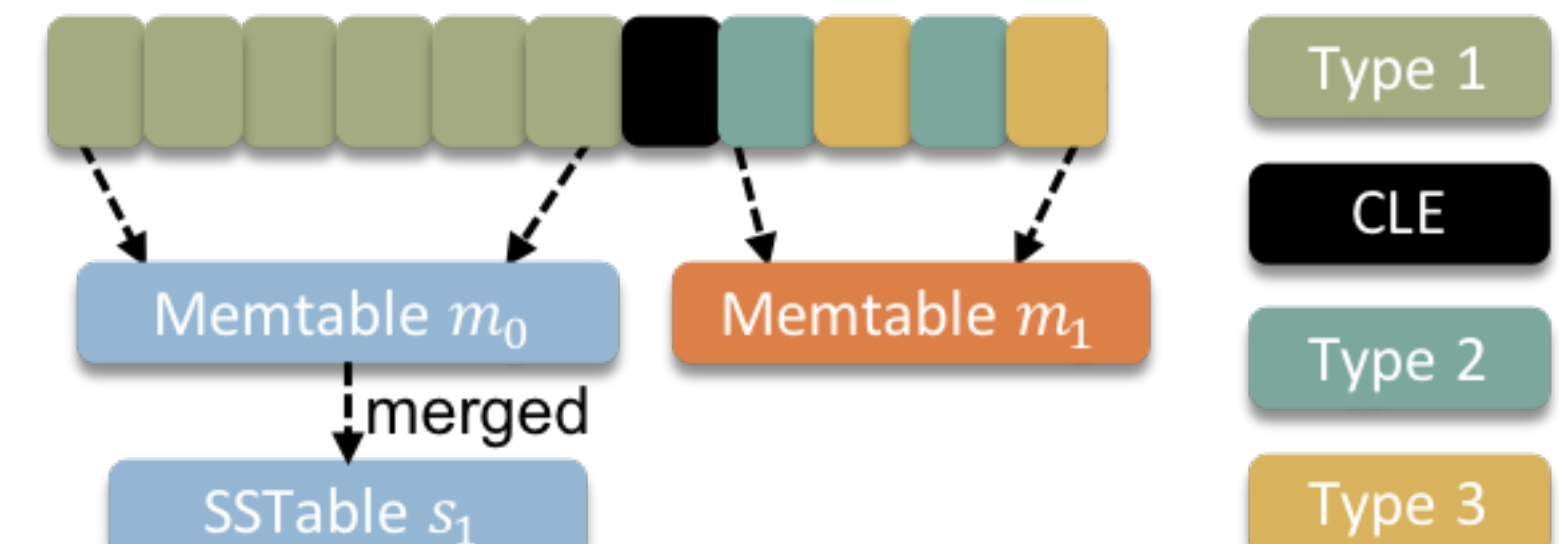
Compaction

- (1) Initiate
 - Create a new Memtable m_1
 - Wait until no one writes the old Memtable m_0 any more
- (2) Start
 - Create a new version of SSTable (s_1) by merging s_0 and m_0



Concurrency Control

- (1) Type 1: validate and write on m_0
- (2) Type 2: validate on m_0 and m_1 , write on m_0
- (3) Type 3: validate and write on m_1



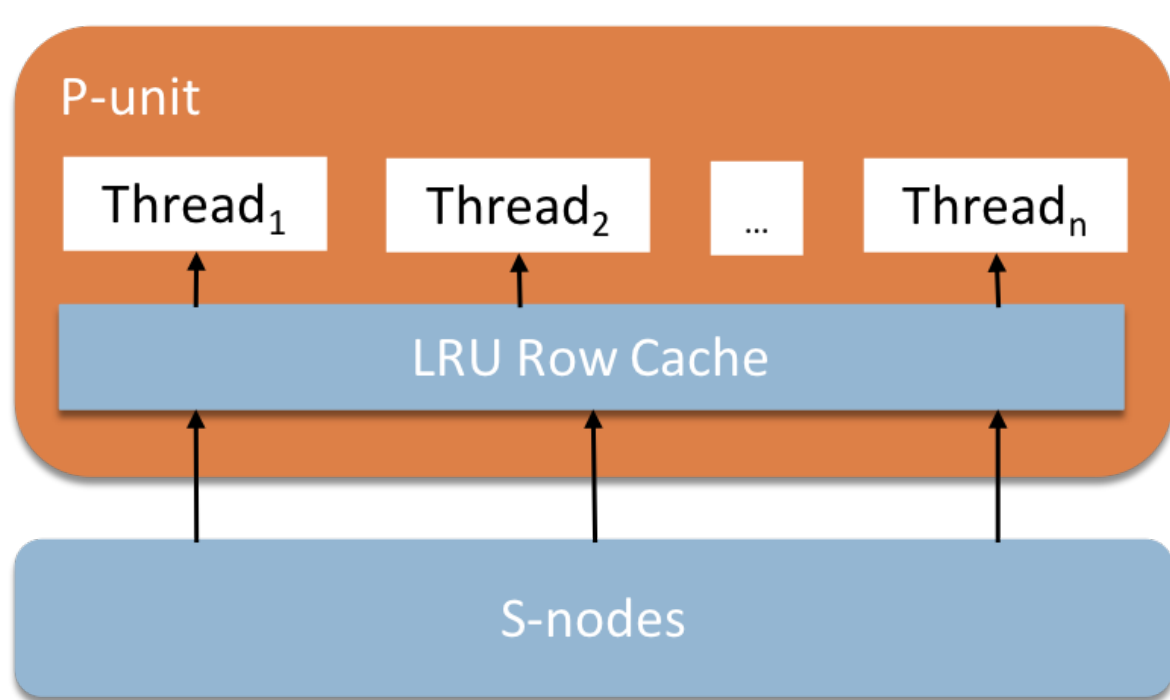
Data Recovery

- (1) Persist a compaction log entry (CLE) when the data compaction is started
- (2) Data recovery starts from the CLE of the last finished data compaction

Data Access Optimizations

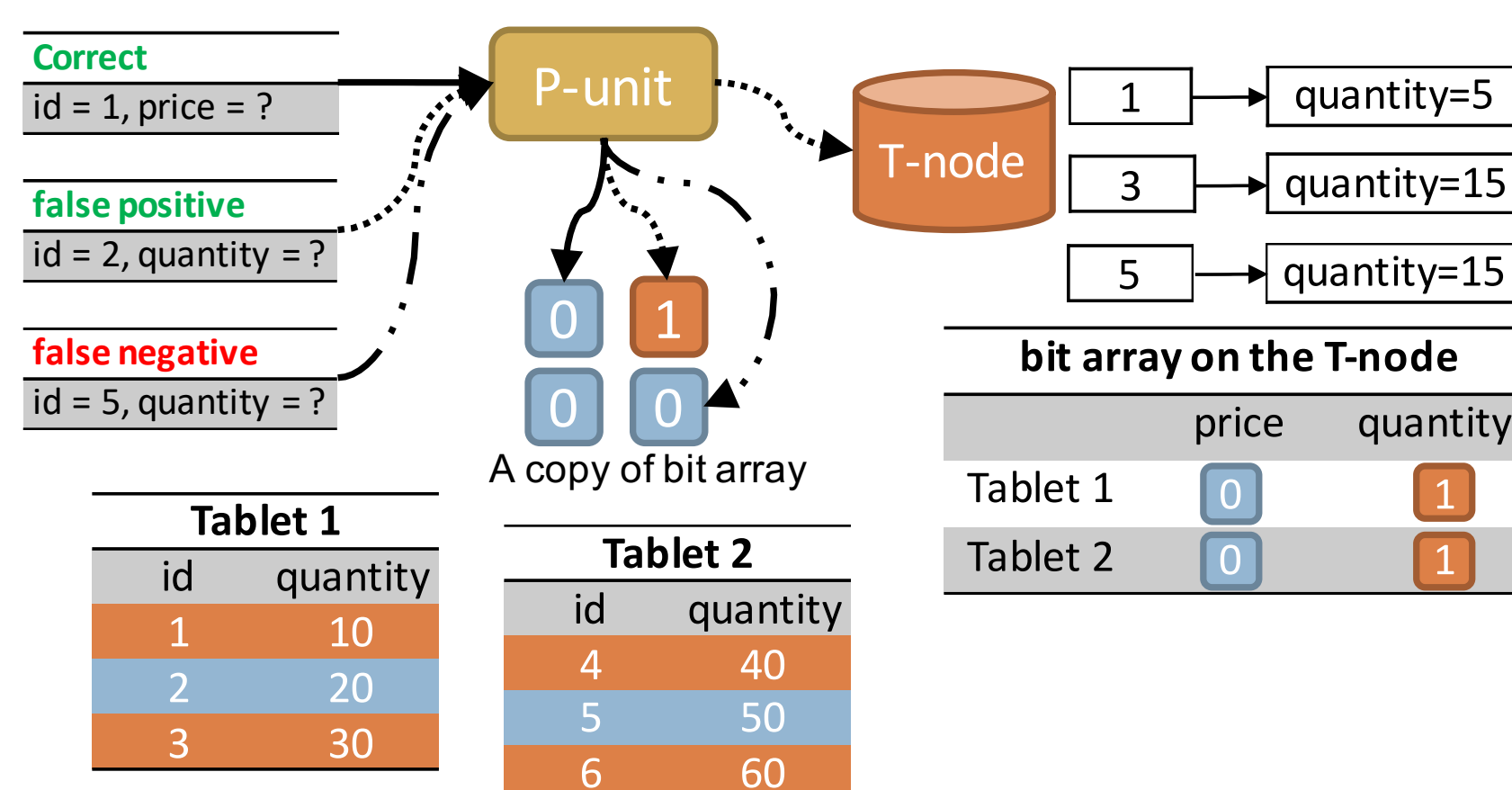
SSTable Cache

- Avoid duplicate S-node read
- Cache immutable SSTable on P-units



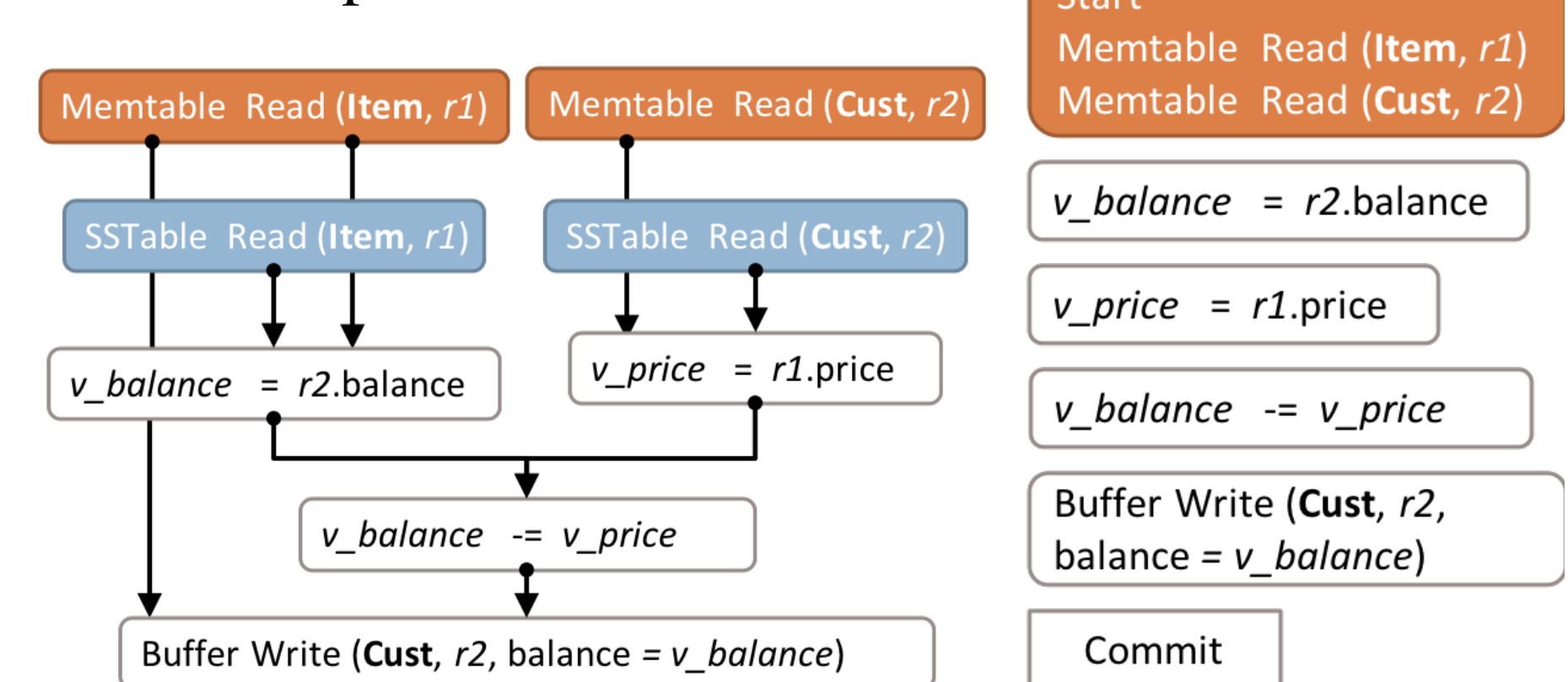
Asynchronous Bit Array

- Filter useless T-node read



Transaction Compilation

- Pre-execute S-node read
- Group T-node read



System Evaluation

System:

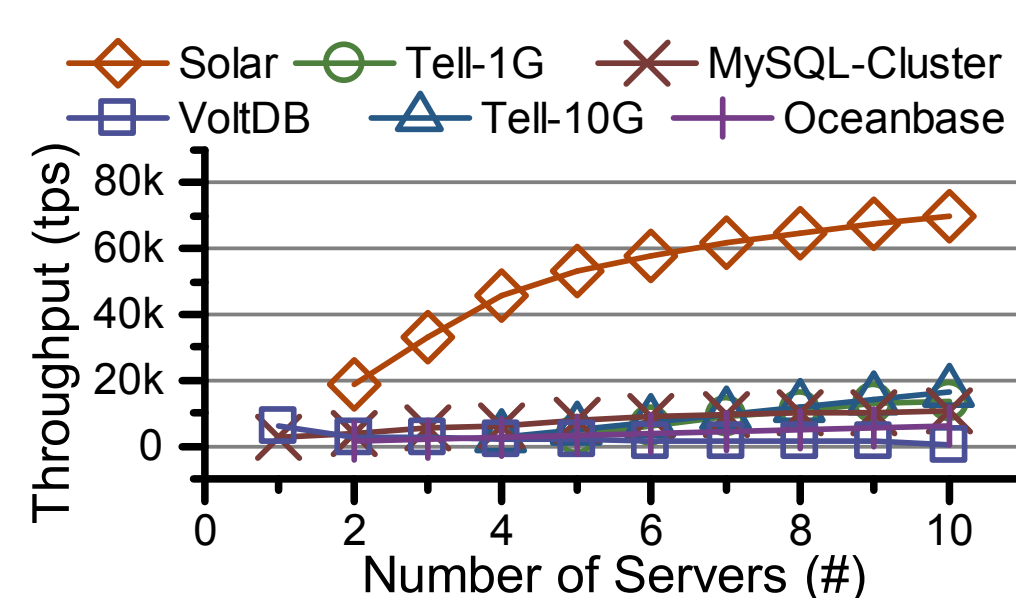
- (1) Solar
- (2) VoltDB
- (3) Tell
- (4) MySQL-Cluster
- (5) Oceanbase

Benchmark:

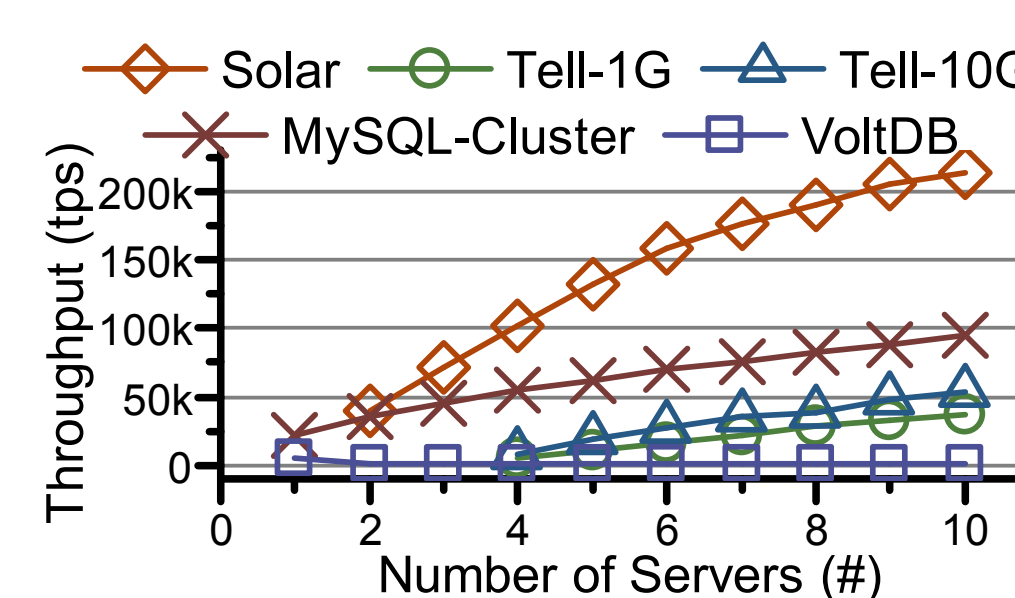
- (1) TPC-C
- (2) Smallbank
- (3) E-commerce

Hardware:

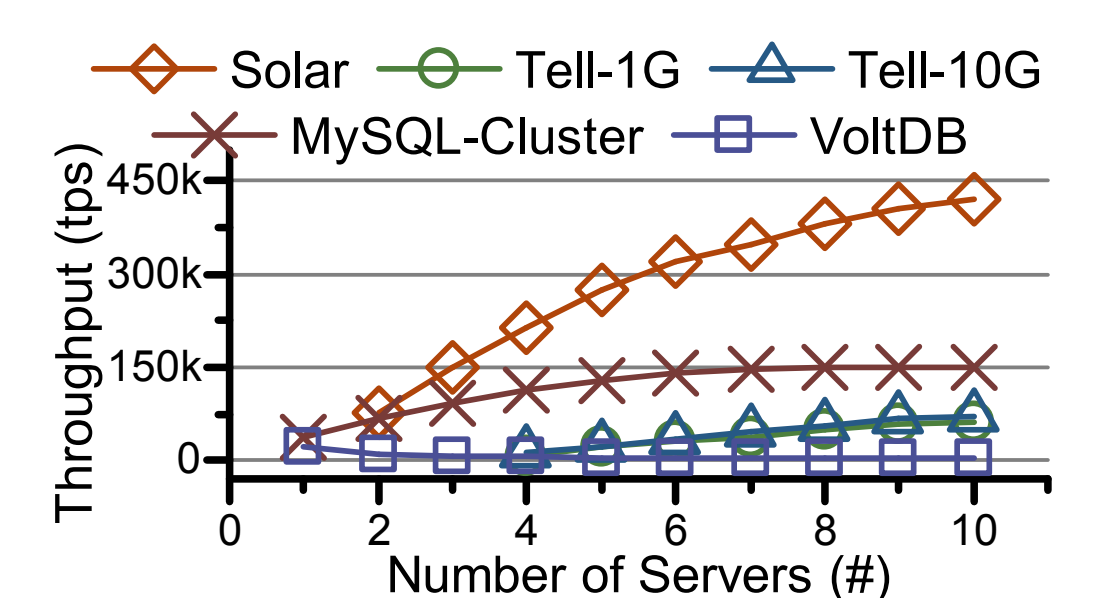
- (1) 11 nodes
- (2) 1/10 Gigabytes switch



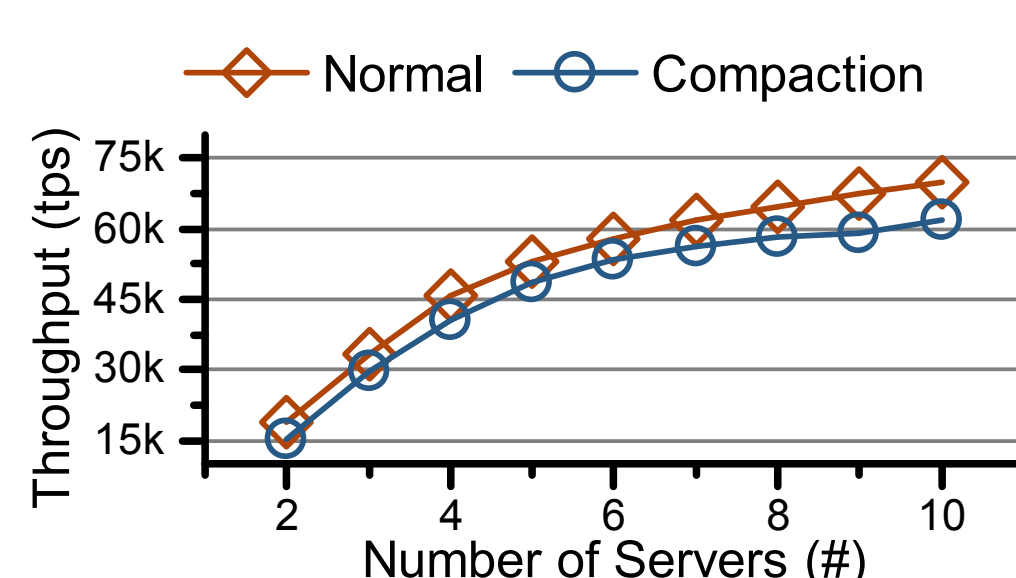
(a) TPC-C



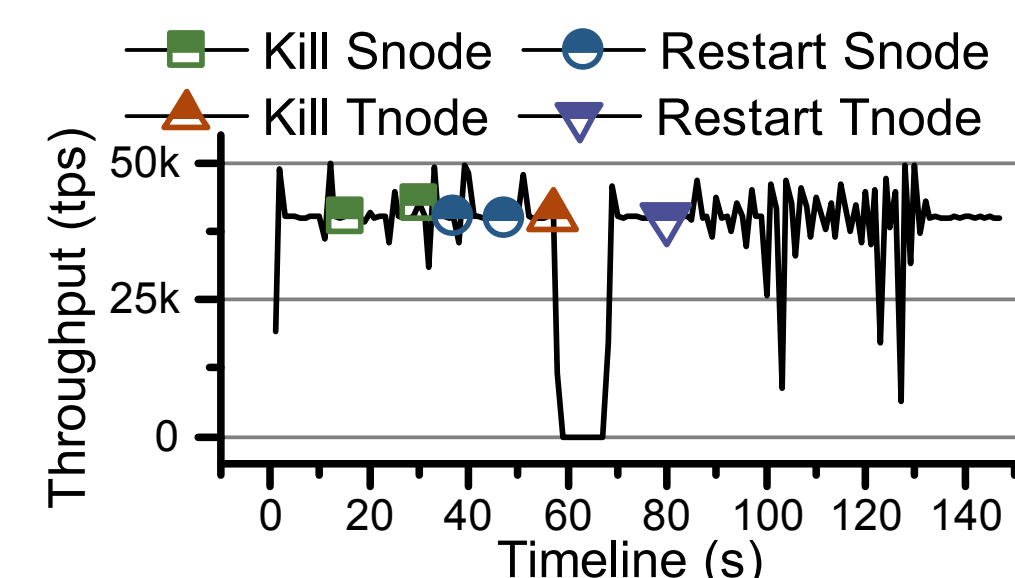
(b) Smallbank



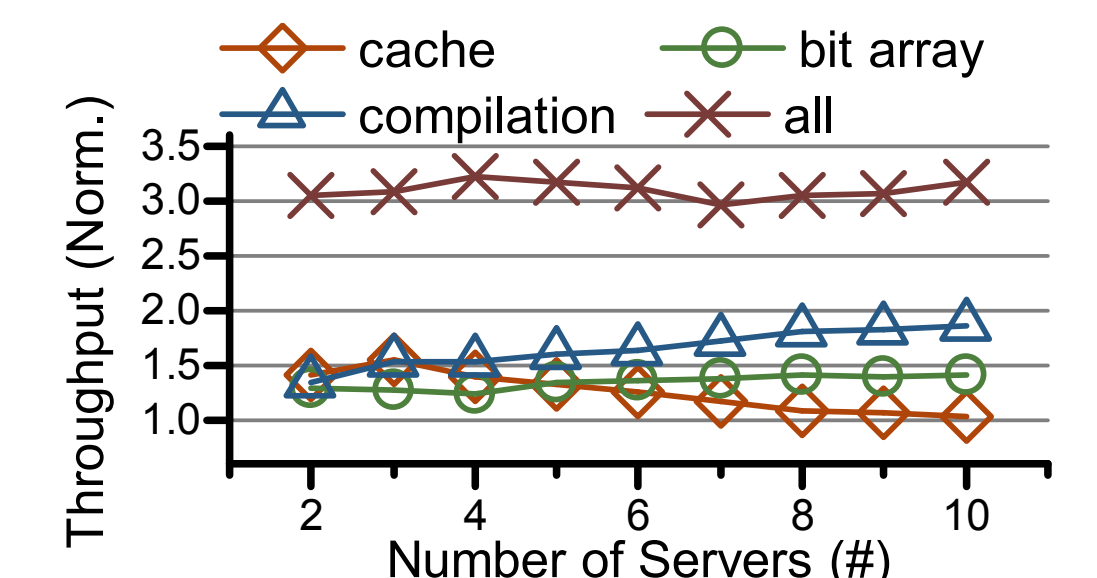
(c) E-commerce



(d) Compaction



(e) Recovery



(f) Data access optimizations