Rich Uhlig is going to be there. He doesn't want to see that stuff.

What?

He has two kids. He's a family man.

OK I'll see what I can do.
Andy Pavlo’s
Mostly Professional and Collected Presentation about
Getting Down & Dirty in OLTP Databases with Intel’s NVM SDV

September 25, 2015
The Story So Far

2014: Comparing Existing DBMSs

2015: Evaluating Storage Architectures
2014: Existing DBMSs

Comparison of disk vs. main-memory DBMSs running on Intel NVM SDV.

Found that logging is (still) the main bottleneck in both systems.

Paper: ADMS @ VLDB’14
2015: Storage Architectures

Evaluated storage and recovery methods for OLTP DBMSs.

Developed NVM-optimized methods that achieve 5.5x better throughput with 2x fewer writes.

Paper: SIGMOD’15
### 2015: Storage Architectures

<table>
<thead>
<tr>
<th>In-Place</th>
<th>Table Storage</th>
<th>Logging</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<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>RocksDB</td>
</tr>
</tbody>
</table>
YCSB :: 10/90 RW :: 2x Latency

Throughput (txn/sec)

- In-Place
- Copy-on-Write
- Log-Structured

Traditional vs NVM-Optimized
YCSB :: 10/90 RW :: 2x Latency

- **Traditional**
- **NVM-Optimized**

<table>
<thead>
<tr>
<th>NVM Stores</th>
<th>In-Place</th>
<th>Copy-on-Write</th>
<th>Log-Structured</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>150,000,000</td>
<td>100,000,000</td>
</tr>
<tr>
<td>50,000,000</td>
<td></td>
<td>190,000,000</td>
<td>150,000,000</td>
</tr>
<tr>
<td>100,000,000</td>
<td></td>
<td>200,000,000</td>
<td>180,000,000</td>
</tr>
<tr>
<td>150,000,000</td>
<td></td>
<td>210,000,000</td>
<td>190,000,000</td>
</tr>
<tr>
<td>200,000,000</td>
<td></td>
<td>220,000,000</td>
<td>210,000,000</td>
</tr>
</tbody>
</table>
PCOMMIT Evaluation

Weakly-ordered sync primitive that retains data in the flushed cached lines.

Emulated with RDTSC and PAUSE instructions on NVM SDV.

Summer 2015: ~10,000 PCOMMIT invocations per second per CPU core.
YCSB // In-Place Engine

Throughput (txn/sec)

Sync Primitive Latency (ns)

- 90/10 RW
- 50/50 RW
- 10/90 RW

 ISTC BIG DATA
New Stuff

NVM vs. SSD

Multi-level Anti-Caching

DRAM+NVM storage manager
NVM vs. SSD

Two-level Storage Hierarchy

Disk-oriented vs. Memory-Oriented
- Caching (MySQL)
- Anti-caching (H-Store)
Disk vs. Memory Oriented DBMSs

Caching

DRAM → Durable Storage

Anti-Caching

Durable Storage → DRAM
YCSB :: 90/10 RW :: 4x Latency

- Anti-Caching (NVM)
- Anti-Caching (SSD)
- Caching (NVM)
- Caching (SSD)

Throughput (txns/sec):
- I/O Bound
- CPU Bound

Skew Factor (Low→High):
- 0.5
- 0.75
- 1
- 1.25

Fewer DRAM Misses
Lock Contention
YCSB :: Byte-Addressable Access

Throughput (txn/sec)

Block Size

Anti-Caching (NVM)

Anti-Caching (SSD)

Single Tuple

Merging Overhead

ISTC Big Data
Voter :: 4x Latency

Anti-Caching (NVM)

Anti-Caching (SSD)

DRAM Evictions

Elapsed Time

Throughput (txn)

/sec

0

75000

150000

225000

0

75000

150000

225000

16
Multi-Level Anti-Caching

OLTP Workload

Current Investigation:
- Eviction Policies
- Retrieval Policies
- Access Interfaces
- Data Organization
Multi-Level Anti-Caching

OLTP Workload

- Lin Ma
- Michael Giardino
- Sam Zhao
- Joy Arulraj
- Dana Van Aken
- Prashanth Menon
- Ugur Cetintemel
- Justin DeBrabant
- Kshitij Doshi
- Subramanya R. Dulloor
- Aaron Elmore
- Samuel Madden
- David Maier
- Michael Kaminsky
- Tim Kraska
- Jeff Parkhurst
- Andrew Pavlo
- Michael Stonebraker
- Nesime Tatbul
- Donald Trump
- Kristin Tufte
- Stanley Zdonik

1 Brown, 2 Intel Labs, 3 MIT, 4 Portland State University, 5 CMU, 6 Univ. of Chicago
Voter :: Multi-Level :: 2x Latency

Transactions per Second

Memory (MB)

Slow as SSD
DRAM+NVM DBMS

Building a new storage manager for our new DBMS that will seamlessly incorporate NVM as an extension to its address space.

Upper-levels of the system are oblivious to “true” location of data.
END
@ANDY_PAVLO