**Query Fresh: Log Shipping on Steroids**

Tianzheng Wang*  Simon Fraser University  
Ryan Johnson, Ippokratis Pandis  Amazon Web Services

Part of ERMIA: [https://github.com/ermia-db/ermia](https://github.com/ermia-db/ermia)

---

**What?** Hot standby solutions often give *stale reads* with no strong safety

**Why?** Slow network, often-serial replay, and *data redundancy* (log + the “real” DB)

**How?** *Append-only storage* (fast replay) + RDMA over NVRAM (fast log shipping)

---

**High availability by log shipping**

**Primary:** Read + Write  
**Backup(s):** Read + Failover

---

**Infeasible: synchronous log shipping**

---

**The freshness gap**

```
Primary: 9:40 $0  
Backup(s): 9:41 $50  
9:41 $0  
9:42 $0  
...  
9:50 $50
```

**Key reasons:**
- Asynchronous log shipping due to slow network
- Heavyweight (serial) replay due to dual-copy architecture

---

**Existing approaches vs. Query Fresh**

---

**Query Fresh = Append-only storage + Fast RDMA over NVRAM**

---

**Append-only storage for log shipping**

- LSN == address in the log
- Redo-only logging
- Log records == data records

**Log buffer in NVRAM (NV-DIMMs or 3D XPoint)**
- RDMA over persistent log buffers
- Fast persistence – no storage I/O on critical path

---

**Leveraging modern hardware**

- RDMA over fast network (e.g., InfiniBand)
  - Network no longer the slowest part
  - Enables synchronous log shipping
- Log buffer in NVRAM (NV-DIMMs or 3D XPoint)
  - RDMA over persistent log buffers
  - Fast persistence – no storage I/O on critical path

---

**Single-copy + quick replay = fresh reads**

- Append-copy storage: Log == Database
  - Index: key ➔ permanent record ID (RID)
  - Indirection array: RID ➔ record address
- Fast replay: simply set up indirection arrays
  - No record creation, no index ops (except inserts)
  - Parallel and reuse existing recovery machinery

---

**Safe, fast primary and fresh backups**

8 x 16-core (2-socket) Intel E5-2650, 64GB RAM  
56Gbps InfiniBand RDMA and 10Gbps Ethernet TCP

**Primary: full TPC-C**

---

**Network saturated**  
56Gbps BW / 11.37Gbps logs  
~4.9 backups

---

**Backups:** Read-only StockLevel and OrderStatus

---

**4 replay threads only**  
12 query threads / 16 total threads = 75% utilization