Removing the I/O Bottleneck in Enterprise Storage

WALTER AMSLER, SENIOR DIRECTOR
HITACHI DATA SYSTEMS

AUGUST 2013
Agenda

- Enterprise Storage
  - Requirements and Characteristics
  - Reengineering for Flash – removing I/O bottlenecks

- Measuring Performance
  - Application Performance Metrics vs Synthetic Benchmarks Numbers

- Summary
Requirements for Storage Systems

EFFECTIVELY ADDRESSING TECHNOLOGY AND BUSINESS CHALLENGES

- What Customers are demanding
  - Reduce cost, optimize service-level delivery via scale-up, dynamic provisioning and tiering across different media types
  - Management abstraction to enable ease of use, speed and automation
  - 24x7xforever application availability, eliminate planned and unplanned outages
  - Reliability, Availability Serviceability (RAS)

- Recent Trends in High-End Storage
  - Storage Subsystems are designed for the Virtual data center
  - Storage Infrastructure is transformed in Storage Services
  - Exploitation of loosely coupled vs tightly coupled Architectures
Characterizing Storage systems

ANOTHER FORM OF RAS: REDUNDANCY, ARCHITECTURE, SCALABILITY

- Storage Architectures and design – different value propositions
  - Modular Architecture vs Enterprise Architecture
  - Component/Site Redundancy

- Performance
  - Time is Money – must cope with peak demands and satisfy strict SLA’s

- Functionality
  - Virtualisation
  - Dynamic Tiering
  - In-System Snapshots and Clones
  - 2DC and 3DC Sync and Asynchronous Replication
  - Rich GUI/CLI Management Capabilities – Ease of Use
Modular vs Enterprise Architecture

BALANCING COST, SCALABILITY, PERFORMANCE AND CAPACITY

Modular-Architecture

Enterprise-Architecture

Applications

Connectivity

Processors

Cache Memory

Backend

Storage Media Types

SSD
FC / SAS
SATA

Storage Pool

Storage Array

Connectivity

Processors

Cache Memory

Backend
Modular storage growth – Scale-Out

ADD MORE OF THE SAME – BUT BEWARE OF ISLANDS

Applications

Connectivity

Processors

Cache Memory

Backend

Storage Media Types
- SSD
- FC / SAS
- SATA

Storage Array

Connectivity

Processors

Cache Memory

Backend

Storage Pool

Loosely coupled Architecture
Enterprise storage growth – Scale-Up

EXPAND CAPACITY, CONNECTIVITY AND PROCESSING POWER

Applications

Connectivity

Processors

Cache Memory

Backend

Storage Media Types

SSD

FC / SAS

SATA

Storage Pool

Storage Pool

Storage Pool

Storage Pool

Tightly coupled Architecture
When things go wrong

FAILURE IMPACT - GOOD ENOUGH VS BULLET PROOF

- Availability depends on failure domains and the choice of component/site redundancy options
  - Bulletproof storage array: [http://www.youtube.com/watch?v=Gnjb1WVkhmU](http://www.youtube.com/watch?v=Gnjb1WVkhmU)
I/O Bottleneck in Enterprise Storage

BUILT FOR FLASH FROM THE GROUND UP VS RE-ENGINEERED

- Traditional Storage Arrays
  - originally designed for hundreds, then thousands of HDD’s
  - Ever larger DRAM Cache and sophisticated Algorithms mitigate/hide HDD performance characteristics
    - Works great for sequential read/write
    - Works very well for Random I/O with good Locality of reference

- The IO Gap
  - Moore’s Law - processor speed has increased dramatically
  - HDD Speed (Seek and RPM) has virtually stayed the same
  - server virtualization randomizes I/O, LOR is lost, aka «I/O Blender»

- The Emergence of Flash demands a new approach
Read IO Operation; Cache Hits and Misses

Data found in cache = ‘Hit’

Performance is media independent

No data found = ‘Miss’

S3 Sweet Spot: SSD vs HDD = 10 : 1
30+ fundamental software changes to turbo-charge performance with Flash
- New “express” I/O processing
- New Cache Slot Allocation method
- Reduced ucode Overhead and path length

Significant performance impacts
- Up to 65% reduction in response time
- Up to 4X Random IO scalability

Non-disruptive installation and transparent to current applications
Flash Memory Summit 2013
Santa Clara, CA

Flash Acceleration Impact for all flash array

145 PATENTS RELATED TO HITACHI FLASH TECHNOLOGY

- Backend Codepath reduction, logic and ASIC optimization
  - Version 1: Basic Design for HDD – non optimized
  - Version 2: BE/FE Job Integration, Cache Buffer Slot Management
  - Version 3: use DXBF, avoid CTL to CTL communication Improve CPU L1 Cache Hit Rate for Instructions

![Graph showing read processing time and throughput for different versions of HUS VM V01, V02, and V03.]

- Throughput (KIOPS): 240KIOPS, 500KIOPS, 1000KIOPS
- Read Processing Time (%): HUS VM V01, HUS VM V02, HUS VM V03
Latency and what does it really mean

MEASURING PERFORMANCE - RELEVANCE TO YOUR BUSINESS

- **Vendor Provided Measurement Data**
  - Objective is to show «champion numbers»
  - Customers need to have a complete understanding of what was measured and how, for example:
    - 80 usec Latency: single 512Byte Block Read measured at Fibre Channel Port with a Fibre Channel Analyzer
    - 1 Million IOPS: 4KB Random Reads measured by IOMETER
  - Interesting, but not relevant from an application perspective

- **Need a different approach**
  - Include and consider all the different technology layers of entire platform
  - Example: Oracle Database Platform Architecture
**Oracle Database Platform Architecture**

**Application Network** (IP-based)
Bandwidth, latency during remote database mirroring (sync, async) due to switches and sql*net and tcp/ip stack (frame size, ...).

**Oracle Database**
Different versions, patches and options, about hundred configuration parameters.

**Volume & File Management**
Different volume managers (VxVM, ASM) and file systems (UFS, VxFS, ext3, JFS, ZFS, raw devices), different I/O methods (async, direct), a lot of config parameters (#LUNS, queue depth, max i/o unit), software striping and/or mirroring, multipathing.

**Storage Network** (IB-, FC- or IP-based)
Bandwidth, latency during remote storage mirroring (sync, async) due to switches, hubs and distance.

**Server & Operating System**
Different server systems, processors and CPU architectures, (x86, IA-64, UltraSparc, SPARC64, Power), #cores, multithreading, main memory, bus architecture. Different operating systems and patches, over hundred configuration parameters, virtualization of resources.

**Storage System**
Different storage systems, storage tiers and storage technology: spindle count and speed, RAID management, cache management, server interface technology, storage system options like remote copy, hardware striping and/or mirroring, virtualization of resources.
Measuring Oracle Application Performance

Measuring Hardware

Measuring complex Application I/O or customer reality

Measuring Server and Storage and «Mindless» IO
### Benchware Approach

Library of Oracle benchmark tests - implemented in PL/SQL, Java and SQL

<table>
<thead>
<tr>
<th>CPU Performance</th>
<th>OLTP systems</th>
<th>DWH systems</th>
<th>Metrics Efficiency</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU-bound Oracle operations</td>
<td>★★</td>
<td>★★</td>
<td>throughput</td>
<td>[ops]</td>
</tr>
<tr>
<td>All operations in Level 1, 2, 3 CPU cache</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server Performance</th>
<th>OLTP systems</th>
<th>DWH systems</th>
<th>Metrics Efficiency</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU-bound Oracle operations</td>
<td>★★</td>
<td>★★</td>
<td>throughput</td>
<td>[bps] [tps] [rps]</td>
</tr>
<tr>
<td>All operations in RAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database Performance</th>
<th>OLTP systems</th>
<th>DWH systems</th>
<th>Metrics Efficiency</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed resource usage: CPU, memory, storage</td>
<td>★★★</td>
<td>★★</td>
<td>throughput</td>
<td>[bps] [tps] [rps]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Storage Performance</th>
<th>OLTP systems</th>
<th>DWH systems</th>
<th>Metrics Efficiency</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O-bound Oracle operations</td>
<td>★★</td>
<td>★★★</td>
<td>speed</td>
<td>[rps] [tps] [qpm]</td>
</tr>
</tbody>
</table>

| **•** sequential I/O | | | throughput service time virtualization tiering | [MBps] [GBps] [iops] [ms] |
|----------------------|| | | | |
| 1 MByte, read | write | | | |

| **•** random I/O | | | | |
|------------------|| | | |
| 32 block size, read | write | | | |

**[s]** seconds  
**[ms]** milli seconds \((10^{-3})\)  
**[us]** micro seconds \((10^{-6})\)  
**[ns]** nano seconds \((10^{-9})\)  

**[bps]** buffers per second  
**[rps]** rows per second  
**[tps]** transactions per second  
**[ops]** operations per second  

**[MBps]** mega bytes per second  
**[GBps]** giga bytes per second  
**[iops]** i/o operations per second  
**[qpm]** queries per minute  

**[less important]**  
**[important]**  
**[very important]**
Measuring Datawarehouse Workload

SEQUENTIAL READ, MULTIPLE PROCESSES – TYPICAL FOR DWH

Number at measuring point: Avg service time within Oracle for 1 MByte I/O read request (128 x 8 kByte)

Throughput in [MBps]

Degree of parallelism (dop)
Measuring OLTP Workload

8KB RANDOM READ; 100% CACHE MISS - TYPICAL FOR OLTP

Number at measuring point:
Avg service time within Oracle for 8 kByte single block random read

HDS VSP with 16 FMDs
- single instance
- multi instance
What does it mean to your business?

KEY PERFORMANCE METRICS LEAD TO SERVICE LEVEL AGREEMENTS

- The measured server/storage platform will deliver:
  - 8GB/sec sequential Read throughput for your DWH
  - 250,000 8KB Random Reads with Zero Cache Hits for your OLTP application with a Response Time of less than 3 Milliseconds

- Note: Oracle Measurements for Random Read IO
  - Oracle currently does not understand «Microseconds»
  - Response Time for Random Read is reported in Milliseconds, and data is rounded e.g. 0 MS or 1 MS

- R/T for high Random Read I/O Rates generally at 1-4 MS
  - This also applies to All Flash Appliances/Arrays
Summary

- Enterprise Storage today has a lot to offer
  - RAS: Reliability, Availability, Serviceability
  - Superior Performance
  - Seamless Scale-Up Architecture

- Flash Storage Exploitation
  - Value of Re-engineering equals «Built from scratch»
  - In addition you get the functionality and EoU you need

- Performance and Latency Claims
  - Must understand what is being measured and how
  - The key is the mileage you get for your application!
Questions and discussion
Thank you