VDBENCH Overview


Steven A. Johnson

SNIA Emerald™ Training
SNIA Emerald Power Efficiency Measurement Specification, for use in EPA ENERGY STAR®,

July 14-17, 2014
Agenda

- Introduction to VDBENCH – iodriver
- Purpose of VDBENCH for EPA program
- Performance 101
  - Overview of things that effect performance
- Overview of VDBENCH scripts format
  - SD, WD and RD parameters
- Detailed discussion of SD, WD, and RD parameters
- Discussion of the output of VDBENCH
Quick overview of performance terms

- **Scale-ability** – able to increase in throughput or performance with increasing application demands
- **Utilization** – How busy a resource is during a period of time. Generally expressed as a percent from 0 – 100
- **Service time** – Generally the actual time something take for a specific task
- **Response time** – Usually considered Service time plus queueing time for resource
- **Latency** – the period of time one component in a system is waiting for another component
- **Data transfer time** – The latency required to transfer the requested data from a resource
- **Queueing** – The natural process of things lining up to be services
- **Queueing Theory** – The Mathematical study of Queueing systems
- **Queue depth** – Frequently associated with number of outstanding IOs to a Storage System
- **Cache** – Placing frequently used things in an easily accessible place. For computers, placing data in a place that has much faster access time.
Performance terms (cont)

- **Cache hit** – Information the system is looking for is located in high-speed memory.
- **Cache miss** – Information was not in high-speed memory and had to be found on a slower device.
- **Sequential** – Type of workload that can read or write something one block after another.
- **Logically sequential** – An application may read or write a file from beginning to end.
- **Physically sequential** – While an application may think it is reading physically sequential, generally this is not the case. Dd at the raw level can create physically Seq workloads.
- **Random** – Access pattern moves around a file or physical device.
- **Locality of Reference** – Accesses are concentrated in a particular area (i.e. head of indexes of a data base).
- **Solid State Disk (SSD)** – Storage device with no moving parts. A disk drive whose storage capability is provided by solid state storage.
Performance terms (cont)

- RAID – Redundant Array Independent (Inexpensive) Disks
- RAID 0 – No Redundancy – maybe striped across many drives (rarely used)
- RAID 1 – Also know as mirroring. Data is mirrored to two drives
- RAID 10 – A variation of RAID 1. Will stripe across more than two drives.
- RAID 5 – A complex scheme of storing Parity blocks to recreate data if one device fails
- RAID 6 – Similar to RAID 5 except there are two parity blocks and can survive a double drive failure. Important to new SATA drive technologies where during the drive rebuild process a second failure is likely.
- Bottleneck – a term used to discuss what is holding the system back from performing better. Bottlenecks can be in Processors, HBAs, Controllers or Disk drives.
Overview of components of a storage subsystem

Client

User Space
System Boundary
Kernel Space

Device Drivers

Fibre Channel
SAS

Power Source
Power Meter

Hardware
Software

System Under Test (SUT)

Controller

Disk
Power cables

Power Feed
Power Distribution Unit
VDBENCH

▷ An application that simulates a controlled IO load on a storage system
▷ It is written in 99% Java and 1% C for exceptional efficiency
▷ Designed to execute a workload on a storage system
▷ Performance output can be thought of as a simple equation: $f(\text{Workload, Config}) = \text{Performance} + \text{Power}$
Workload Dimensions

- Workload is a very complex multi-dimensional problem
  - Number of threads or queue depth to storage
  - Transfer size
  - Read to write ratio
  - Sequential vs random
- Cache hit or cache miss
Configuration Dimensions

- Configuration is equally complex multi-dimensional problem
  - Number/type of drives
  - Capacity of configured system
  - Raid Level
  - Size of RAID set
  - Size of Stripe
  - Controller or JBOD
  - Number/type of back-end connections
  - Number/type of front-end connections
  - Volume Manager configuration
  - System parameters that affect storage (sd_max_throttle, max_contig, multi-pathing software, etc)
  - cache mirroring
  - broken hardware (failed controller, disk drive, path, etc)
  - Accessed RAW or Buffered
  - Tiering software active
  - Compression enable / disabled
Performance outputs

- IOs per second for small block workloads
- MB per second for large block workloads
- Average response time in ms
- Combined with the Power Meter
  - Average Watts over the interval
  - Average Amps over the interval
- Generally shows the peak performance of some system resource bottleneck
VDBENCH (cont.)

- VDBENCH is an IO driver that allows for a workload targeted to specific storage and reports performance
  - vdbench has three basic statements to the script
    - SD - Storage Definition - defines what storage to be used in the run
    - WD - Workload Definition - Defines the workload parameters for the storage
    - RD - Run Definitions - determines what storage and workload will be run together and for how long. Causes IO to be executed and report IOPS, Response Times, MB/sec, etc
  - Output from vdbench is a web browser friendly .html file.
Simple 3 line VDBENCH Script

* Author: Henk Vandenberghe.

* Example 1: Single run, one raw disk

* SD: Storage Definition
* WD: Workload Definition
* RD: Run Definition
* Solaris style Raw Disk

sd=sd1,lun=/dev/rdsk/c6t0d0s4
wd=rr,sd=sd1,xfersize=4096,rdpct=100
rd=run1,wd=rr,iorate=100,elapsed=10,interval=1

* Single raw disk, 100% random read of 4k records at i/o rate
* of 100 for 10 seconds
Storage Definitions

- This part of the script defines the storage to be used in this script
- SNIA/EPA workload is designed to run against “RAW” Storage. No buffering.
- Make sure you select the right storage, it will destroy everything on the disk. This includes your root or C: disk.
- Make sd name unique. SD=unique_name
# RAW vs Buffered

<table>
<thead>
<tr>
<th>OS</th>
<th>RAW</th>
<th>Buffered</th>
</tr>
</thead>
</table>
| Windows | lun=\\.\d:  
         | lun=\\.\PhysicalDrive4                   | d:                                            |
| Solaris | lun=/dev/rdsk/c3t0d2s4  
         | lun=/dev/vx/rdsk/c3t0d2s4                  | lun=/dev/dsk/c3t0d2s4  
         |                           | lun=/dev/vx/dsk/c3t0d2s4                 |
| Linux  | lun=/dev/sdb,openflags=o_direct         | lun=/dev/sdb                                 |
| AIX    | lun=/dev/rsatathin1                     | ???                                           |

```
sd=default, size=300g
sd=sd1, lun=/dev/rdsk/c6t3d0s0
sd=sd2, lun=/dev/rdsk/c7t1d0s0, size=200g
sd=sd3, lun=/dev/rdsk/c8t6d0s0, size=200g```
Workload Definitions

- Each WD name must be unique: \( wd = wd\_unique \)
- Parameters include:
  - \( sd \): devices to run against
  - \( seekpct \): Pct time to move location
  - \( rdpct \): read pct
  - \( xfersize \): transfer size
  - \( skew \): Percent of workload for this definition
  - \( threads \): number of threads this definition
  - \( wd \): default setup defaults for the following \( wd \)

- \( hotband = (10, 18) \) execute hot band workload against a range of storage

\[
\begin{align*}
wd &= \text{HOTwd\_uniform}, \text{skew} = 6, sd = sd\_*, \text{seekpct} = 100, rdpct = 50 \\
wd &= \text{HOTwd\_hot1}, sd = sd\_*, \text{skew} = 28, \text{seekpct} = \text{rand}, \text{hotband} = (10, 18)
\end{align*}
\]
Run Definition

- Each run definition name must be unique  \( rd=rd\_unique \)
- Parameters include:
  - \( wd= \) which workload definitions to run now
  - \( iorate= \) define either io/sec or the keyword “max” or “curve”
  - \( warmup= \) define period where ios do not count towards average (30 or 5m or 12h)
  - \( elapsed= \) define length of run

\[ \begin{align*}
  \text{rd} &= \text{rd1\_hband}, \text{wd} = \text{HOTwd*}, \text{iorate} = \text{MAX}, \text{warmup} = 30, \text{elapsed} = 6H, \text{interval} = 10, \text{pause} = 30, \text{th} = 200 \\
  \text{rd} &= \text{rd1\_seq}, \text{wd} = \text{wd\_seq}, \text{iorate} = \text{max}, \text{forrdpct} = (0, 100), \text{xfer} = 256K, \text{warmup} = 30, \text{el} = 20m, \text{in} = 5, \text{th} = 20
\end{align*} \]
Vdbench summary report, created 13:09:26 Mar 13 2013 MST

Link to logfile: logfile
Run totals: totals
Copy of input parameter files: parmfile
Copy of parameter scan detail: parmsgan
Link to errorlog: errorlog
Link to flatfile: flatfile
Link to HOST reports: localhost
Link to response time histogram: histogram
Link to SD reports: sd1 sd2
Link to workload report: wd_mixed
Link to workload report: wd_seq
Link to Run Definitions: rdl_mixed000 For loops: rdpct=0.0 xfersize=8k threads=48.0
rdl_mixed100 For loops: rdpct=100.0 xfersize=8k threads=128.0
rdl_seq For loops: rdpct=0.0 xfersize=256k threads=48.0
rdl_seqR For loops: rdpct=100.0 xfersize=256k threads=128.0

13:09:31.014 Starting RD=rdl_mixed000; I/O rate: 1000; elapsed=1800; For loops: rdpct=0.0 xfersize=8k threads=48.0

<table>
<thead>
<tr>
<th>Mar 13, 2013</th>
<th>interval</th>
<th>i/o rate</th>
<th>MB/sec</th>
<th>bytes</th>
<th>read</th>
<th>resp</th>
<th>read</th>
<th>write</th>
<th>resp</th>
<th>resp max</th>
<th>stddev</th>
<th>depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:10:31.415</td>
<td>1</td>
<td>991.57</td>
<td>7.75</td>
<td>8192</td>
<td>0.00</td>
<td>1.708</td>
<td>0.000</td>
<td>1.708</td>
<td>23.479</td>
<td>1.093</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>13:11:31.294</td>
<td>2</td>
<td>1002.72</td>
<td>7.83</td>
<td>8192</td>
<td>0.00</td>
<td>1.804</td>
<td>0.000</td>
<td>1.804</td>
<td>4.974</td>
<td>0.638</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>13:12:31.271</td>
<td>3</td>
<td>994.57</td>
<td>7.77</td>
<td>8192</td>
<td>0.00</td>
<td>1.797</td>
<td>0.000</td>
<td>1.797</td>
<td>4.780</td>
<td>0.634</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>13:13:31.290</td>
<td>4</td>
<td>1000.23</td>
<td>7.81</td>
<td>8192</td>
<td>0.00</td>
<td>1.802</td>
<td>0.000</td>
<td>1.802</td>
<td>5.278</td>
<td>0.645</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>13:14:31.339</td>
<td>5</td>
<td>1004.10</td>
<td>7.84</td>
<td>8192</td>
<td>0.00</td>
<td>1.805</td>
<td>0.000</td>
<td>1.805</td>
<td>29.753</td>
<td>0.654</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>13:15:31.280</td>
<td>6</td>
<td>1002.27</td>
<td>7.83</td>
<td>8192</td>
<td>0.00</td>
<td>1.848</td>
<td>0.000</td>
<td>1.848</td>
<td>46.246</td>
<td>0.682</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>13:39:31.214</td>
<td>30</td>
<td>995.17</td>
<td>7.77</td>
<td>8192</td>
<td>0.00</td>
<td>1.800</td>
<td>0.000</td>
<td>1.800</td>
<td>5.211</td>
<td>0.633</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>13:39:31.222</td>
<td>avg_2-30</td>
<td>999.11</td>
<td>7.81</td>
<td>8192</td>
<td>0.00</td>
<td>1.816</td>
<td>0.000</td>
<td>1.816</td>
<td>83.487</td>
<td>0.652</td>
<td>1.8</td>
<td></td>
</tr>
</tbody>
</table>
Running vdbench

Parameters to vdbench

- `-f` file(s) to be part of script
- `-o` output directory (add a “+” to keep from overwriting earlier runs)
- `-e` elapsed time override
- `-i` interval time override
- `-w` warmup time override
- `-s` simulate execution (open storage, check syntax)

```
/vdbench/vdbench -f comp_25.txt t5a_config.txt script.txt -o t5_comp_25+
/vdbench/vdbench -i 10 -f one_file_script.txt -o simple_test+
```
Steven.A.Johnson@oracle.com