The Magic and Mystery of In-Memory Apps

Taufik Ma – Industry Insight
Shaun Walsh - Marketeer
Contents

- The Use In Memory Applications?
- Evolution towards & Role of In-Memory Computing
- Role of Storage in In-memory solutions
- Customer Trends
- Emerging Technologies & Some Predictions
- Summary
Magic and In-Memory Applications

Shaun Walsh - Marketeer
Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke
The Evolution of Storage Tiers

NVM will Accelerate Both Meta-Data & Application Data
NVDIMM Acceleration Segments

<table>
<thead>
<tr>
<th>NVDIMM Type</th>
<th>Presentation</th>
<th>Access Method</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-N</td>
<td>DRAM</td>
<td>Byte</td>
<td>Consistent</td>
</tr>
<tr>
<td>-F</td>
<td>Storage</td>
<td>Block</td>
<td>Variable</td>
</tr>
<tr>
<td>-P</td>
<td>DRAM and/or Storage</td>
<td>Byte &amp; Block</td>
<td>Variable</td>
</tr>
</tbody>
</table>

- Data Base Log Files
- Clustering
- Cache Synchronization
- In-Memory DBs
- MemCacheD
- RAID
- De-Dupe

NVDIMM-N

NVDIMM-F
NVDIMM-P
3D-XPoint

Meta Data Acceleration
NVM-DIMM – fills growing DRAM-NAND gap

- In Memory Applications are driving a new class of Storage Class Memory (SMC)

- Latency and persistence are as important as absolute bandwidth

- Byte and Block address flexibility is vital to scaling In-Memory Applications (IMA)

Source: Objective Analysis, 2015
The Future of Business Intelligence

Bandwidth & Capacity
• Old performance was data rates (GB/s) & capacity (TB)
• Store Everything, Sort Later
• Higher Cost, Slow Decisions

Latency & Persistence
• Real-Time is Business Critical
• Major Players Driving NMV
• Store the Vital & Analyze now

Latency and Persistence are the new value currency for real-time applications & storage
Procter & Gamble - Real-Time Reporting & Business Decisions

400% Increase in decision support systems performance

35,000 Retail, supply chain and business users supported

55% Reduced database from 36TB to 16TB all in memory

P&G achieved faster, more reliable reporting and analytics


© 2015 G2M COMMUNICATIONS. All rights reserved.
McLaren Group – Faster Formula 1

- Faster and more consistent lap times
- Improved down force for better grip
- Real-time telemetric analysis
- More World Championships
The Art and Science of In Memory Applications

Taufik Ma
Industry Insight
Evolution of Databases & Analytics

1980s
- RDBMS
- Operational (OLTP, ERM)
- Data Warehousing (Data mining, DSS, Analytics)

1990s
- RDBMS
- Oracle, MS SQL, Sybase
- EDW/OLAP
- Teradata, Oracle, SAS, etc

2000s-2015
- RDBMS
- MySQL, Postgres
- NoSQL
- MongoDB, Cassandra
- Hadoop
- MapReduce, HBase
- EDW/OLAP
- IBM Netezza, EMC Greenplum
Ongoing Evolution & Specialization…

Real-time, Online Operations
- RDBMS
- NoSQL

Batch, Offline Analytics
- EDW/OLAP
- Hadoop

Structured Data, Relational
Unstructured, Schema-less

OLTP, ERM
- Purchases, clicks
- User profiles, reviews
- Content Management

User Segmentation
- Daily offer recommendation
- Ad serving engine
- Fraud Detection
Ongoing Evolution & Specialization…

Real-time, Online Operations
- **RDBMS**
  - Structured Data, Relational
- **NoSQL**
  - Unstructured, Schema-less

Real-time analytics
- **In-Memory Database**
  - Hana, Exalytics, MemSQL, etc
- **In-Mem Data Processing**
  - Spark, Hadoop in-mem

Batch, Offline Analytics
- **EDW/OLAP**
  - OLTP, ERM
  - Purchases, clicks
  - User profiles, reviews
  - Content Management
- **Hadoop**
  - Financial risk/value analysis
  - Fraud Prevention
  - Real-time recommendations
  - Profitability analysis
  - User Segmentation
  - Daily offer recommendation
  - Ad serving engine
  - Fraud Detection
## Multiple Tools Within A Customer

<table>
<thead>
<tr>
<th></th>
<th>Customer Profiles (G2M Survey)</th>
<th>$500M+ Retail</th>
<th>$500M+ Pharma</th>
<th>$1B+ Manufacturing</th>
<th>$1B+ Pharma</th>
<th>$1B+ SaaS</th>
<th>$250M+ Healthcare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>MongoDB</td>
<td>Yes</td>
<td>No plans</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No plans</td>
<td>No plans</td>
</tr>
<tr>
<td>Spark</td>
<td>Yes</td>
<td>No plans</td>
<td>Considering</td>
<td>Yes, in 6 months</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes, in 6 months</td>
</tr>
<tr>
<td>SAP HANA</td>
<td>No plans</td>
<td>Yes</td>
<td>Considering</td>
<td>Yes</td>
<td>No plans</td>
<td>No plans</td>
<td>Considering</td>
</tr>
<tr>
<td>Microsoft Hekaton</td>
<td>No plans</td>
<td>No plans</td>
<td>Considering</td>
<td>Yes, in 6 months</td>
<td>No plans</td>
<td>No plans</td>
<td>Yes, in 12 months</td>
</tr>
<tr>
<td>memSQL</td>
<td>No plans</td>
<td>No plans</td>
<td>Considering</td>
<td>Yes, in 6 months</td>
<td>No plans</td>
<td>No plans</td>
<td>Yes, in 12+ months</td>
</tr>
<tr>
<td>Oracle Exalytics</td>
<td>No plans</td>
<td>No plans</td>
<td>Yes</td>
<td>Yes</td>
<td>No plans</td>
<td>No plans</td>
<td>Yes, in 12+ months</td>
</tr>
</tbody>
</table>

“Specialized Tools for Specific Needs”
(Or “Too Many Data Islands”?)
How many in-memory applications do you (or will you) run?

- 1-5
- 6-10
- More than 10
### Key Enabler of In-Memory Computing: Today’s Technologies

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Time to get data</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU L1 cache</td>
<td>0.001 usec</td>
</tr>
<tr>
<td>DRAM</td>
<td>0.01 usec</td>
</tr>
<tr>
<td>NAND</td>
<td>100 usec</td>
</tr>
<tr>
<td>HDD</td>
<td>10,000 usec</td>
</tr>
</tbody>
</table>

On a human scale…

If I complete 50 operations in 50 seconds, then have to wait for data…

- **DRAM** = getting food from the fridge (10’s of seconds)
- **NAND** = taking the day off
- **HDDs** = hiking the Pacific Coast Trail (months)
## Performance Comes at a Price

<table>
<thead>
<tr>
<th>Storage</th>
<th>Time to get data</th>
<th>Price / GB</th>
<th>Cost for 100TB</th>
<th># 2U Servers Req’d to Hold 100TB*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>0.01 usec</td>
<td>$5.60</td>
<td>$560,000</td>
<td>130</td>
</tr>
<tr>
<td>NAND</td>
<td>100 usec</td>
<td>$0.35</td>
<td>$35,000</td>
<td>5</td>
</tr>
<tr>
<td>HDD</td>
<td>10,000 usec</td>
<td>$0.03</td>
<td>$3,000</td>
<td>2-3</td>
</tr>
</tbody>
</table>

**DRAM**
- 32G DIMM for $179 ea, Samsung Registered DDR4, M393A4K40BB0-CPB0
- 3125 x 32G DIMMs

**NAND**
- 2.5” 1TB SSD, $350 ea, Intel 540S
- 100 x 2.5” 1TB SSD

**HDD**
- 3.5” 4TB SATA HDD for $120 ea, Seagate ST4000DM000
- 25 x 3.5” 4TB SATA HDD

*Assuming 24 DIMM slots, 24x 2.5” drives or 12x 3.5” drives
Location of Data & Tasks

**Hadoop: MapReduce / HDFS**
- **Input File**
- **Chunks**
- **Parallel Tasks**
  - DISK 1
  - MEM 2
  - MEM 3
- **Sends tasks to data nodes**
- **JobTracker / Name Node**

**Spark / Tachyon**
- **Input File**
- **Partitions (RDDs)**
- **Parallel Tasks**
  - MEM 1
  - MEM 2
  - MEM 3
- **Sends tasks to worker nodes**
- **Spark Driver**

**SAP Hana**
- **Input File**
- **User Partitioning**
- **Local Tasks**
  - Master
  - Slave(s)
  - Standby
Surviving Failures

**Hadoop: MapReduce / HDFS**
- Input Files: 1 2 3
- Chunks
- Parallel Tasks: 1 2 3
- DISK
- 3-fold Replication

**Spark / Tachyon**
- Input Files: 1 2 3
- Partitions (RDDs)
- Parallel Tasks: 1 2 3
- MEM
- To persistent storage
- Lineage: Record of transformations that created an RDD from its “parent”

**SAP Hana**
- Input Files: 1 2
- User Partitioning
- Local Tasks: 1 2
- Ext Storage
- Logs & savepoints
No such thing as 100% In-Memory

Hadoop: MapReduce / HDFS

- Input Files
  - 1
  - 2
  - 3
- Chunks
  - a
  - b
  - c
- Parallel Tasks
  - RAM
    - 1
  - DISK
    - 2
  - SSD
    - 3

Spark / Tachyon

- Input Files
  - 1
  - 2
  - 3
- Partitions (RDDs)
  - a
  - b
  - c
- Parallel Tasks
  - MEM
    - 1
  - SSD
    - 2
  - HDD
    - 3

SAP Hana

- Input Files
  - 1
  - 2
  - a
  - b
- User Partitioning

HDFS2.0 Heterogeneous Storage

- Storage Types & Policies
  - Files/directories assigned policies
  - (e.g. Lazy_persist, All_SSD)

Tachyon Tiered Storage

(for Off_heap Spark RDDs)

- Auto or manual

SAP HANA Dynamic Tiering

- Data spec’d as either Hot or Warm

* ARCHIVE tier not shown
Customer In-Memory Computing Trends (based on G2M survey)

### SIZE
- Cluster sizes similar to big data solutions
  - ½ respondents > 500 servers, 1/3 at >50
  - And not just for Spark
- With datasets that fit available DRAM capacity
  - 1/3 at >100TB, 1/3 at >10TB

### GROWTH
- ~Half with 10-20%+/yr dataset growth
- Majority use/want tier-ing when dataset > DRAM
  - Only minority would rely on scale-out only
- Mixed on whether tier-ing should be transparent or not
  - Some want it transparent to developer; Rest want developer to have control via policy

### EFFICIENCY
- ~Half believe “my storage capacity forces me to have more compute capacity then I need”
- Majority have or have plans for consolidated data silos
  - OLTP+IMDB, Spark+Hadoop, NoSQL+Hadoop
Emerging Technologies: High-speed Fabrics & Disaggregated Storage

- Data Center Ethernet speeds ramping faster than drive speeds: 10/25/40/50/100G
- RDMA-over-Ethernet technologies
- Multi-host PCIe fabrics emerging (e.g. OCP Lightning) albeit w/ less scalability

- Ethernet or PCIe based fabric
- DAS-like performance Local or SAN
- Map any drive to any host
- Scale each storage tier separately from compute
- Early proof points: EMC DSSD, SanDisk InfiniFlash, DriveScale

![Diagram showing high-speed fabrics and storage tiers](image)
## Emerging Technologies: Storage Class Memory

<table>
<thead>
<tr>
<th>Storage</th>
<th>Persistence</th>
<th>Time to access data</th>
<th>Price / GB</th>
<th>Cost for 100TB</th>
<th># 2U Servers Req’d to Hold 100TB*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>N</td>
<td>10ns+</td>
<td>$5.60</td>
<td>$560,000</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3125 x 32G DIMMs</td>
<td></td>
</tr>
<tr>
<td>NV-DIMM -N</td>
<td>Y</td>
<td>10ns+</td>
<td>$10+</td>
<td>$1,000,000+</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>If 2X+ DRAM</td>
<td></td>
<td>16G NVDIMM, supercap</td>
</tr>
<tr>
<td>3DXP DIMM</td>
<td></td>
<td>100ns Rd 500ns Wr</td>
<td>$2+</td>
<td>$190,000+</td>
<td>~50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>If 1/3+ DRAM</td>
<td></td>
<td>assuming 96 or 128GB DIMMs</td>
</tr>
<tr>
<td>NAND</td>
<td>Y</td>
<td>100 usec</td>
<td>$0.35</td>
<td>$35,000</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.5&quot; 1TB SSD, $350 ea, Intel 540S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDD</td>
<td>Y</td>
<td>10,000 usec</td>
<td>$0.03</td>
<td>$3,000</td>
<td>2-3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.5&quot; 4TB SATA HDD for $120 ea, Seagate ST4000DM000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Assuming 24 DIMM slots, 24x 2.5” drives or 12x 3.5” drives
In-Memory Computing Predictions / Trends

1. **3DXP DIMMs used for “Jumbo Memory”** – value in lower $/GB vs DRAM, not persistence
   - Mix of 3DXP & DRAM DIMMs in server nodes
   - Tier-ing will be tuned to accommodate slower writes & reads
   - Spark, In-mem Hadoop, MemSQL, Hana, etc
   - NV-DIMM –P might have similar adoption but predictable latency is a concern

2. **Increasing use of NVMe SSDs as “Far Memory”** – as next tier (below DRAM/3DXP)
   - Priority on $/TB, not persistence. Resiliency still via Lineage, logs, etc
   - Remove "last-inch" of latency via BLKB (block-layer/kernel bypass) stacks (e.g. EMC libflood, SPDK)
   - Implemented as a fabric-disaggregated cluster to enable efficiency & independent scalability
   - Longer-term, HW-based paging of near-memory to far-memory

3. **Use of “Persistent Memory” for In-Mem computing will evolve**
   - For 3DXP & NV-DIMM –N
   - Industry progress on pmem file systems (Linux, Windows)
   - Does persistence replace or complement lineage/logs?
   - Need low latency replication across nodes (PMoF)
- In-memory solutions growing in adoption – driven by real-time analytics
- Co-existence of structured (e.g. Hana) and unstructured frameworks (e.g. Spark)
- Confluence of big-data & real-time analytics drives increasing adoption of tier-ing
- Newer technologies on horizon will continue to create disruptions to in-memory computing architectures