An overview of the VMmark benchmark on HP ProLiant servers and server blades

Executive summary

This paper explains the purpose and importance of VMmark, how the benchmark is conducted, and how to interpret the results. This document is an attempt to provide VMmark information for a less technically trained audience of decision makers than served by the official VMmark Guide on the VMware website.

What VMmark measures

The VMmark benchmark is intended to measure the performance of virtualized servers on a system under test (SUT) so that customers can compare the capabilities of different platforms for virtualization. VMmark represents the performance of virtual machines within a server running VMware ESX and a set combination of operating systems and specially tuned applications reflecting a typical datacenter environment. VMmark uses a collection of 'sub-tests' derived from commonly used load-generation tools as well as from benchmarks developed by the Standard Performance Evaluation Corporation (SPEC®). VMmark is an open standards effort that is agnostic toward hardware platforms and different virtualization software systems. VMmark uses workloads that represent common applications in datacenters. It is important to note that VMmark is designed to benchmark the performance of the virtualization software and the hardware, and is not designed as a benchmark of any other software component.

The importance of VMmark

VMmark is the first virtualization benchmark in the industry. Traditional benchmarks are designed to stress a single system within a test run, attempting to achieve maximum performance or scalability. Results of these benchmarks do not provide information about the scalability of virtualized systems. In a virtualized system, multiple benchmarks can be run on the same number of processors and other components as a non-virtualized system. By definition, then, these benchmarks running concurrently would not be able to achieve comparable results as when they are run singly on a system to which all processing and other resources can be devoted in an undivided manner to the single benchmark. A new benchmark was needed to standardize the ability to measure and compare virtualized system performance. A great deal of effort was invested in the design of the VMmark benchmark to provide objective and useful results.

Virtualization software overhead

The performance of a Virtual Machine (VM) can not be predicted by taking a single-server benchmark score and dividing that by the number of VMs in a planned virtualized configuration of the single server platform. As virtualization software consumes some system resources, there is overhead in running the virtualized environment.

Choice of workloads and operating systems

VMmark developers carefully considered what applications common to a datacenter environment could be used. Part of the design also included choosing what combinations of operating systems might commonly be run on a virtualized server in a consolidation effort or new server deployment.

Tuning for virtualization and for the benchmark

After the workloads were chosen, developers needed to determine how to tune these applications to be measurable within a virtualization framework. Virtualization has an impact on other applications being run on a system. The impact is different for each application. Different factors affect performance. In a virtualized system, in addition to virtualization software overhead, there is virtualized OS overhead. Also, the VMs are sharing resources such as processor, disk I/O, network bandwidth, and memory, all of which impact individual VM and total system performance. A non-virtualized server can achieve a higher SPECweb score than a virtualized server. Why is there a difference? A non-virtualized server and the application running on it can access to all resources, whereas in a virtualized environment these resources are shared with multiple virtual machines with their individual operating system instances and applications.

Standardization

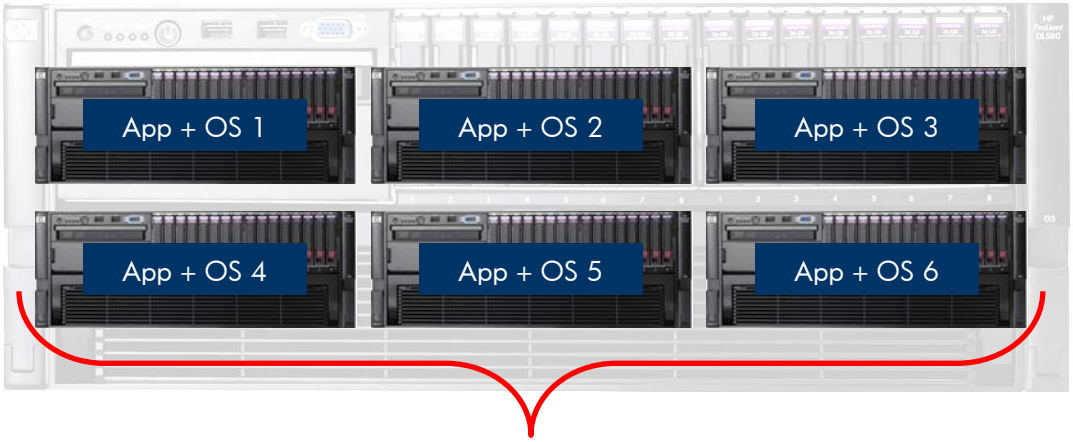
In order for any benchmark to be useful in comparing systems, there must be requirements about how to set up the different aspects of the benchmark consistently, for both hardware and software. VMmark specifies how the packages need to be built, such as how much memory, disk, and CPU are allowed. Vendors running the benchmark may not make any changes to the packages as modified for VMmark or to the virtual machine configuration rules.

VMmark workload suite and key calculations

Tile – the unit of work

VMmark uses sets of 6 virtual machines to run the workloads, and refers to one set of 6 virtual machines with workloads as a 'tile.' The two most important numbers in the results are the performance of each individual workload and the total number of tiles that a system can run. The total number of tiles that a system can run gives an estimate of the system's capacity for consolidation. A good reason for using the word 'tile' is that one visual representation looks similar to a tile mosaic.

Figure 1. Tile: 6 virtual machines each running a different workload and separate operating system[1]



The VMmark benchmark can be set up to run just one tile, or multiple tiles. In Figure 2, the server is running eight tiles.

Decisions about how many tiles to attempt running depend on the platform chosen. Platforms with more processors and more cores can run more tiles. Expectations about how many tiles can be run are not easily predictable. Test engineers estimate the maximum number of tiles and start in that range, and keep adding a tile at a time to the test until the run is not successful. That shows the maximum number of tiles achievable, and the test is then officially run with the maximum successful number of tiles.

[1] VMware VMmark Benchmarking Guide, page 22.

Figure 2. Visual representation of a single server running more than one tile[2]

8 groups of 6 virtual machines = 8 VMmark tiles

The virtual machine configurations for a VMmark tile

Table 1 shows the workloads and applications being run with each VMmark tile. Note that the standby server virtual machine does not run an application as it functions to answer a heartbeat during the test run; however, it does run an operating system and is configured as 1 CPU with a specified amount of memory and disk space.

Table 1. VMmark workload summary per tile

Workload	Application	Virtual Machine Platform
Mail server	Exchange 2003	Windows 2003, 2 CPU, 1 GB RAM, 24 GB disk
Java server	SPECjbb2005-based	Windows 2003, 2 CPU, 1 GB RAM, 8 GB disk
Web server	SPECweb2005-based	SLES 10, 2 CPU, 512 MB RAM, 8 GB disk
Database server	MySQL	SLES 10, 2 CPU, 2 GB RAM, 10 GB disk
File servers	dbench	SLES 10, 1 CPU, 256 MB RAM, 8 GB disk
Standby server	*None*	*Windows 2003, 1 CPU, 256 MB RAM, 4 GB disk*

The following list shows just one example for each application regarding how it is tuned for VMmark:

- **Mail server** – LoadSim from Microsoft simulates Exchange mail server users. To make LoadSim appropriate for VMmark, the number of mail server users is set at 500.
- **Java Server** – SPECjbb2005 is used to measure a system's ability to run Java applications. SPECjbb2005 does multiple short runs while increasing the size of the database. VMmark required a steady load and simulation of a long-running application; therefore, the database size was modified to be set to the maximum and one long run is performed.
- **Web Server** – SPECweb2005 measures the number of simultaneous user sessions in querying and accessing web pages. For VMmark, the test was modified from performing three shorter iterations of the benchmark run to

[2] http://www.vmware.com/products/vmmark/overview.html

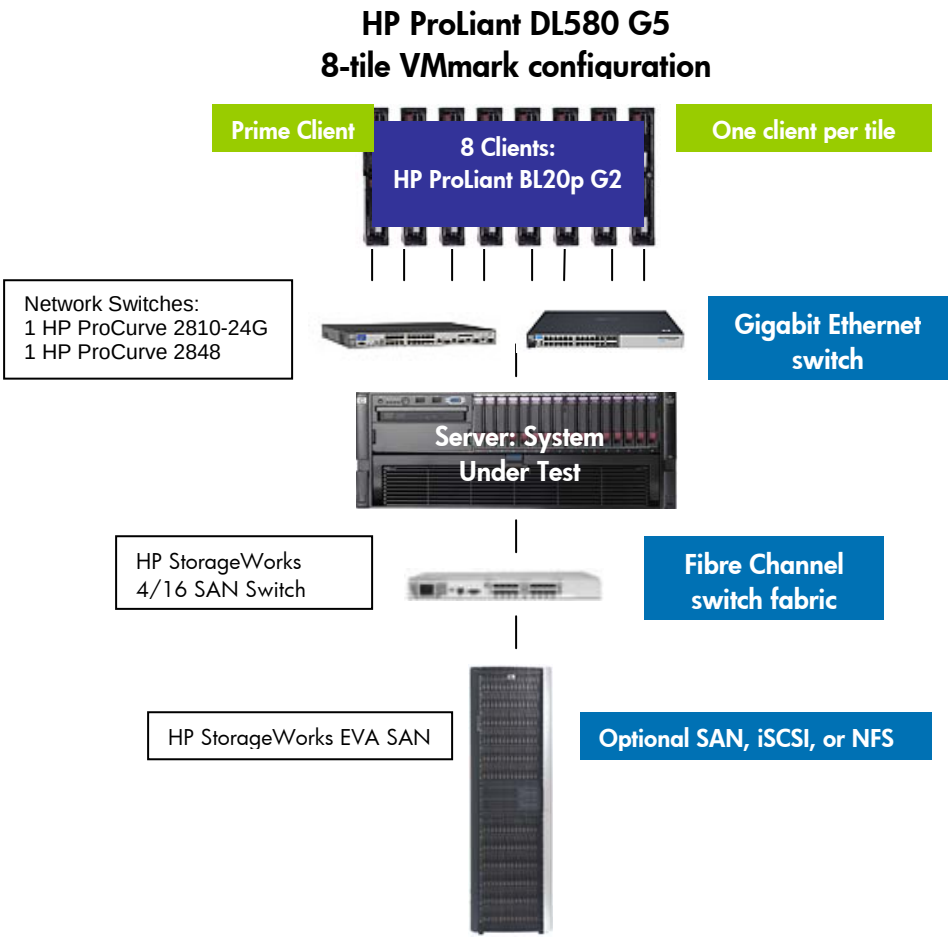
perform a single long iteration, to avoid multiple ramp up and ramp down periods during the VMmark benchmark.

- **Database Server** – SysBench, an open-source database benchmarking tool, is used in VMmark, along with MySQL as the underlying database. For VMmark, SysBench is set up to simulate 16 users, and the database instance is approximately 2.5GB.
- **File Server** – Derived from the industry-standard NetBench benchmark, dbench is used to measure the ability of a file server to service requests from clients. The dbench application in single system benchmarking has a relatively short running time. VMmark modifies it to run repeatedly for the duration of each VMmark run.

VMmark client systems

Client systems work in conjunction with the VMmark test configurations to drive the workloads on the tile. Each tile requires a client system with specific configuration rules and software. When more than one tile is run, one client is set up as the 'primary client.' In the figure below, the system had 8 tiles and needed 8 clients.

Figure 3. VMmark Benchmark network configuration for the HP ProLiant DL580 G5 result[3]



[3] See the results page on the VMmark website listed at the end of this paper.

Key calculations

After a VMmark run, each workload reports a performance metric. Table 2 shows what metric is reported by each workload.

Table 2. VMmark workload summary

Workload	Application	Metric
Mail server	Exchange 2003	Actions/minute
Java server	SPECjbb2005-based	New orders/minute
Web server	SPECweb2005-based	Accesses/minute
Database server	MySQL	Commits/minute
File server	dbench	MB/second

Metrics are collected every 60 seconds over a run designed to last for 3 hours. Upon completion of the test run, there is a large set of metrics reported. The results are a snapshot of how the 5 workload VMs are performing at a point in time, compared to the scores to the reference system, and computed as a ratio. To get consistent scores, the first and last half hour of the test period are not used in the score computation; only the middle steady-state period scores are used. This middle two-hour period is divided into three 40-minute periods, and the scores within each period are averaged to achieve 3 scores for the application. The median of these 3 scores is used as the final score for an application.

Each application's final score is normalized by comparison to the results of a reference system capable of successfully running a single tile, to obtain a ratio. Then, a geometric mean of the normalized scores is computed as the final score for the tile. In a system running more than one tile, the per-tile scores are summed to create the final metric. The reference system for all VMmark results was run on an HP ProLiant DL580 G2 single-tile benchmark. The set of reference scores obtained by this system is shown in Table 6 in the Appendix. These reference scores will not be re-run for future updating. If they were updated, than all scores posted up to that time would have to be recalculated based on any update.

Single-tile workload

The tables below show example metrics for a single-tile test and a multiple-tile test run.

Table 3. Example workload scores for SUT and reference system, single tile (artificial data)

Workload	Score for SUT	Score for Reference System	Ratio
Mail server	950 actions/minute	1000 actions/minute	950/1000 = **.95**
Java server	940 new orders/minute	1000 new orders/minute	940/1000 = **.95**
Web server	1020 accesses/minute	1000 accesses/minute	1020/1000 = **1.02**
Database server	1100 commits/minute	1000 commits/minute	1100/1000 = **1.10**
File server	20 MB/second	10 MB/second	20/10 = **2.00**

Using a geometric mean, the normalized scores are combined to arrive at the score for the tile:

(0.95 * 0.94 * 1.02 * 1.10 * 2.00) ^ (0.2) = **1.15**

The score for VMmark for this run would be stated as **1.15 @ 1 tile**.

Multiple-tile workload

In comparable platforms, if the SUT runs more than one tile, the workload scores within each tile will be lower than for systems running just one tile, but the aggregate score will be higher.

Table 4. VMmark workload summary for multi-tile benchmark, same reference system score for ratio (artificial data)

Workload	SUT Tile 1	SUT Tile 2	SUT Tile 3	SUT Tile 4
Mail server	900	920	910	890
Java server	840	850	850	840
Web server	1020	1000	990	1030
Database server	950	980	930	970
File server	8	7	8	8
Using a geometric mean, the normalized scores are combined to arrive at the score for the tile.				
Tile scores	**0.90**	**0.88**	**0.89**	**0.90**

The overall score for the multi-tile system under test would be the sum of the normalized or geometric mean scores for the four tiles:

0.90 + 0.88 + 0.89 + 0.90 = 3.58.

The score for this VMmark run would be stated as **3.58 @ 4 tiles**.

Interpreting and comparing the results

To interpret results, you must look at both numbers within the score. Within any result set with the same number of tiles, the system with the higher score (left number) has achieved a better result. A higher score for the left number will always be associated with a higher number of tiles. A score in a test run of 5 tiles and the same processors across two different platforms will be similar. Being able to achieve more tiles is desirable. VMmark scores **should** be similar for similar platforms, because VMmark is designed to test the processors; therefore, identical processor types should achieve equivalent results in terms of both score and number of tiles achievable. Comparisons are most useful between systems with the same type of processors, for example, between quad-core and quad-core, not between quad-core and dual-core. Higher scores always correlate with higher numbers of tiles, so both numbers are important in evaluating performance. Results with different numbers of tiles can be compared in the sense that a platform able to run more tiles was able to run more concurrent workloads than other platforms, which is desirable. However, the only useful comparisons are between similar platforms, such as two quad-core two-socket platforms, not a quad-core four-socket platform versus a quad-core two-socket platform.

Dell published a paper earlier touting the advantages of virtualizing on their 2-socket quad-core PowerEdge 2950 system as compared to HP's four-socket quad-core ProLiant platforms, claiming better price/performance among other advantages. However, HP can make the same claim, that our 2-socket quad-core platforms are less expensive for many purposes than our 4-socket quad-core platforms. Two-socket and four-socket platforms are both valid choices, each suited to an ideal environment.

HP recommends that customers look into the details of system configuration for any results posted. For example, it is important to look at what type of RAID is configured. HP best practices would recommend a fault tolerant configuration, which may use up some system resources and result in a lower score than a system not implementing data protection. Another aspect to look at is whether the benchmark has been run with a shipping version of VMware ESX Server. Some of the recent results posted by other vendors have been run with beta versions of VMware ESX Server. The HP ProLiant DL580 G5 and BL680 G5 servers posted the industry's **first** 16-core results and so far the BL680 G5 is the **only** server blade results on the VMmark benchmark. Results show that the HP ProLiant DL580 G5 is the highest VMmark result overall for systems running released virtualization software.

Figure 6. VMmark results obtained with released versions of VMware ESX Server as of 12-07-07

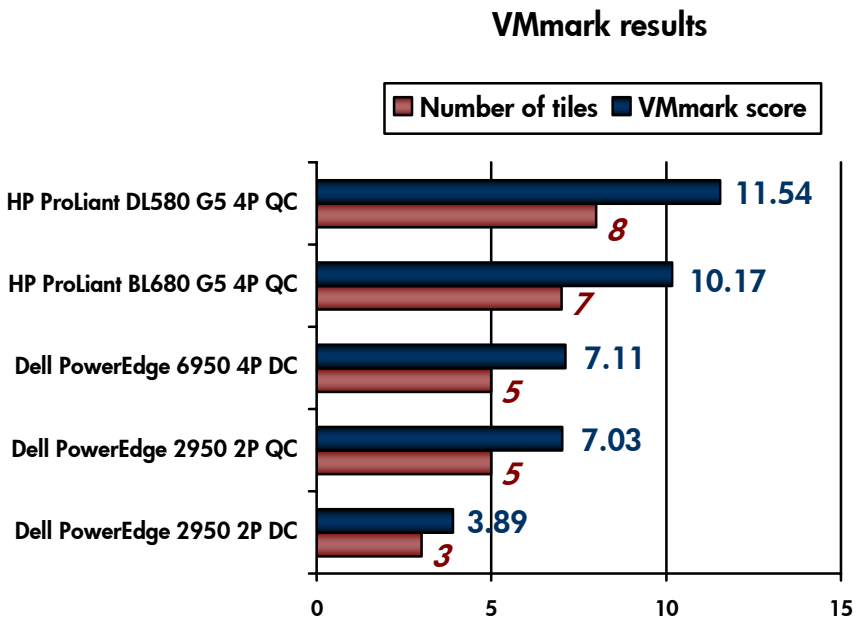


Table 5. VMmark configuration for system results in Figure 6

Submitter/System Description	VMmark score
16 cores	
HP ProLiant DL580 G5: **4xQuad-Core Intel Xeon X7350 2.933 GHz** (4 sockets/4 cores per socket/1 thread per core), 2 x 4 MB L2 cache, 64 GB memory	**11.54 @ 8 tiles**
HP ProLiant BL680 G5: **4xQuad-Core Intel Xeon E7340 2.4 GHz** (4 sockets/4 cores per socket/1 thread per core), 2 x 4 MB L2 cache, 64 GB memory	**10.17 @ 7 tiles**
8 cores	
Dell PowerEdge 6950: 4xDual-Core AMD Opteron(tm) Processor 8222SE 3.0 GHz (4 sockets/2 cores per socket/1 thread per core), L1 cache 64KB(I) + 64KB(D) on chip/core, L2 cache B(I+D) on chip/core, 64 GB memory	**7.11 @ 5 tiles**
Dell PowerEdge 2950: 2xQuad-Core Intel Xeon X5365 3.0 GHz (2 sockets/4 cores per socket/1 thread per core), L1 cache 32KB(I) + 32KB(D) on chip/core, L2 cache 8MB(I+D) on chip/per chip, 4 MB shared/2 cores, 32 GB memory	**7.03 @ 5 tiles**
4 cores	
Dell PowerEdge 2950: 2xDual-Core Intel Xeon 5160 3.0 GHz (2 sockets/2 cores per socket/1 thread per core), L1 cache 32KB(I) + 32KB(D) on chip/core, L2 cache 4MB(I+D) on chip/per chip, 32 GB memory	**3.89 @ 3 tiles**

HP ProLiant servers, BladeSystem, and VMmark

Partnership between HP and VMware

HP is proud that the HP ProLiant DL580 server platform was chosen to be the reference system in the development of the VMmark benchmark. More than a dozen ProLiant servers are certified for VMware. HP can help your business plan, implement and operate a virtual infrastructure with VMware. HP qualifies a wide range of ProLiant servers, StorageWorks storage, and integrated HP management software. For a quick overview, download our Solutions Guide (pdf), or visit www.hp.com/go/vmware for more information. HP offers a total of 41 VMware ESX Server 3.0 certified servers, more than IBM, Dell, and Sun.[4]

HP market leadership

HP ProLiant servers and server blades are a vital part of the HP success story. HP is the #1 vendor in worldwide server shipments. HP increased its worldwide server unit shipments by 10 times the total of all other vendors combined in the third calendar quarter of 2007, according to figures released on 11-29-07 by industry analyst firm IDC.[5]

HP proven performance

Proven performance is part of the reason that HP is #1 in server shipments. HP has posted hundreds of benchmark results on the most commonly used benchmarks on hundreds of ProLiant servers and blades, helping customer to identify reasons to be confident in HP.

Appendix

The reference system used for normalizing all benchmark results was the HP ProLiant DL580 G2 running VMware's ESX Server 3.0.1, build 32039 (with patch ESX-6075798). The system contained two 2.2-GHz single-core Intel Xeon CPUs with hyper-threading support, and was configured with 16GB of memory. Storage was provided by an EMC Clariion CX500 disk array connected via a 1Gb/s fiber channel link and containing five 10,000 RPM disks configured in RAID5. The load-generating client was an HP ProLiant DL385 with two 2.6 GHz single-core AMD Opteron CPUs and 4 GB of memory running 32-bit Microsoft Windows Server 2003 operating system with Service Pack 2. The client and the reference system were connected through a single 1 Gigabit Ethernet link.

Table 6 shows the actual scores of the reference system.

Table 6. Reference System Workload Scores

Workload	Score
Mail server	1096.80 actions/minute
Java server	16,k613.58 new orders/minute
Web server	1018.95 accesses/minute
Database server	1,492.38 commits/minute
File server	12.83 MB/second
Standby server	*Not applicable*

[4] Same cross-generational count used for competitor platforms. For the most up to date list visit: www.hp.com/go/vmware and http://www.vmware.com/pdf/vi3_systems_guide.pdf. The VMware systems guide was last updated October 29, 2007. Information valid as of 11-02-07.

[5] IDC, Q307 Worldwide Quarterly Server Tracker, November 2007.

For more information

For more information on VMware for HP ProLiant servers:
http://h18004.www1.hp.com/products/servers/vmware/index.html

HP VMware information:
http://www.hp.com/go/vmware

Home page for VMware's VMmark:
http://www.vmware.com/products/vmmark/overview.html

VMmark FAQ:
http://www.vmware.com/products/vmmark/faq.html

VMmark Guide:
http://www.vmware.com/vmtn/resources/573

Full Disclosure Reports for DL580 G5 and BL680 G5, the two top results and the only 16-core results posted as of date of publication: http://www.vmware.com/files/pdf/vmmark_hp1.PDF and http://www.vmware.com/files/pdf/vmmark_hp2.PDF

December 2007