Let’s Talk About Storage & Recovery Methods for Non-Volatile Memory OLTP Database Systems

Andy Pavlo + Joy Arulraj
Carnegie Mellon University
Winter 2013: First Blood

- Initial evaluation of existing DBMSs on Intel NVM SDV
- Results published ADMS@VLDB’14
MySQL vs. H-Store

90% Reads / 10% Writes

50% Reads / 50% Writes
Summer 2014: First Blood, Part II

• Evaluate storage and recovery methods for NVM.

• Preparing SIGMOD’14 submission.
DBMS Testbed

• Custom lightweight DBMS.
  – Uses NUMA & PMFS interfaces.
  – No volatile DRAM.

• Partition-based locking CC.

• Pluggable architecture:
  – Supports different storage engines.
Engine #1 — In-place Updates

- Apply change to tuples directly.
  - VoltDB with ARIES.
  - Table storage + write-ahead log.
  - STX B+Tree
Engine #2 — Copy-on-Write Updates

• Make new copy before updating:
  – Shadow paging using LMDB Persistent B+Tree.
  – No logging.
  – Background garbage collection.
Engine #3 — Log-based Updates

- Changes only written to log.
  - Based on LevelDB’s LSM.
  - No table storage.
  - Background level compaction.
# Storage Engines

<table>
<thead>
<tr>
<th>Storage Style</th>
<th>Table Storage</th>
<th>Logging</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Place</td>
<td>Yes</td>
<td>Yes</td>
<td>VoltDB</td>
</tr>
<tr>
<td>Copy-on-Write</td>
<td>Yes</td>
<td>No</td>
<td>LMDB</td>
</tr>
<tr>
<td>Log-based</td>
<td>No</td>
<td>Yes</td>
<td>LevelDB</td>
</tr>
</tbody>
</table>
NVM Optimized Engines

• Refactored engines to be “pointer-oriented”.

• Extended Intel’s *libpmem* allocation library.
  – Added arena-based allocation.
  – Significantly improved throughput.
Experimental Evaluation

• Yahoo! Cloud Serving Benchmark:
  – 2 million records (~2GB)
  – Two workload mixtures
  – Two skew settings
  – 1 million transactions
Experimental Evaluation

• NVM Latency Configuration:
  – 2x DRAM (≈200ns)
  – 8x results not shown.

• 8 partitions on 8 cores.
Throughput

90% Reads / 10% Writes

50% Reads / 50% Writes
NVM Reads/Writes

90% Reads / 10% Writes

50% Reads / 50% Writes
Recovery Time

Zero Recovery
Discussion

• NVM engines outperforms “traditional” engines:
  – Higher throughput
  – Reduced wear on device.

• In-place performs best overall.
Fall 2014: N-Store

- First DBMS for NVM-only operating environment.
- OLTP/OLAP hybrid
  - Column-store that supports fast in-place updates.
- Indexing + Many-Core
END

@andy_pavlo