

VDBENCH Overview

Steven Johnson

SNIA Emerald[™] Training

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

June 24-27, 2013









- 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





- 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.





- 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 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





- 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.











- 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(Workload, Config) = Performance + Power





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, multipathing software, etc)
- cache mirroring
- broken hardware (failed controller, disk drive, path, etc)
- Accessed RAW or Buffered
- Tiering software active
- SNIA Emerald[™] Training ~ June 24-27, 2013



- 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 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.





```
* Author: Henk Vandenbergh.
* 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
```





- 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





| OS | RAW | Buffered |
|---------|---|---|
| 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

sNIA Emerald[™] Training ~ June 24-27, 2013 solution www.sniaemerald.com



- 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
 - SNIA Emerald™

wd=HOTwd_uniform, skew=6, sd=sd*, seekpct=100, rdpct=50

wd=HOTwd hot1,sd=sd*,skew=28,seekpct=rand,hotband=(10,18) SNIA EmeratedTM Training ~ June 24-27,2013 www.sniaemerald.com

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



- 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

- interval= time between reporting statistics in seconds
- threads= number of threads per lun or concatenated storage
- forrdpct= range of pct read to execute

rd=rd1_hband,wd=HOTwd*,iorate=MAX,warmup=30,elapsed=6H,interval=10,pause=30,th=200 rd=rdt_seq,wd=wd_seq,iorate=max,forrdpct=(0,100),xfer=256K,warmup=30,el=20m,in=5,th=20 SNIA EmeraldTM Training ~ June 24-27, 2013 www.sniaemerald.com

Performance outputs summary.html



| Vdbench summary report, created | 13:09:26 Mar 13 2013 MST | | | | |
|----------------------------------|--|--|--|--|--|
| Link to logfile: | <u>logfile</u> | | | | |
| Run totals: | totals | | | | |
| Copy of input parameter files: | <u>parmfile</u> | | | | |
| Copy of parameter scan detail: | parmscan | | | | |
| Link to errorlog: | errorlog | | | | |
| Link to flatfile: | flatfile | | | | |
| Link to HOST reports: | localhost | | | | |
| Link to response time histogram: | <u>histogram</u> | | | | |
| Link to SD reports: | <u>sd1</u> <u>sd2</u> | | | | |
| Link to workload report: | wd_mixed | | | | |
| Link to workload report: | <u>wd_seq</u> | | | | |
| Link to Run Definitions: | <u>rd1_mixed000 For loops: rdpct=0.0 xfersize=8k threads=48.0</u> | | | | |
| | <u>rd1_mixed100 For loops: rdpct=100.0 xfersize=8k threads=128.0</u> | | | | |
| | <u>rd1_seq For loops: rdpct=0.0 xfersize=256k threads=48.0</u> | | | | |
| | rd1 segR For loops: rdpct=100.0 xfersize=256k threads=128.0 | | | | |

13:09:31.014 Starting RD=rd1_mixed000; I/O rate: 1000; elapsed=1800; For loops: rdpct=0.0 xfersize=8k threads=48.0

| Mar 13, 2013 | interval | i/o rate | MB/sec 1024**2 | bytes i/o | read pct | resp time | read resp | write resp | resp max | resp stddev | queue depth |
|--------------|----------|-------------|-------------------|--------------|-------------|--------------|--------------|---------------|-------------|----------------|----------------|
| 13:10:31.415 | 1 | 991.57 | 7.75 | 8192 | 0.00 | 1.708 | 0.000 | 1.708 | 23.479 | 1.093 | 1.7 |
| 13:11:31.294 | 2 | 1002.72 | 7.83 | 8192 | 0.00 | 1.804 | 0.000 | 1.804 | 4.974 | 0.638 | 1.8 |
| 13:12:31.271 | 3 | 994.57 | 7.77 | 8192 | 0.00 | 1.797 | 0.000 | 1.797 | 4.780 | 0.634 | 1.8 |
| 13:13:31.290 | 4 | 1000.23 | 7.81 | 8192 | 0.00 | 1.802 | 0.000 | 1.802 | 5.278 | 0.645 | 1.8 |
| 13:14:31.339 | 5 | 1004.10 | 7.84 | 8192 | 0.00 | 1.805 | 0.000 | 1.805 | 29.753 | 0.654 | 1.8 |
| 13:15:31.280 | 6 | 1002.27 | 7.83 | 8192 | 0.00 | 1.848 | 0.000 | 1.848 | 46.246 | 0.682 | 1.9 |
| | | | | | | | | | | | |
| 13:39:31.214 | 30 | 995.17 | 7.77 | 8192 | 0.00 | 1.800 | 0.000 | 1.800 | 5.211 | 0.633 | 1.8 |
| 13:39:31.222 | avg_2-30 | 999.11 | 7.81 | 8192 | 0.00 | 1.816 | 0.000 | 1.816 | 83.487 | 0.652 | 1.8 |



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



SNIA Emerald[™] Training ~ June 24-27, 2013

www.sniaemerald.com