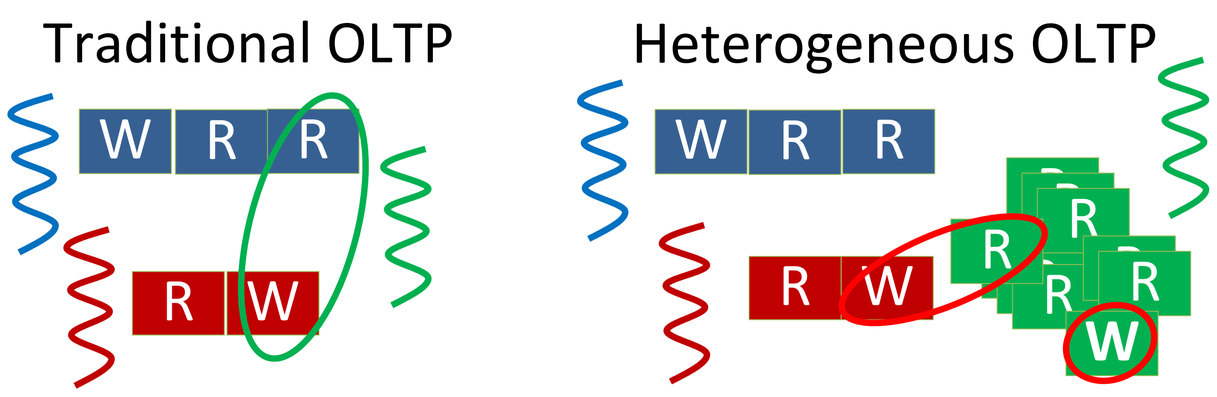
ERMIA: Fast Memory-Optimized Database System for Heterogeneous Workloads

https://github.com/ermia-db/ermia

Kangnyeon Kim, Tianzheng Wang — *University of Toronto*
Ryan Johnson — *LogicBlox*
Ippokratis Pandis — *Amazon Web Services*

What? ***Heterogeneous workloads*** are coming, but existing MMDBMS isn't good at them
Why? "Wrong" concurrency control in use: ***aborts*** too many ***read-mostly*** transactions
How? ***Fair and robust CC*** (snapshot isolation + certifier) + ***scalable physical layer***

Read-mostly analytical components



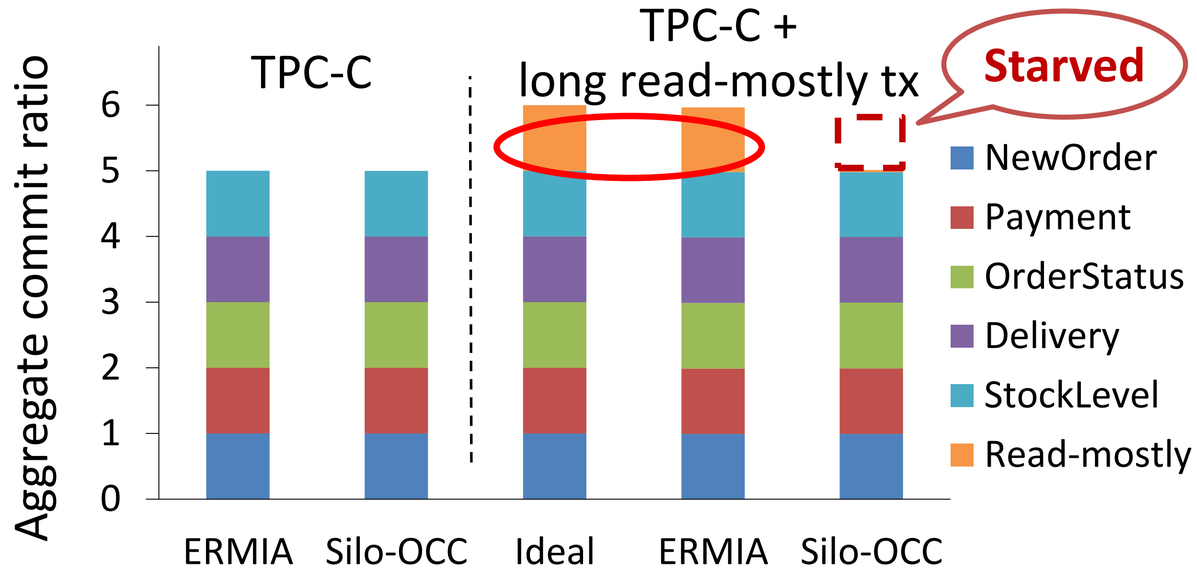
Not only…

- Short write-intensive
- Short read-only

But also:

- Short write-intensive
- Longer read-***mostly***

OCC: not always the best



Robust CC needed for heterogeneous OLTP

ERMIA = Snapshot Isolation + Serial Safety Net + Scalable Physical Layer

Fair and robust logical layer

❶ Read-friendly snapshot isolation
- Abort on write-write conflicts, preserves reads

❷ Serializability with Serial Safety Net *
- Cheap certifier on top of any CC >= RC (e.g. SI)
- Maintains fairness and robustness

* *T. Wang, R. Johnson, A. Fekete, I. Pandis. "**S**erial **S**afety **N**et: Efficient Concurrency Control on Modern Hardware", DaMoN '15*

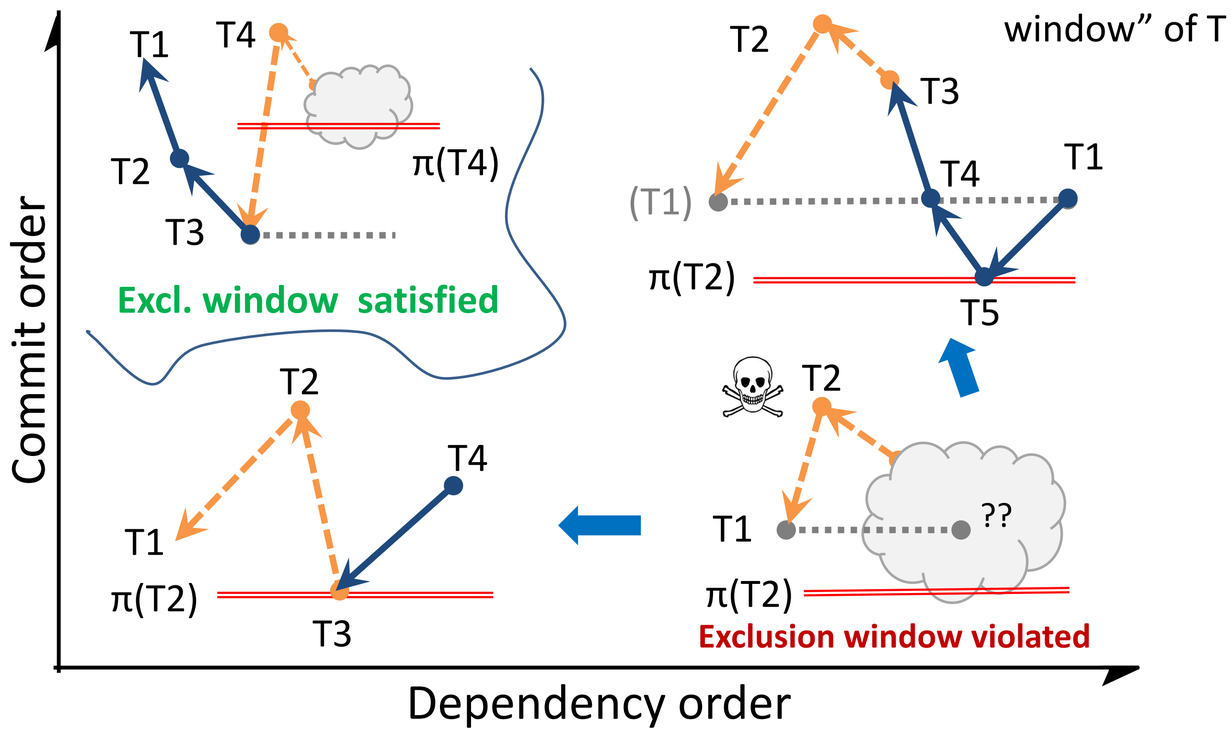
Scalable physical layer

❶ Minimal global communication
- One atomic-fetch-add per tx for global ordering
- Eases implementation of snapshot isolation
- Simplifies logging/recovery

❷ Easy maintenance via indirection
- Fast recovery, single-hop index update, etc.

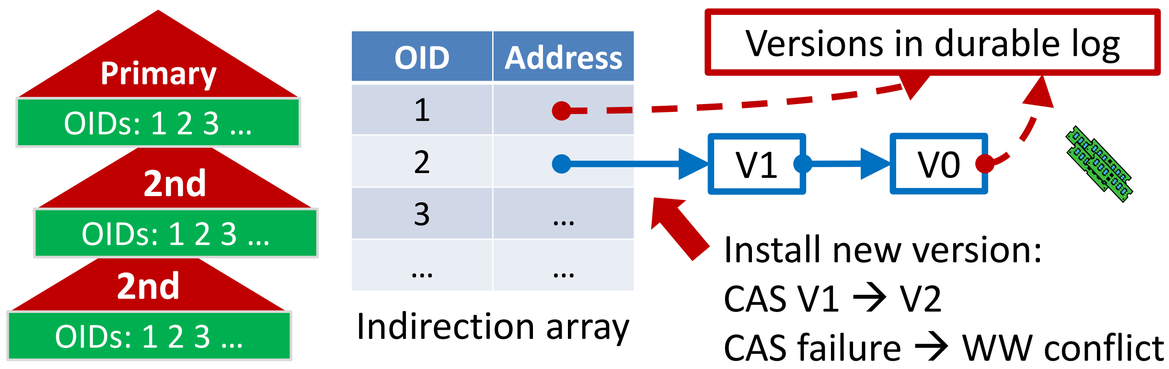
Serial Safety Net

- **def:** c(T) = T's commit time, π(T) = ***earliest*** successor
- **Forbid** any pred P with π(T) ≤ c(P) ≤ c(T) → "Exclusion window" of T



Indirection array

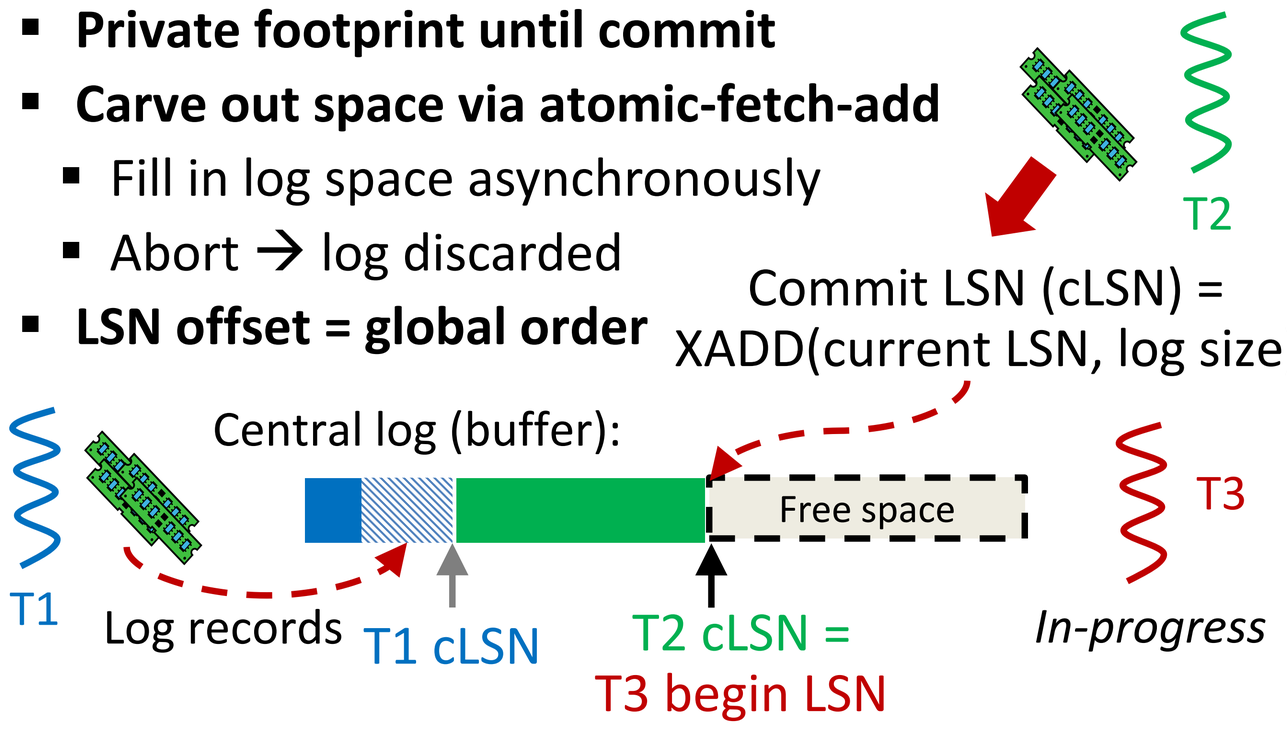
- **Object IDs** (instead of pointers) at leaf level
- **Updates:** no index update needed



- **Recovery:** load header information only

Scalable centralized redo logging

- **Private footprint until commit**
- **Carve out space via atomic-fetch-add**
  - Fill in log space asynchronously
  - Abort → log discarded
- **LSN offset = global order**



Robust to "convenient" & real workloads

HW: 4-socket 6-core Intel E7-4807, 64GB RAM