VDBench + Script

Steven Johnson – Oracle
Carlos Pratt - IBM

SNIA Emerald™ Training

SNIA Emerald Power Efficiency Measurement Specification,

Version 2.1

July 20-21, 2015

SNIA™ Green Storage Initiative

SNIA Emerald™

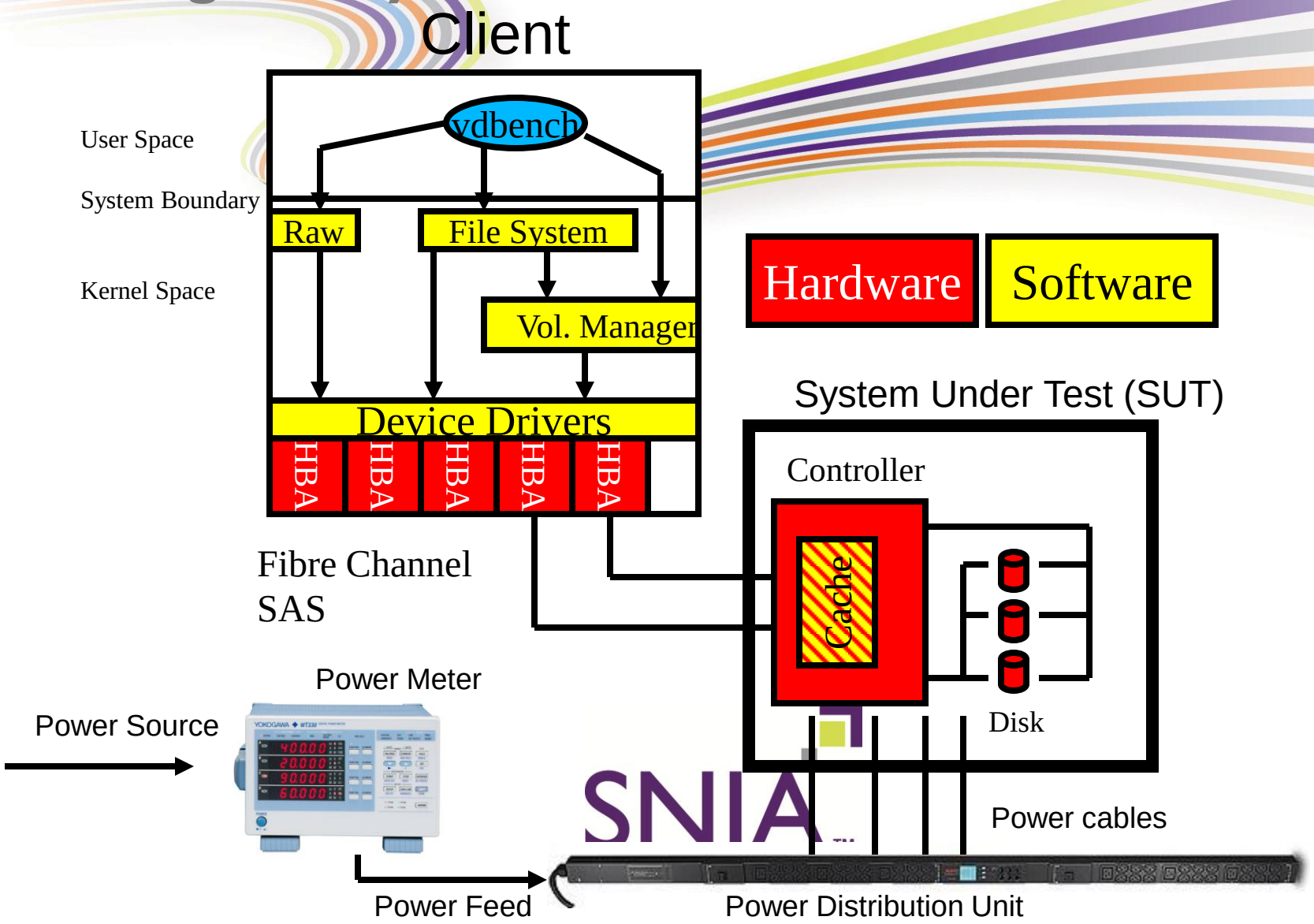
Agenda

- VDBench IO driver overview
- Workload + configuration = performance + power consumption
- Defining storage
- Review Emerald Script
- VDBench 50403 – new features
- Questions

VDBENCH Overview

- Download at: http://www.oracle.com/technetwork/server-storage/vdbench-downloads-1901681.html
- Included in the zip file is the vdbench.pdf manual
- 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**

Overview of components of a storage subsystem



Workload Dimensions

- Number of threads or queue depth to storage
- Transfer size
- Read to write ratio
- Sequential vs random
- Cache hit or cache miss

Configuration Dimensions

- 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

```
*
* 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
```

SNIA™
Green Storage Initiative

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

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/rhdisk9	???

```
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

```
wd=HOTwd_uniform,skew=6,sd=sd*,seekpct=100,rdpct=50
wd=HOTwd_hot1,sd=sd*,skew=28,seekpct=rand,hotband=(10,18)
```

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
  - 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=rd1_seq,wd=wd_seq,iorate=max,forrdpct=(0,100),xfer=256K,warmup=30,el=20m,in=5,th=20
```

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+
```

SNIA™
Green Storage Initiative

Performance output summary.html

Vdbench summary report, created 10:16:02 Apr 07 2015 PDT

Link to logfile: logfile
Run totals: totals
Copy of input parameter files: parmfile
Copy of parameter scan detail: parmscan
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 sd3 sd4 sd5 sd6 sd7 sd8

Link to workload report: HOTwd_uniform
Link to workload report: HOTwd_hot1
Link to workload report: HOTwd_99rseq1
Link to workload report: HOTwd_99rseq2
.
.
Link to workload report: wd_fill

Link to Run Definitions: rd1_hband_warm For loops: threads=58
rd1_hband_final For loops: threads=58

10:21:08.000 Starting RD=rd1_hband_final; I/O rate: Controlled MAX; elapsed=600; For loops: threads=58

Apr 07, 2015	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	cpu% sys+u	cpu% sys
10:22:08.372	1	42179.43	1192.99	29657	65.80	17.730	17.838	17.524	115.524	19.987	748.1	24.3	20.6
10:23:08.312	2	42335.78	1196.99	29647	65.82	17.768	17.982	17.356	120.671	20.265	752.2	24.2	20.7
10:24:08.352	3	42303.88	1197.53	29682	65.85	17.782	17.930	17.496	122.466	20.154	752.3	24.4	21.0
10:25:08.312	4	42341.05	1199.77	29712	65.80	17.767	17.796	17.710	122.914	20.104	752.3	24.3	21.0
.													
.													
10:31:08.532	10	42508.40	1203.51	29687	65.78	17.700	17.682	17.734	124.943	20.099	752.4	23.5	20.3
10:31:08.622	avg_2-10	42380.09	1199.41	29675	65.81	17.751	17.857	17.547	131.292	20.181	752.3	24.1	20.7

10:31:09.372 Vdbench execution completed successfully

Emerald Script

- A VDBench script has been developed for Emerald Testing
- This script needs editing to fit your specific environment

Defining Storage for the test

```
#Version 2015_07_17 Draft version for WebEx, Add 4K support, example Linux, AIX SDs

# Any resulting script is run with the Emerald_System_Configuration through this example command line:
#      vdbench -f Emerald_test_script.txt –o out_dir
#
#Version 2015_07_17 Draft version for WebEx, Add 4K support, example Linux, AIX SDs
concatenate=yes
compratio=2.00

###################
# Begin Storage Designator Section
# Change sd's to Match Storage Configuration
###################
###################
#  Example Storage Definition (sd) (Windows)
###################
#sd=sd1,lun=\\.\PhysicalDrive2
#sd=sd2,lun=\\.\PhysicalDrive3
# .
# .
#sd=sdN,lun=\\.\PhysicalDriveN
###################
#  Example Storage Definition (sd) (Linux)
###################
#sd=sd1,lun=/dev/sdb,openflags=o_direct
#sd=sd2,lun=/dev/sdc,openflags=o_direct
# .
# .
#sd=sdN,lun=/dev/sdN,openflags=o_direct
###################
#  Example Storage Definition (sd)  (AIX)
###################
#sd=sd1,lun=/dev/rhdisk2
#sd=sd2,lun=/dev/rhdisk3
# .
# .
# sd=sdN,lun=/dev/rhdiskN
```

Selecting the Physical device block size

- Large disk drives are changing their smallest amount of data they can transfer.
- Older systems historically have used 512 byte block devices.
- New devices are moving to a 4K byte block. The script needs to be changed for 4K devices.

```
# Default transfer sizes for native 512 Byte Block devices
wd=default,xfersize=(8k,31,4K,27,64K,20,16K,5,32K,5,128K,2,1K,2,60K,2,512,2,256K,2,48K,1,56K,1),rdp
# Uncomment next line for Default transfer sizes for native 4K Byte Block devices
#wd=default,xfersize=(8k,31,4K,31,64K,20,16K,5,32K,5,128K,2,60K,2,256K,2,48K,1,56K,1),rdpct=70,th=1
```

Streams and Threads

- Each system will need load changes in the script

```
# Sequential 4 Corners workload
# Replace Change_a2 defines the number of streams across the concatenated storage space
wd=wd_seq,sd=sd*,seekpct=0,streams=Change_a2

# Pre=fill storage workload
# Replace Change_a1 defines the number of streams across the concatenated storage space
# Hint: Normally, Change_a2 equates to Change_a1
wd=wd_fill,sd=sd*,seekpct=eof,streams=Change_a1

###################
#Pre-fill and conditioning Run Definitions
###################
# Pre-fill Test Phase Test phase that fills storage.
# Replace Change_y1 with the optimal number of threads that the system can handle and fill the sto
#  The number of threads (Change_y1) for the pre-fill workload shall be a multiple of Change_a1
#  Hint: After tuning Change_y2 below Eguate Change_y1 to Change_y2
#  PREFILL NOT PART OF POWER TESTING

rd=rd_prefill,wd=wd_fill,iorate=max,rdpct=0,xfersize=256K,elapsed=5000m,interval=60,th=Change_y1
# START OF POWER TESTING
# Conditioning Test Phase
# Test phase to condition and stabilize the storage system
# Replace Change_x1 to optimal number of threads for system.  Recommend ~8 per physical drive in sy
# After tuning to determine Change_x2 below Change_x1 Shall = Change_x2
rd=rd_conditioning,wd=HOTwd*,iorate=MAX,warmup=10m,elapsed=12H,interval=60,th=Change_x1
```

Streams and Threads (cont)

```
##################
# Active Run Definitions
##################
#default parameters used for all active run definitions
rd=default,iorate=MAX,elapsed=31m,interval=60
# Hot Band test phase
# Replace Change_x2 to optimal number of threads for system.  Recommend ~8 per physical drive in system
rd=rd_hband_final,wd=HOTwd*,th=Change_x2
# Random writes test phase
# Replace Change_x3 to optimal number of threads for system.  Recommend ~4-8 per physical drive in syste

rd=rd_rw_warm,wd=wd_mixed,rdpct=0,xfersize=8k,elapsed=10m,th=Change_x3  #added section for warmup period

rd=rd_rw_final,wd=wd_mixed,rdpct=0,xfersize=8k,th=Change_x3
# Random reads test phase
# Replace Change_x4 to optimal number of threads for system.  Recommend ~8 per physical drive in system

rd=rd_rr_warm,wd=wd_mixed,rdpct=100,xfersize=8k,elapsed=10m,th=Change_x4  #added section for warmup peri
rd=rd_rr_final,wd=wd_mixed,rdpct=100,xfersize=8k,th=Change_x4
# Sequential write test phase
# Replace Change_y2 with the optimal number of threads for the system under test Recommend 2-3 per physi

# The number of threads (Change_y2) for the sequential workload shall be a multiple of Change_a2
rd=rd_sw_warm,wd=wd_seq,rdpct=0,xfersize=256K,elapsed=10m,th=Change_y2
#added section for warmup period of 10 minutes
rd=rd_sw_final,wd=wd_seq,rdpct=0,xfersize=256K,th=Change_y2
# Sequential read test phase
# Replace Change_y3 with the optimal number of threads for the system under test Recommend 2-3 per physi

# The number of threads (Change_y3) for the sequential workload shall be a multiple of Change_a2

rd=rd_sr_warm,wd=wd_seq,rdpct=100,xfersize=256K,elapsed=10m,th=Change_y3
rd=rd_sr_final,wd=wd_seq,rdpct=100,xfersize=256K,th=Change_y3
```

VDBench 5.04.03 new features

- Several enhancements have been made to the latest VDBench
  - Correct errors in number of threads used during hot banding tests (resulted in th=1 being th=13)
  - Added new skew table to assist in tuning system

Skew.html (cont.)

		i/o	MB/sec	bytes	read	resp	read	write	resp	resp	queue	skew	skew	skew
.962	WD:	rate	1024**2	i/o	pct	time	resp	resp	max	stddev	depth	requested	observed	delta
.962	HOTwd_uniform	1605.38	45.04	29421	50.00	0.489	0.767	0.210	20.481	0.670	0.8	6.00%	6.00%	
.962	HOTwd_hot1	7492.27	210.02	29393	70.00	0.614	0.786	0.213	70.233	0.772	4.6	28.00%	28.00%	
.963	HOTwd_99rseq1	1337.04	38.49	30182	100.00	0.781	0.781	0.000	26.672	0.865	1.0	5.00%	5.00%	
.963	HOTwd_99rseq2	1339.16	38.40	30068	100.00	0.779	0.779	0.000	24.775	0.863	1.0	5.00%	5.00%	
.963	HOTwd_99rseq3	1338.30	38.40	30089	100.00	0.778	0.778	0.000	26.929	0.862	1.0	5.00%	5.00%	
.963	HOTwd_99rseq4	1337.66	38.40	30105	100.00	0.778	0.778	0.000	32.428	0.862	1.0	5.00%	5.00%	
.963	HOTwd_99rseq5	1337.80	38.37	30072	100.00	0.778	0.778	0.000	25.847	0.863	1.0	5.00%	5.00%	
.963	HOTwd_hot2	3747.54	105.06	29395	70.01	0.605	0.773	0.212	24.763	0.761	2.3	14.00%	14.01%	
.963	HOTwd_hot3	1872.71	52.54	29417	70.00	0.609	0.778	0.212	25.724	0.766	1.1	7.00%	7.00%	
.963	HOTwd_hot4	1338.08	37.52	29405	69.99	0.611	0.782	0.212	15.626	0.771	0.8	5.00%	5.00%	
.963	HOTwd_99wseq1	1337.99	38.36	30059	0.00	0.227	0.000	0.227	16.860	0.223	0.3	5.00%	5.00%	
.963	HOTwd_99wseq2	1336.11	38.37	30109	0.00	0.227	0.000	0.227	62.127	0.225	0.3	5.00%	4.99%	
.963	HOTwd_99wseq3	1338.07	38.39	30081	0.00	0.227	0.000	0.227	12.720	0.222	0.3	5.00%	5.00%	
.964	Total	26758.13	757.36	29678	65.81	0.588	0.780	0.218	70.233	0.754	15.7	100.00%	100.00%	

SNIA™
Green Storage Initiative

Skew.html

18:32:57.961		i/o	MB/sec	bytes	read	resp	read	write	resp	resp	queue
18:32:57.961	Slave:	rate	1024**2	i/o	pct	time	resp	resp	max	stddev	depth
18:32:57.961	host2-0	1663.34	47.05	29662	65.79	0.590	0.783	0.220	29.308	0.756	1.0
18:32:57.961	host2-1	1665.35	47.13	29673	65.82	0.590	0.783	0.220	32.428	0.757	1.0
18:32:57.961	host2-2	1671.87	47.30	29667	65.84	0.590	0.782	0.219	24.805	0.757	1.0
18:32:57.961	host2-3	1668.72	47.21	29665	65.83	0.589	0.781	0.219	25.410	0.757	1.0
18:32:57.961	host2-4	1668.76	47.26	29694	65.80	0.589	0.781	0.218	26.672	0.756	1.0
18:32:57.961	host2-5	1668.76	47.23	29675	65.80	0.589	0.781	0.219	62.127	0.757	1.0
18:32:57.962	host2-6	1673.15	47.34	29667	65.81	0.589	0.781	0.219	25.847	0.755	1.0
18:32:57.962	host2-7	1665.62	47.15	29685	65.81	0.590	0.783	0.220	70.233	0.757	1.0
18:32:57.962	Total	13345.56	377.67	29674	65.81	0.590	0.782	0.219	70.233	0.757	7.9

SNIA™
Green Storage Initiative

Questions?

SNIA™
Green Storage Initiative