Jan Karlsson & Patrik Eriksson
jan3karlsson@gmail.com & patrik.carl.eriksson@gmail.com
April 18, 2014

How the choice of Operating System can affect databases on a Virtual Machine.

Contact Information:

Author(s):

Jan Karlsson, Patrik Eriksson

E-mail: jan3karlsson@gmail.com & patrik.carl.eriksson@gmail.com

University advisor:

Nina Dzamashvili Fogelström

School of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona
Sweden

Internet : www.bth.se/com
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Jan Karlsson & Patrik Eriksson
jan3karlsson@gmail.com & patrik.carl.eriksson@gmail.com
April 18, 2014

Abstract

As databases grow in size, the need for optimizing databases is becoming a necessity. Choosing the right operating system to support your database becomes paramount to ensure that the database is fully utilized. Furthermore with the virtualization of operating systems becoming more commonplace, we find ourselves with more choices than we ever faced before. This paper demonstrates why the choice of operating system plays an integral part in deciding the right database for your system in a virtual environment. This paper contains an experiment which measured benchmark performance of a Database management system on various virtual operating systems. This experiment shows the effect a virtual operating system has on the database management system that runs upon it. These findings will help to promote future research into this area as well as provide a foundation on which future research can be based upon.

Jan Karlsson & Patrik Eriksson
jan3karlsson@gmail.com & patrik.carl.eriksson@gmail.com
April 18, 2014

Thesis content

Chapter 1 - Introduction

Database management systems(DBMS) are usually the back-end of most software on the market. Therefore new programs that are being developed have a crucial choice to make. Which Database management system should they use with their program and which operating system is going to run this server with the Database management system.

The choice might be a difficult one based on many factors but how does one make the right choice? How do you know what to avoid? These questions are even more important as many Experts have different Database management system and operating system preferences.

The question here lies in the impact of operating systems on database management systems. How do different operating systems affect a database management system and more importantly, how significant this impact is. We believe that this information should be readily available for anyone. The primary goal here is to increase awareness and provide proof on how Database management systems affect different operating systems. There is very limited information on this subject available today and we aim to fill this gap of information.

According to Haran Boral, “A user faced with the task of selecting a system would ideally like to rate the systems on a "DBMS Richter Scale".”[1] Which is very true, a user would like to have some sort of measurement of which Database management systems is better. Just like a person buying an electronic device would like to have information on which one of the devices are better. The only measurement available to the user is some sort of popularity measurement. However while this measurement can be helpful sometimes, it does not indicate why any given database management system is more popular. Also popularity does not usually mean better as well. Usually databases and operating systems are chosen based on how many years of support is offered as well as convenience. We try to broaden this by adding more factors to this comparison.

The three major factors that contribute into the decision to choose the right system for your Database management system:

1. **Performance**: The operating system might also affect performance in a negative way. Slowing down the response time of the Database management system.

2. **User experience**: The installation and maintenance of the Database management system on any particular system can also have a huge impact on the choice.

3. **Scalability**: Database management systems today are put through a lot of stress as software gets bigger and more demanding.

We will focus on the Performance aspect of this comparison. We feel that performance is a key factor in choosing the right Database management system because it will show us the impact of the operating system very clearly. As databases become larger, the need to optimize performance is increasing substantially. According to Scott Fertig and David H. Gelernter from Yale University the need for larger databases becomes bigger in the field of advanced AI[3]. This shows that as databases become bigger, the need for optimizing the databases becomes more important.

Operating systems are the base of any server however more and more companies choose to virtualize their operating systems. The benefits of virtualization typically outweigh the cost for most companies. There seems to be a lack of information on database performance on virtual machines which we will also fill.

We expect that we will see a difference in performance between the different operating systems. The degree of impact is a different matter entirely and pretty difficult to predict. This information will be useful for future research in this area.

In Chapter 2 we will explain and present concepts and terms that are important as well as further elaborate on the issue that this thesis addresses. Then in Chapter 3 we will present and discuss the research questions and research design. Chapter 4 will be dedicated to the literature review results and design. Following that we will present the result as well as design of our experiment in Chapter 5. Chapter 6 will be dedicated to discussing our results and chapter 7 is our conclusion.

Chapter 2 – Background

First we will need to define the terms we are going to use in this paper.

Operating system

An operating system is a program that handle the computer resources and distribute them to other programs. The operating system manage this distribution with a scheduler to avoid conflicts[7].

Difference between OS

Since all of the tested operating systems are different distributions of Linux the key differences are how each of them has chosen to handle optimizations, installations, updates and package handling. Things like file-system and basic memory management are the same in all of the used operating systems.

Furthermore since Ubuntu is known to be the most popular Linux it has the most support and active forums for help. Ubuntu desktop uses a lot of already installed packages that are not necessary for a database server. Those packages are not installed in Ubuntu server. Furthermore it has a long term support of 3 years on desktop and 5 years on server[21]. Ubuntu is based on Debian and uses dpkg as its package manager which gives a broader variety of packages.

CentOS is a popular Linux distribution and its widely used as a server. Similarly to Ubuntu it has a desktop version and a server version. Since its focus is server usage, it sports 10 year long support for the system[22]. It also has very few installed packages. However CentOS is also very conservative in its updates and uses older but more stable packages because it has a focus on stability before performance. CentOS is based on RedHat and unlike Ubuntu uses “RPM Package Manager” as its package manager.

Fedora only comes with a desktop version (note that the server version is currently under development) and similarly to CentOS, it is based on RedHat. However unlike CentOS it updates much more frequently and it is not as conservative as CentOS is. It has no long term support which leads to it not being considered as stable as CentOS. Lastly Fedora is based on RedHat and also uses “RPM Package Manager” as its package manager.

GUI or CLI

All operating systems have a way to communicate either with a GUI (Graphical user interface) or with a CLI (Command line interface).
A CLI uses only text and commands to execute tasks on the operating system while the GUI can be used with a mouse and pictures to ease the process of configuring and executing tasks on the operating system. Although the GUI is easier to understand, most of the operating systems still have a CLI in their system to be able to support all the commands that are unsupported in GUI systems. The GUI is easier to understand but I comes at a cost of performance because the computer is running the GUI as a process that takes up resources.

Linux can support both GUI as well as CLI. All the Unix based systems have a CLI included even if they have a GUI.

Ubuntu uses Unity as a standard GUI. Fedora and CentOS Desktop use GNOME instead.
The differences between GNOME and Unity are marginal but GNOME has a slight performance advantage over Unity. [22]

DBMS
Databases are a collection of data on an electronic storage device, which is specially organized for rapid search and retrieval by a computer[8].

The Database managements system is a software that handles and computes databases. Its a piece of software interface between the user and the database. It is used for quick search and retrieval of information from a database. The DBMS also determines how data are stored and retrieved[9].

There are several standards on how the data is stored. The most common one is the relational database, it is a DBMS-type that saves the data in tables.

A DBMS also provides a language to act as a interface to the data in the database model. SQL (Structured query language) is an attempt to standardize a language for relational databases. This language is used for adding, Searching and removing data in the database[10].

Difference in DBMS
The way to communicate with the database is by a set of commands to select insert and remove data, also known as queries. The database stores all the data on external disks and saves large I/O buffers in memory for handling different queries.

Databases in general are sometimes affected by the operating systems processes. When the database tries to manage the memory and the operating system memory scheduler blocks it or other processes takes the memory, then performance problems might occur. Even the smallest tasks may affect the database performance negatively.

MySQL
MySQL is one of the most popular database management systems in the world. It is a relational databases that uses the SQL standard for their queries. MySQL is open source and its owned by oracle since 2010.
Similarly to other database management systems MySQL has a I/O buffer where it stores data and indexing in memory. It utilizes this memory as much as it can to avoid paging as much as possible.

Paging is when the process is in need of more memory than the server has, which causes it to store blocks of data in the secondary memory (the hard drive). This is to avoid waiting for memory to be free but it is a costly operation for the server.

TPC

TPC is an organization that creates and manage different standards of measuring databases also known as benchmarks.

They have several different benchmarking standards and their two biggest are the TPC-C and TPC-H. The key difference between TPC-C and TPC-H is how they measure their performance as TPC-C uses transactions per second and TPC-H uses queries per second because of this the results between TPC-C and TPC-H should not be compared to each other.

One of their benchmarking standards is TPC-H which examines large volumes of data and illustrates decision support.

TPC-H is a benchmarking tool for decision making in databases. It executes 22 queries with a high degree of complexity in a database modeled after the schema shown in Figure 1.0. This benchmark is made to give answers to critical business questions.

Figure 2: The TPC-H Schema

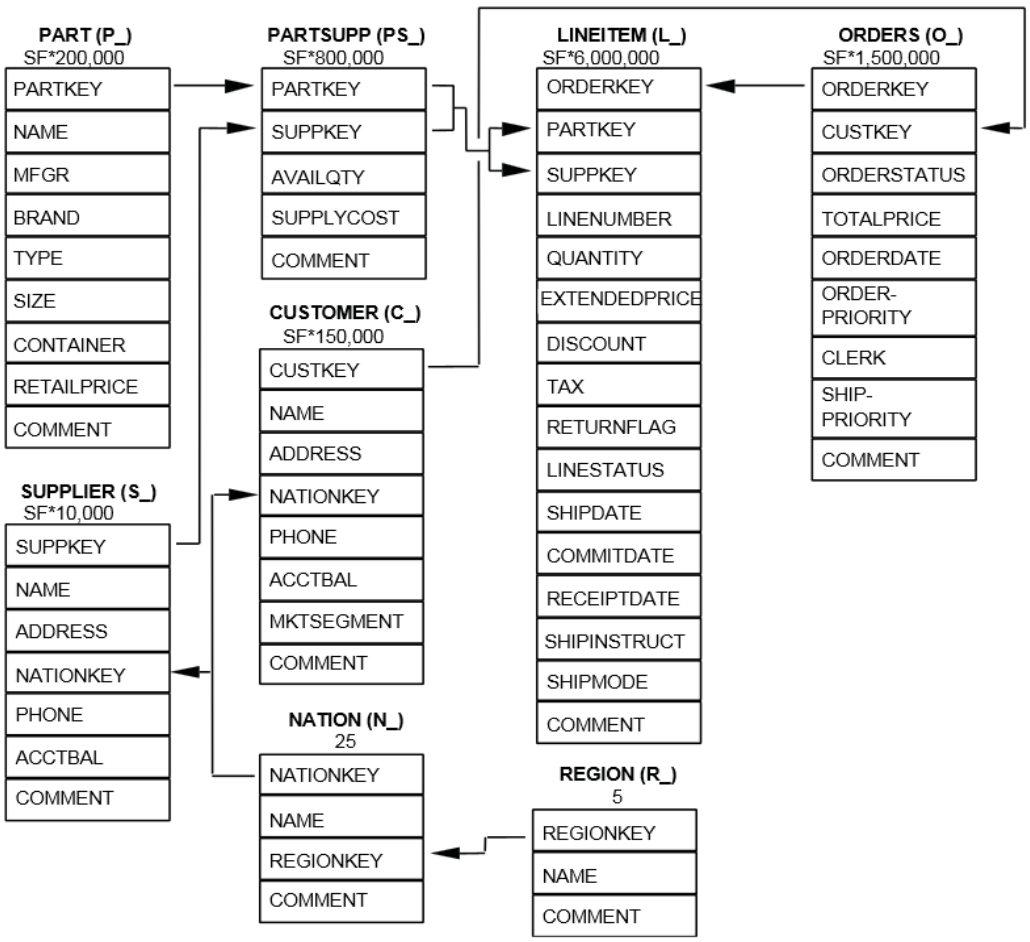

Table 1.0

The number above the tables are the number of rows(or tuples) in that table and SF is the scale factor of the database (how many times the the tables are multiplied to give a larger database).
For a specific definition of TPC and TPC-H look at
http://www.tpc.org/default.asp

HammerDB
HammerDB is a graphical tool to execute TPC-C or TPC-H on different databases like oracle and MySQL. It creates scripts and connections to a given database type and benchmarking model. It provides an easy interface for running this benchmark.
It support scale factor number of threads and logging of the queries time.
It can be used both remotely and locally on the same computer as the database.
See Appendix [19] on how to install and use HammerDB.

Virtual Machines
Virtual machines (VM:s) are an abstract layer between the OS and another computing environment. It gives the possibility to have several OS on one

machine as well as the ability to divide resources between those machines as you please.

A physical computer environment consist of hardware with a software layer on top where the OS and applications are running. A virtual machine is similar except it is running on top of another computing environment. With virtual machines the system is getting the advantages of portability to be packed down and moved to a new location and can help with isolating services from each other[14][18].

Hypervisor VM:s
A Hypervisor VM uses a virtual machine software layer called the hypervisor layer in between the OS and the hardware and gives the appearance to the OS that they are the only running OS on the machine[13].

Hosted VM
Hosted VM uses an application on a physical machine OS and on top of that the virtual OS is provided. So it uses the virtual layer in between the OSes instead of between the OS:S and the hardware as Hypervisor VM:s uses.

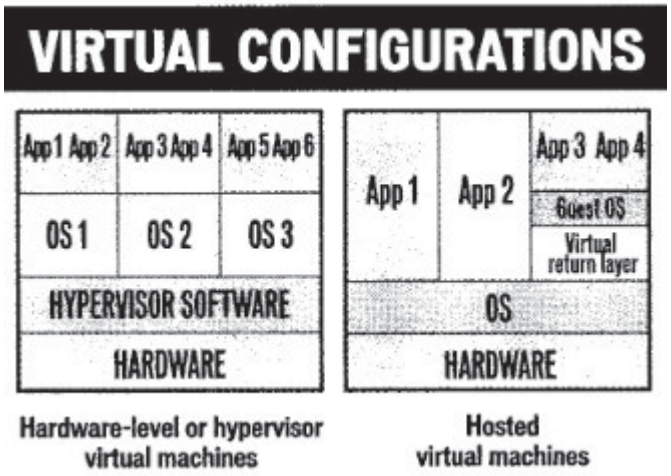

Table 1.1 Picture showing the difference on the layers between Hypervisor VM and Hosted VM (Picture from [13]).

Jan Karlsson & Patrik Eriksson
jan3karlsson@gmail.com & patrik.carl.eriksson@gmail.com
April 18, 2014

Chapter 3 – Research questions and research design

Our research questions are:
Literature study:

1. Does the operating system affect the DBMS?
2. How is it beneficial to use a virtual machine over a physical machine?
3. What type of virtual machine is preferred in terms of performance?

Experiment:

1. Does operating system Ubuntu/CentOS/Fedora affect the DBMS in terms of performance on a hosted type virtual machine.
2. How does Operating system Ubuntu/CentOS/Fedora affect the DBMS in terms of performance on a hosted type virtual machine.

We will use snowball sampling to determine papers that answer our research questions. We will look for keywords that are relevant to our topic and use snowball sampling to broaden the different papers we examine.

Design for the empirical part

Now the question that we want to answer is does the choice of operating system affect the DBMS on a hosted type virtual machine.

We will conduct an experiment on multiple operating systems and how they perform with a database management system. This will give us a clear results on whether or not operating systems have an impact on the performance of the database. The operating systems that we will choose are Fedora, Ubuntu and CentOS. Since we where limited in our resources, we had to use 3 free OS in our tests but this should not affect the outcome as we are interested in the differences between any OS.

As for the choice of DBMS there are 3 leading database management systems right now according to db-engines.com [4]. This site factors in number of mentions on websites as well as Google trends and stack overflow/stack exchange mentions. It also checks for job offers and profiles in professional networks and bases its popularity value on these factors.
The leading DBMS according to this website are:

- Oracle
- mySQL
- MSSQL

Unfortunately MSSQL does not work on Linux, which leaves Oracle and MySQL. MySQL is free which is convenient and it also shows an upwards trend on db-engines.com. These are the reasons we will use MySQL for our experiment.

This experiment will be conducted with the help of virtual machines. The reason for using virtual machines is that virtual machines are being used more and more these days. There has been a surge of companies converting to the more

convenient virtual machines the last couple of years. The reason is not only convenience. According to Umar Farooq Minhas from the University of Waterloo "Virtual machine technologies are increasingly being used to improve the manageability of software systems, including database systems, and lower their total cost of ownership." [2]

We will be using Virtual Box to conduct our experiment. Virtual box provides a thick virtualization. Some might argue that a thin hypervisor environment is better however according to Steve G. Langer and Todd French [6], " the best performance was often seen from a thick virtualization tool (Virtual Box) rather than the thin hypervisor environment."

Our experiment will be similar to Umar Farooq Minhas experiment[2]. We will be using the TPC benchmark just like the aforementioned paper. We will be using TPC-H to test performance.

TPC-H is a benchmark that reflects multiple aspects of the capability of a system. TPC-H provides the framework for the experiment. The TPC-H framework provides data as well as a data structure for the experiment. Once TPC-H has created the database we will use HammerDB to measure the performance of the system based on the 22 queries. HammerDB provides the interface to run and build the TPC-H script. We will measure the amount of time it takes to run the TPC-H script in how many seconds those queries took divided by the number of users that carried out these queries.

Total time for all queries / Virtual users

TPC-H uses 22 queries to measure this query-per-hour as well as has an option to configure how many threads are run at the same this. All 22 TPC-H queries are designed to test as much of the system as possible. The number of threads simulates multiple virtual users that run these 22 queries at the same time. For example with 10 threads the 22 queries are run parallel 10 times. Then the result is divided by 10 to normalize it against the single thread results.

First we need to test the difference between remote and local. This is important because we need to understand what impact remote has on the server. Then we will compare CentOS DB versus CentOS Desktop. This is done to understand the difference in impact between DB and Desktop variants. Since the only difference between a DB version and a desktop version is often the software that is installed we expect only a minor hit in performance due to the GUI running in the background. Once we have gathered this information we will begin with the main test. This test includes all the three major databases which are all run and compared.

All of these tests are run with 1, 5 and 10 virtual users. This is done to get many results as well as provide information on result difference between load and thread count. It also simulates multiple queries happening at the same time which is pretty common for commercial databases. First the test is run once to

warm up the systems buffer pool as well as the file-system cache. Then all 22 queries are run 3 times per number of thread count and the median is taken.

Setup
Virtualbox = All the systems was setup with 64 bit versions. The system is getting 2GB ram and 12GB virtual hard drive with VDI(Virtual Box Disc Image) and is dynamically allocated.
MySQL = After MySQL are installed we changed the file my.cnf the bind-address to the internal ip instead of local-host so you can connect to it remotely.

Independent Variables:

V1	Virtualized operating system	Ubuntu 14.04	Ubuntu 14.04	Fedora 20.1	CentOS 6.5	CentOS 6.5
v2	Desktop/Server	Desktop	Server	Desktop	Desktop	Server
v3	Remote/Local	Local	Remote	Local	Local/Remote	Remote
v4	Nr of threads	1/5/10	1/5/10	1/5/10	1/5/10	1/5/10

Controlled Variables:

Controlled		
v5	Database management system	MySQL 5.6.17
v6	Hardware	Intel(R) Core(TM) i7-4700MQ 2.4GHz 8GB RAM
v7	Networking	Qualcomm Atheros AR8171/8175 PCI_E Gigabit Ethernet Controller (NDIS 6.30)
v8	Benchmarking	TPC-H
v9	Benchmarking Software	Hammer DB 2.16
v10	Virtualisation Software	Virtualbox 4.3.8
v11	Hosted Operating System	Windows 8 64x

Dependant Variables:
Performance
We will be using HammerDB to measure performance by comparing TCP-H script runtime. HammerDB runs all 22 queries and takes the total time it takes to complete them. This unit of time is what we will measure to show the speed of the database.

Validity Threats

Unfortunately we have limited resources which has led to not being able to execute the experiment on physical machines. Further research can be made into exploring the relationship between physical machines and virtual machines.

One Threat is that operating systems have different configuration options in their installation. The most important configuration is that some of the operating systems that we are using have a special OS version that is specifically designed for database/server use. To avoid this validity threat we will be running tests for both the server/database system as well as the desktop version. This will give us the difference between Desktop and server as well as guarantee that our data is usable. Furthermore we will also make sure to use the exact same configuration and the same hardware on all systems.

Furthermore our data collection will be mostly carried out by one person. This leads to the risk that a lone reviewer can interpret the data incorrectly. Therefore we make sure that results are discussed between at least 2 researchers to avoid this threat.

Another threat is managing sheer number of variables that we have in our experiment. We have to make sure that all of our controlled variables are really controlled and do not change from one run to the next. We put a lot of effort in researching all of the variables that can affect database performance and we made sure that we control all variables that can affect performance.

Chapter 4 – Literature review results

For this thesis we used three search engines to find our results. These search engines are very versatile and use many different databases for their results. The three databases that we used are:

- Google Scholar
- BTH Summon (Blekinge institute of Technologies)
- Encyclopedia Britannica

These search tools include hundreds of databases. However the most important for our purpose are:

- ACM Digital Library
- Acronym finder
- IEEE Xplore
- Springer
- CiteSeerx

For a more detailed list please refer to [5].

The aforementioned databases are very relevant when it comes to the subject of computer science. They specialize in engineering and therefore seems to yield to most relevant results.

For our search process we used multiple keywords. The following keywords were used to find references:

- DBMS
- Database Management system
- Virtual Machine
- VM
- Operating system
- OS

These keywords were used to find relevant references that would help strengthen our statement and answer our research questions. The process we used to determine relevance of any given paper was to read the abstract and conclusion. When reading both the beginning and the end of the paper we could form an educated guess on what the paper was trying to accomplish.

The search strings that were used were:

Search Phrase	Result	Snowball	DB	Total
DBMS		2	2 Bth sum	
DBMS Survey			Bth sum	
most used DBMS			Bth sum	
Operating System, Virtual Machine		2	Bth sum	
Virtual machine		1	Bth sum	
database management system		1	Google Scholar	
database system on virtual machines		1	Google Scholar	
database virtual machine		2	Google Scholar	
database operating system		1	Google Scholar	
Database performance on Virtual machines		2	6 Bth sum	
operating system database data structures		1	Bth sum	
most popular DBMS		1	Google	
operating system		1	1	
Database		1	Encyclopedia Britannica	
SQL		1	Encyclopedia Britannica	
operating system		1	Encyclopedia Britannica	
		18	9	27

Table 1.2

As the picture above shows, database performance on virtual machines yields the most results with DBMS as second.

We have categorized the papers that we have found during our literature study in three categories:
DBMS, Virtual Machine
DBMS, Operating System
Operating System, Virtual Machine

Number of papers	16
Number of papers about OS and DBMS	4
Number of papers about VM and DBMS	8
Number of papers about OS and VM	4
Oldest paper	Operating system support for database management published 1981
Newest paper	Virtual machine consolidation based on interference modeling published 2013

Table 1.3 The table gives a short overview of the papers that we have found during our literature study.

For detailed info about the papers see appendix A.

Does the operating system affect the DBMS?

According to Giceva et al.[11]. database managements systems and the operating system has a fragile relationship as even small OS-related tasks can impact the performance of an otherwise scalable database, because the DBMS is unaware of these tasks. As both are using the same resources but want to achieve two different goals with them.

The operating system just manages the resources but has little knowledge of the programs requirements.
On the other hand the DBMS tries to maximize the performance by having deep knowledge of the transactions and its data.

The complexity in this relationship between the DBMS and the operating system is an old issue that is already presented by Michael Stonebraker [12] in 1981.

In Mohiuddin Ahmed et al.:s[16] experiment he showed that MySQL:s performance was affected by the choice of operating system and that Ubuntu showed better performance than Fedora.

Advantages and disadvantages of Virtual Machines
Since the virtual machine doesn't have full control over the computer, the effects to the database can be uncontrollable.

But according to Umar Farooq Minhas et al.[2] the cost of using a virtual machine on a DBMS is low and that the computer hardware they are using isnt fully optimized for virtual technologies, in their experiments they are using a Hypervisor type of virtual machine[17]. However they come to the conclusion that the overhead of virtual machines is outweighed by the benefits.

According to Lixi Wang et al.[15] hosting a database on a virtual machine has great potential to improving the memory and the ease of setting a database up.

What type of virtual machine is preferred?
The best of our knowledge no similar study have been done on a host type of virtual machine and according to Steve G. Langer and Todd French [6] the best performance was often seen from a host type rather than the Hypervisor environment as mentioned earlier.

DBMS, Operating System and Virtual Machine
As most people assume the operating system does affect the DBMS and its performance. What is still an unknown is how the choice of operating system affects the database on a virtual machine. Does the operating system still affects the DBMS or does the virtual machine override the side effects?

The even bigger mystery is whether host type virtual machines affect databases in a similar way as Hypervisor type[6].

This is an unexplored area in terms of effects and performance. What we are doing is scratching the surface of the area in the combination of virtual machines, operating system and DBMS. Our experiment is going to show whether the choice of operating system can effect the DBMS on a virtual machine.

In our studies we have seen that virtual machines are more and more common as they rival physical machines in performance and provide some nice bonus benefits as well. These advantages include the improvement of manageability of

software systems, including DBSM, reduced costs, simplified maintenance. Lastly the virtual machine manager of the host virtual machine provides a layer on top different operating systems. This allows manipulation of resource allocation as well as provides a flexible layer that can be programmed.

We have not found many studies exploring the operating system/DBMS relationship. However those few papers that have explored this fact have concluded that the choice of operating system can have an effect. Mohiuddin Ahmed el al. Showed that Ubuntu 8.04 performed better in certain situations than Fedora 8 for a MySQL database. However one might wonder how this experiment might fare on virtual machines.
We have found that hypervisor virtual machines can have up to 10% overhead because of page faults, system calls and disk I/O, however virtual machines have many other advantages that outweigh this performance hit.

Chapter 5 – Experiment results

To reiterate, The goal of our experiment is to show whether or not the operating system has an impact on the DBMS in a virtual environment. Furthermore it is to show which DBMS is best suited to host MySQL in a virtual environment.

HammerDB was used as a graphical interface for running TPC-H on the databases. TPC-H was run locally on the machines with 1, 5 and 10 threads at once. However HammerDB does not currently support command line executing for MySQL. This limited us in getting only data from Operating systems that had a GUI unless we run the tests remotely from a different machine. We decided to use CentOS DB and CentOS Desktop and run both remotely to see the differences between the two. This will show us the fluctuation that comes with running CentOS remotely over the internet.

For reading the tables:
The results shown are the sum of the runtime of the 22 queries divided by the number of threads. The values are shown in seconds.

NR. of Threads	Local Cent OS desktop (sec)	Remote CentOS desktop (sec)
1	180	185
5	183.6	185.56
10	186.01	189.94

Table 1.4 Comparison between Local and remote controlled CentOS

These results have shown that remote servers slightly increase overhead over local servers which is what we expected. Since the networking between the server and remote access can cause small delays. As traffic goes from one

computer to the other this delay can cause a slight overhead. Now the earlier question was, what is the performance difference between CentOS Desktop and CentOS DB. Both systems were run remotely due to the inability to run locally on CentOS DB.

NR. of Threads	Remote Cent OS desktop (sec)	Remote Cent OS DB (sec)
1	185	174
5	185.56	181.48
10	189.94	183.4

Table 1.5 Comparison between remote Desktop and DB CentOS

The results in the table above show that the database server has a small performance increase most notably with 1 thread. The difference between the server and the desktop version is the package of software that is installed with the OS. The desktop version has many more programs that are installed with the OS which might be the cause of this small performance decrease. Specifically the GUI process that runs during the tests can hit the performance of the machine.

Now we will use the CentOS Desktop(local) results and compare them with Fedora as well as Ubuntu.

NR. of Threads	Local Ubuntu (sec)	Local Fedora (sec)	Local CentOS Desktop (sec)
1	194	166	180
5	210.24	173.48	181.45
10	199.01	173.39	186.01

Table 1.6 Comparison between GUI versions

Our experiment has shown a clear difference in performance across all three platforms. In terms of performance Fedora have shown the best results followed by CentOS and then Ubuntu. The difference in performance shows clearly that operating systems have an impact on the database that is running on top of the system. Ubuntu is showing very slow performance compared to Fedora and CentOS while the difference in performance between CentOS and Fedora is not as substantial.

NR. of Threads	Remote Ubuntu server (sec)	Local Fedora (sec)	Remote CentOS DB (sec)
1	181	166	174
5	175.8	173.48	181.48
10	178.8	173.39	183.4

Table 1.7 Comparison between Ubuntu server,Fedora and CentOS DB.

While the last comparison in Table 1.6 showed a significant difference between the Operating systems this comparison is not as obtuse. Fedora still leading with the best performance followed by Ubuntu server and CentOS. However note that in this comparison Fedora is not remote and therefore cause a delay of around 4 seconds as we have shown in Table 1.4. This would still give Fedora the lead, however a much smaller lead than before.

Fedora also has a GUI running in the background which might cause a performance hit since Desktop versions performed worse in our experiments. Currently a Fedora server version is in development and it will be quite interesting to see the performance of that version.

All of our data can be found in appendix B.

Chapter 6 – Data synthesis and answer to research questions

One of the major questions that this paper explored was whether or not the Operating system affects the DBMS that runs on a virtual operating system. Our experiment clearly shows that this is the case. We have not found any papers discussing the subject of operating systems impact on a virtually hosted Database Server. The few papers that we have found were focused on hypervisor virtual machines instead of hosted virtual machines.

MySQL performed vastly different between the operating system with a 19.32% difference between Fedora Desktop and Ubuntu Desktop.

A number of different sources have found this performance advantage in fedora including the FS-MARK benchmark[20]. The FS-MARK benchmark shows a lot higher results for fedora. FS-MARK is a test designed to measure file-system performance. A better I/O performance can be the advantage that leads to fedora being faster.

As described in chapter 2 Fedora is more frequently updated that CentOS but has no long term support as CentOS has a long term support of 10 years but are more conservative in its updates as its priorities stability before performance. Ubuntu is an in-between here as they have 5 years long term support on their server versions and 3 years on the desktop version.

According to Mohiuddin Ahmed el al. “It is very clear that fine tuning and selection of operating systems plays a very important role on server performance” which our research further confirms[16]. However they have found that Ubuntu 8.04 performed better than FC8 on MySQL. We were using much higher versions of Ubuntu and Fedora which might lead to the conclusion that maybe the newer versions have shifted the results. Another possibility might be that Fedora handles virtualization better since our experiment has been conducted on virtual machines. Note that they are running a transaction based benchmark opposed to our tests which are query based. Which means that one possibility could be that Fedora is better at queries while Ubuntu is better at transactions. Further research could be done on this to determine the exact cause.

Lastly, we have also discovered that server variants of operating systems performed slightly better than their Desktop counterparts on virtual machines.

Jan Karlsson & Patrik Eriksson
jan3karlsson@gmail.com & patrik.carl.eriksson@gmail.com
April 18, 2014

Chapter 7 – Thesis conclusions, future work and contribution

Conclusion

In this paper we have examined three different Linux distributions and their impact on MySQL running on a hosted-type virtual machine. We demonstrated that the choice of operating system has an significant effect on the Database that runs upon it. The DBMS have shown very different results when run on different Operating systems.

In our experiments, we have discovered that Fedora has the best performance with TPC-H. Note that at the time of writing this paper, Fedora does not have a server version and still showed the best results. The Fedora server version is currently under development and these results show promise for the Fedora server variant.

Our experiments have also confirmed that server variants have a slight edge in TPC-H over their Desktop counterparts. We hope that our findings will encourage and promote more research in this area.

Future Work

Possible directions for future work include carrying out this experiment with different benchmarks. For example TPC-C or other workloads. We believe it would be interesting to see whether the results will show a similar difference. Another direction might be to carry out this experiment with physical machines and measuring the impact of operating systems on DBMS with physical machines. Lastly more operating systems or DBMS can be tested and compared. We have used only Linux distributions in our test but it would be interesting to see the results of non-Linux distributions.

Contribution

We have shown that performance difference between operating systems were significant on a virtual machine.

When we started our research a simple question like “Does the operating system affect the DBMS” seams to have a quite obvious answer but that is not the case. We found out that most people just said that yes it affects but there little discussion to why and what is preferable in that case.

This information was mostly missing. We filled this gap of information because we would like to inspire new research in this field. We also want to provide awareness of the issue that the choice of operating systems plays an integral role and further research can help with this choice.

Chapter 8 – References

[1] Haran Boral (1984), A Methodology for Database System Performance Evaluation
http://pages.cs.wisc.edu/~dewitt/includes/benchmarking/sigmod84.pdf

[2] Umar Farooq Minhas, Jitendra Yadav, Ashraf Aboulnaga, Kenneth Salem (2008) Database Systems on Virtual Machines: How Much do You Lose?,
https://cs.uwaterloo.ca/~ashraf/pubs/smdb08overhead.pdf

[3] S Fertig, D Gelernter (1991), "FGP: A virtual machine for acquiring knowledge from cases" http://www.ijcai.org/Past%20Proceedings/IJCAI-91-VOL2/PDF/029.pdf

[4] Popular database toplist http://db-engines.com/en/ranking

[5] BTH summon https://www.bth.se/bib/databaser.nsf/search.xsp/database?lang=en

[6] Steve G. Langer and Todd French (2011) Virtual Machine Performance Benchmarking http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3180542/?report=reader

[7]
The Editors of encyclopedia Britannica
http://global.britannica.com/EBchecked/topic/429897/operating-system-OS

[8]
The Editors of encyclopedia Britannica
http://global.britannica.com/EBchecked/topic/152195/database

[9]
The Editors of encyclopedia Britannica
http://global.britannica.com/EBchecked/topic/152201/database-management-system-DBMS

[10]
The Editors of encyclopedia Britannica
http://global.britannica.com/EBchecked/topic/569684/SQL

[11]
Jana Giceva, Tudor-Ioan Salomie, Adrian Schüpbach
Gustavo Alonso, Timothy Roscoe (2013)
COD: Database / Operating System Co-Design
http://www.inf.ethz.ch/personal/troscoe/pubs/cidr13-cod.pdf

[12]
Michael Stonebraker (1981)
Operating system support for database management
http://dl.acm.org.miman.bib.bth.se/citation.cfm?id=358703

[13] Jan Matlis (COMPUTERWORID April24.2006)
Virtual Machines
http://uv3sv3ds3g.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info:sid/summon.serialssolutions.com&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.atitle=Virtual+Machines&rft.jtitle=Computerworld&rft.au=Jan+Matlis&rft.date=2006-04-24&rft.pub=Computerworld%2C+Inc&rft.issn=0010-4841&rft.volume=40&rft.issue=17&rft.spage=38&rft.externalDocID=1040943191&paramdict=en-US

[14]
Steve Langer, Nick Charboneau, and Todd French (2010)
DCMTB: A Virtual Appliance DICOM Toolbox
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3046693/?report=reader

[15] Lixi Wang, Jing Xu, Ming Zhao, Yicheng Tu, José A. B. Fortes (2011)
Fuzzy Modeling Based Resource Management for Virtualized Database Systems,
http://ieeexplore.ieee.org.miman.bib.bth.se/stamp/stamp.jsp?tp=&arnumber=6005366

[16]
Mohiuddin Ahmed, Mohammad Moshee Uddin, Saiful Azad, Shariq Haseeb(2010)
MySQL performance analysis on a limited resource server: Fedora vs. Ubuntu Linux
http://dl.acm.org.miman.bib.bth.se/citation.cfm?id=1878641

[17]
Umar Farooq Minhas (2007)
A performance evaluation of database systems on virtual machines
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.111.6716

[18]
Steven M. Bellovin (October 2006/Vol. 49, No. 10 COMMUNICATIONS OF THE ACM)
Virtual machines, virtual security?
http://dl.acm.org.miman.bib.bth.se/citation.cfm?id=1164414

[19]
HammerDB Performance Tool
http://hammerora.sourceforge.net/index.html

[20]
Performance between Fedora and Ubuntu
http://openbenchmarking.org/result/1310290-SO-FEDORAUBU35

[21]
Ubuntu Long Term Support

https://wiki.ubuntu.com/LTS

[22]
CentOs Wiki Long Term Support
http://wiki.centos.org/FAQ/General#head-fe8a0be91ee3e7dea812e8694491e1dde5b75e6d

[23]
Performance of gnome versus unity
http://www.linuxuser.co.uk/features/gnome-3-vs-unity-which-is-right-for-you/3