Management of virtualized infrastructure for physics databases

2012 J. Phys.: Conf. Ser. 396 052066

(http://iopscience.iop.org/1742-6596/396/5/052066)

Journal of Physics: Conference Series **396** (2012) 052066 doi:10.1088/1742-6596/396/5/052066

Management of virtualized infrastructure for physics databases

Anton Topurov, Luigi Gallerani, Francois Chatal, Mariusz Piorkowski
CERN, IT Department, Geneva 23, CH-1211 Geneva

E-mail: anton.topurov@cern.ch, luigi.gallerani@cern.ch, francois.chatal@cern.ch, mariusz.piorkowski@cern.ch

Abstract. Demands for information storage of physics metadata are rapidly increasing together with the requirements for its high availability. Most of the HEP laboratories are struggling to squeeze more from their computer centers, thus focus on virtualizing available resources. CERN started investigating database virtualization in early 2006, first by testing database performance and stability on native Xen. Since then we have been closely evaluating the constantly evolving functionality of virtualisation solutions for database and middle tier together with the associated management applications – Oracle's Enterprise Manager and VM Manager. This session will detail our long experience in dealing with virtualized environments, focusing on newest Oracle OVM 3.0 for x86 and Oracle Enterprise Manager functionality for efficiently managing your virtualized database infrastructure.

1. Introduction

This paper addresses CERN experience with virtualization of databases, touching their performance and high availability aspect, management of virtualized environment by means of Oracle VM Manager [1] and Oracle Enterprise Manager [2], and performance tests of Single Root I/O Virtualization (SR-IOV) [3] technology available in the recent networking cards.

1.1. Computer Center challenges

Physics metadata was always playing a crucial part for the functioning of HEP laboratories. With soaring complexity and cost of the engineering tools used, the requirements for metadata availability and size are constantly increasing. Most of the HEP laboratories existing computer centres cannot cope with the physics research growing demands, struggling with lack of physical space and insufficient power supply and cooling. At the same time large numbers of servers running dedicated database services are using only a fraction of the assigned hardware capabilities, thus heavily underutilizing existing computer infrastructure. Combining services on shared hardware is not always feasible due to insufficient isolation concerns.

A possibility to address the above problem is virtualization of database servers, which allows better hardware utilization by giving an acceptable level of isolation for running database services on different virtual machines within a single piece of physical hardware.

2. Databases on Oracle VM

CERN Databases group started testing Oracle databases on virtual machines at 2006, doing proof of usability and stability concept of running Oracle RDBMS 10gR2 databases [4] on Xen [5] with Red Hat Enterprise Linux 4 (RHEL4) [6] guests. Even so the first tests were successful; we could not deploy it in production as no official support from Oracle was available for such setups. A year later, Oracle has announced Xen based Oracle VM (OVM) [7] which enabled us to prepare for future deploy and thus start further tests with database performance and Oracle Real Application Clusters (RAC) [8].

2.1. Database performance on Oracle VM

Since server virtualization introduces an extra layer of abstraction (hypervisor), it was clear that performance of Oracle databases will degrade in comparison to databases running on physical hardware. To quantify this, a set of load tests were done using Swingbench [9]. Both machines were running RHEL4 and Oracle RDBMS 10gR2 with identical configurations.

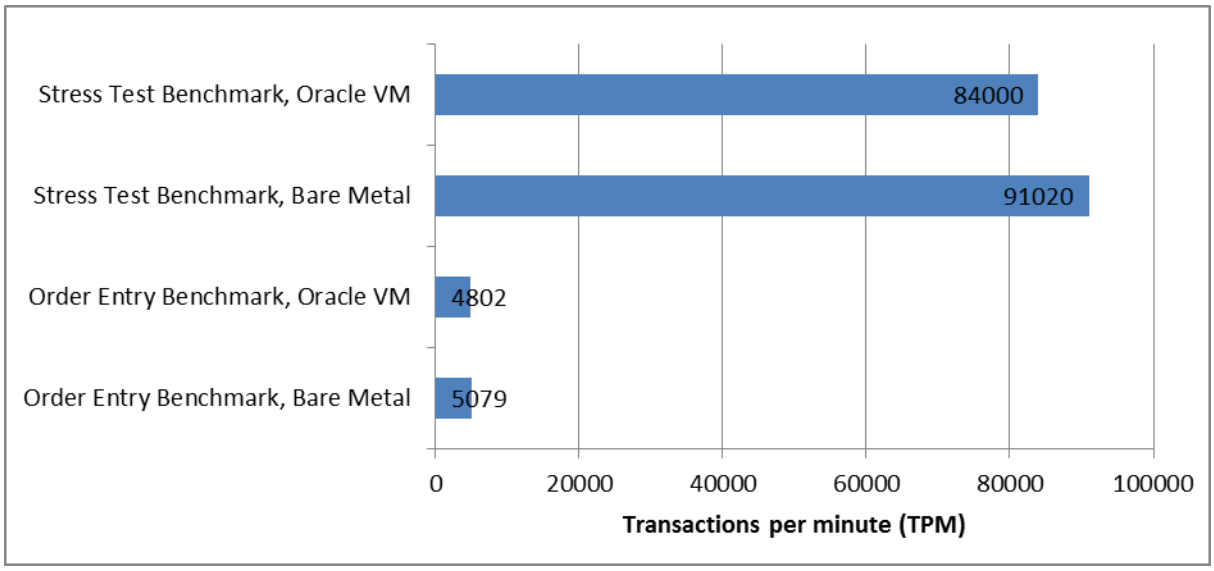

Figure 1. Virtualization overhead on database performance.

As shown on the graph, performance loss varies depending on the type of a load (CPU or I/O intensive), and is on average less than 8%.

2.2. High Availability with Oracle VM

CERN is primarily using Oracle RAC and Oracle Data Guard [10] on physical hardware for assuring high availability of its critical databases. However, less loaded database services do not need the performance scaling of RAC on several machines, the only interest being high availability offered by RAC. In order to consolidate them, RAC can be changed for live migration feature provided by Oracle VM. It allows least possible downtime during move of a virtual machine between different physical hardware. In order to make it to have such migration, all physical machines should have access to shared storage where the image of virtual machine is stored.

On the following graph one can see database performance graphs before, during and after live and standard migrations and respective state of the virtual machine on which the database was running.

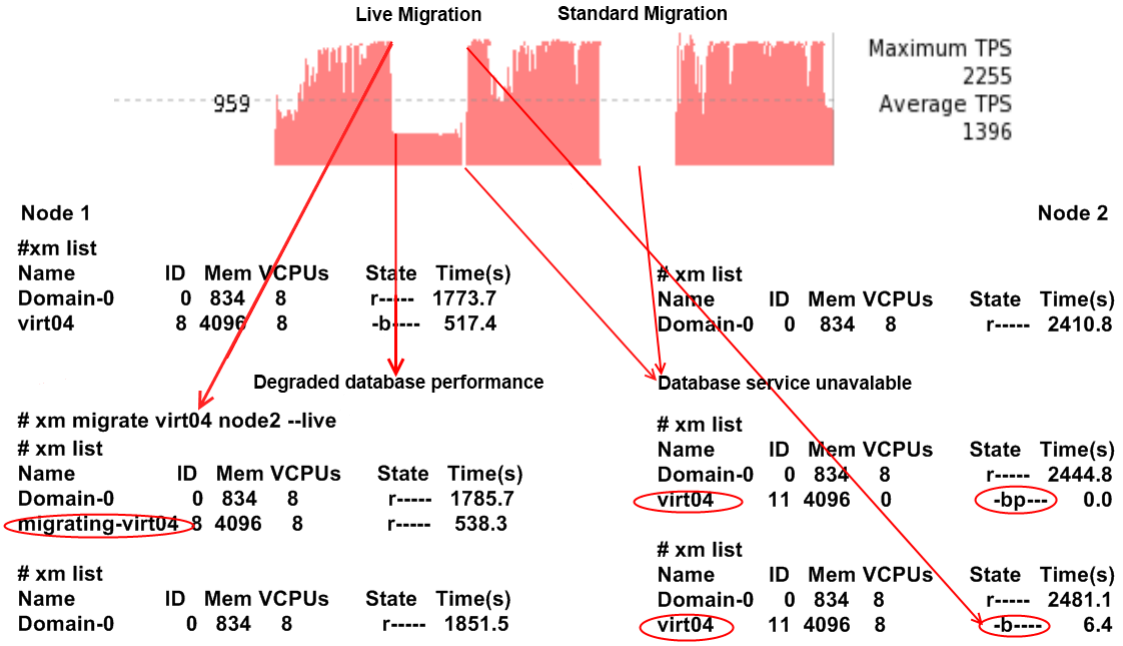

Figure 2. Database performance during live migration.

Just before the migration, the virtual machine is running on the first node. Once live migration is triggered, a new copy of virtual machine is started on second node in the "paused" state (p flag in the state column of the *xm list* output), while the virtual machine on the first node is renamed. Then, a memory image of the virtual machine is transferred to the second node and a virtual machine on the second node becomes active. During the transfer of the memory image the database performance in TPS (transactions per second) degrades significantly, however the database service is still available. During the switch of virtual machines, the database service hangs for few seconds (0 TPS in the graph) ; however connected sessions survive this hang and continue working without the need of reconnect, preserving the state of the transactions. Once the virtual machine on second node takes the active role (the "paused" state is removed), the service performance returns back to normal levels. The second part of the graph shows standard migration; as soon as the migration command is issued, the database becomes unavailable until the migration of virtual machine is finished. More detailed explanation of the output of *xm list* command can be found in the man pages of *xm* [11]

2.3. Management of Virtual Machines for physics databases

There are 3 approaches for managing Oracle VM based virtual machines: Xen command line, Oracle VM Manager, and recently Oracle Enterprise Manager Cloud Control 12c [12] was introduced by Oracle.

Xen command line is flexible and powerful, but is impractical when dealing with bigger quantities of virtual machines. Oracle VM Manager, available since the first release of Oracle VM handles different aspects of virtual machine management like creating and destroying virtual machines, starting, stopping and migrating. The tool has web-based GUI and relies on Oracle VM agent, installed on the target physical machines to execute the commands.

Oracle Enterprise Manager 10.2.0.5 and 11g versions have provided a virtualization management console, which proved to give similar management possibilities to Oracle VM manager, but was lagging from the latter in terms of functionalities, and it was challenging to keep the configuration of these tools synchronized.

For the management of virtual machines, CERN currently uses Oracle VM Manager; however, new Enterprise Manager Cloud Control 12c comes with lots of new features, which proved useful during testing phase at CERN.

The following graph represents the architecture of Oracle Enterprise Manager Cloud Control 12c.

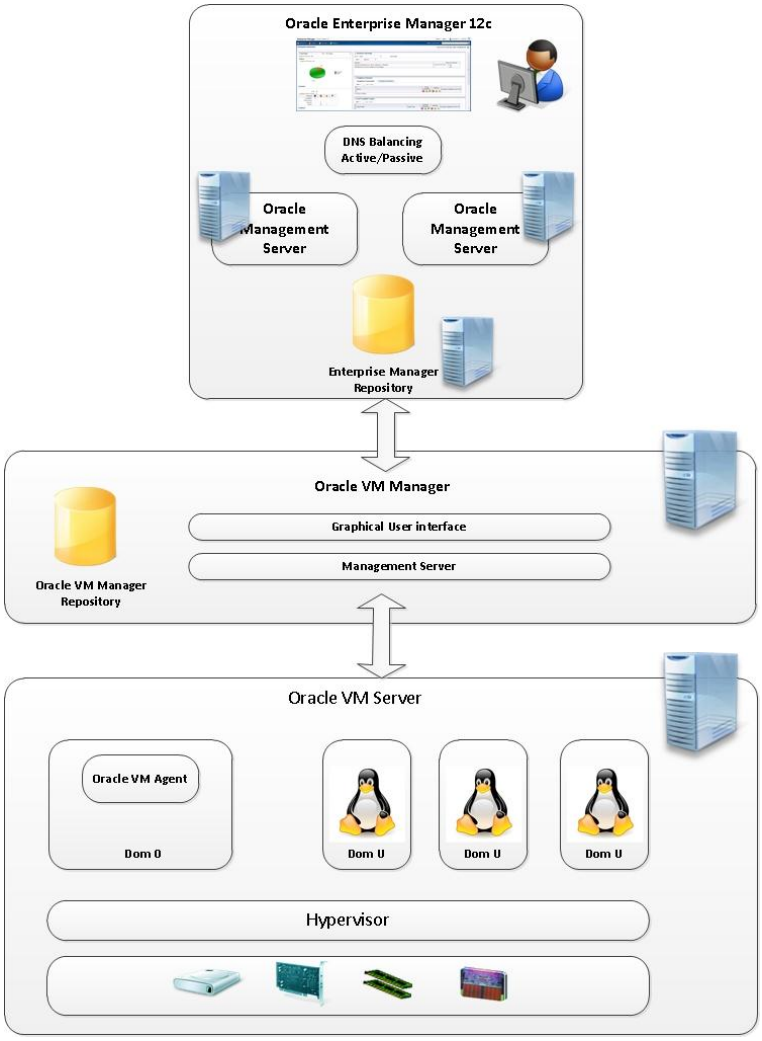

Figure 3. Oracle Enterprise Manager Cloud Control 12c architecture.

The main difference compared with its previous releases is the underneath usage of Oracle VM Manager infrastructure, thus it removes the synchronisation problem of previous versions. The management of networking interfaces of virtual machines is now fully functional, giving the possibility to specify ranges of IPs and MAC addresses to be used during provisioning of new virtual machines. A special attention needs to be paid on the consolidation planner functionality which we found to be very useful. It allows analysis of databases workloads on current physical or virtual

infrastructure and can propose a consolidation plan by mapping this workload to desired destination virtual machines.

2.4. Self Service shift in Oracle Enterprise Manager Cloud Control 12c

Oracle Enterprise Manager Cloud Control 12c comes with a change of management paradigm, focusing on the possibility to offload part of the tasks to power users (self-service users). The following figure is a role diagram of new self-service paradigm.

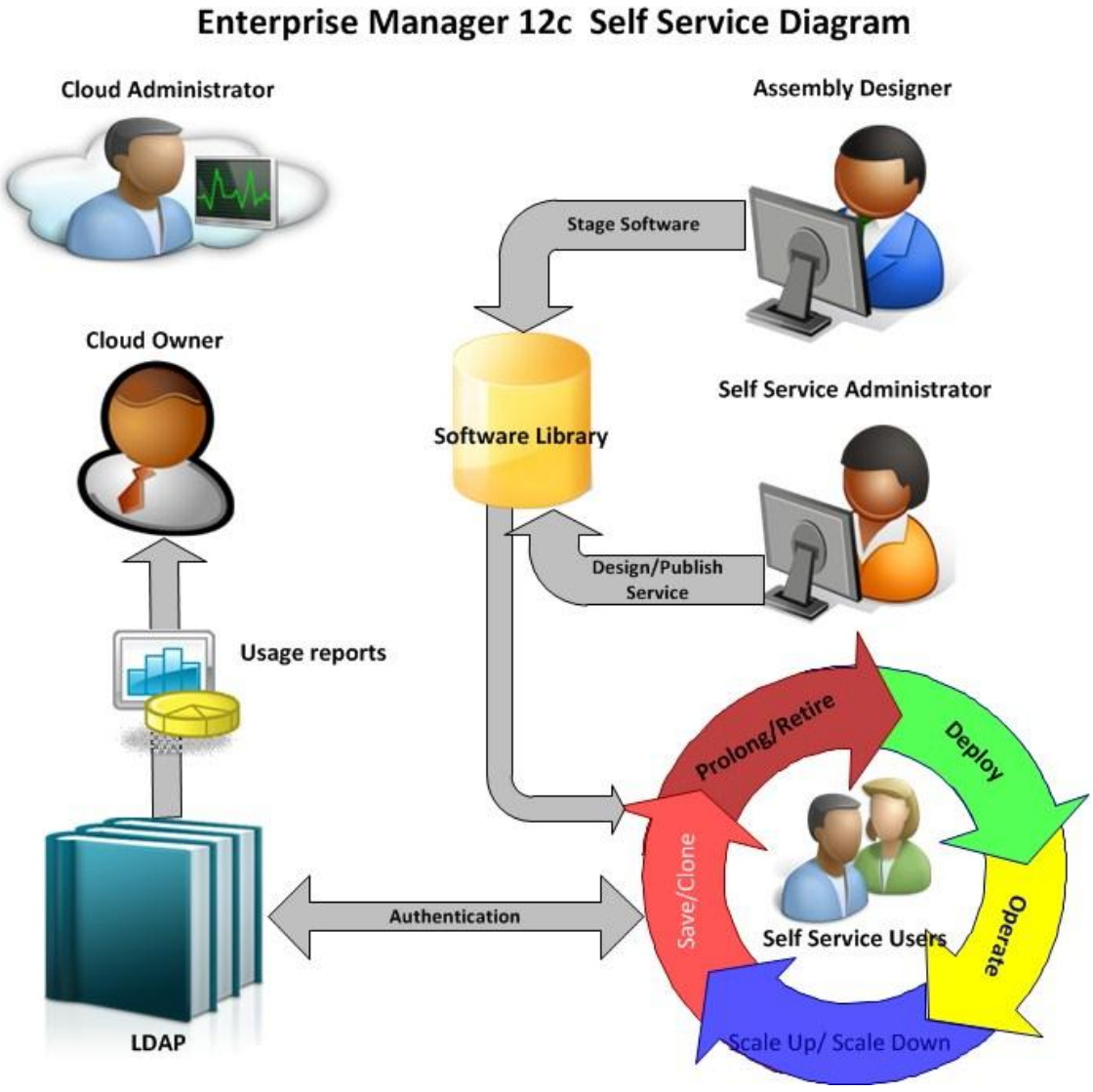

Figure 4. Oracle Enterprise Manager Cloud Control 12c self-service concept.

The setup of the database virtualization environment is done by a person with "Cloud administrator" role, who takes care of installing Oracle VM, grouping physical machines for high availability in pools and zones and configures all other aspects of physical infrastructure. The person with "Assembly designer" role is responsible for staging database and other software, which will be later available for self-service users. The person with "Self-service administrator" role decides on the software availability and quotas for self-service users. Self Service users, depending on the permissions given,

can control the full lifecycle of a database virtual machine. Usage of the self-service infrastructure is reported to the person holding the "Cloud owner" role. More detailed information about the set up procedures and different roles can be found in the Oracle documentation [13].

During the tests of the self-service portal at CERN, it was identified that most of the administrative roles can be safely done by single administrator. Provisioning, operating and retiring virtual machines were stable and aligned to the set-up done on the self-service administrator level.

3. Single Root I/O Virtualization (SR-IOV) with Oracle VM

Since databases are very intensive on I/O subsystem, with most of the storage network based, networking performance of virtual or physical machines is critical for the database services. One of the newest improvements on the virtualization front is Single Root I/O Virtualization (SR-IOV) technology; it enables virtual machines to access network cards directly, without going through the hypervisor, thus improving networking performance. In order to analyse SR-IOV performance, a test environment was set up at CERN, with identical to production database services architecture using virtual machines.

3.1. SR-IOV environment architecture

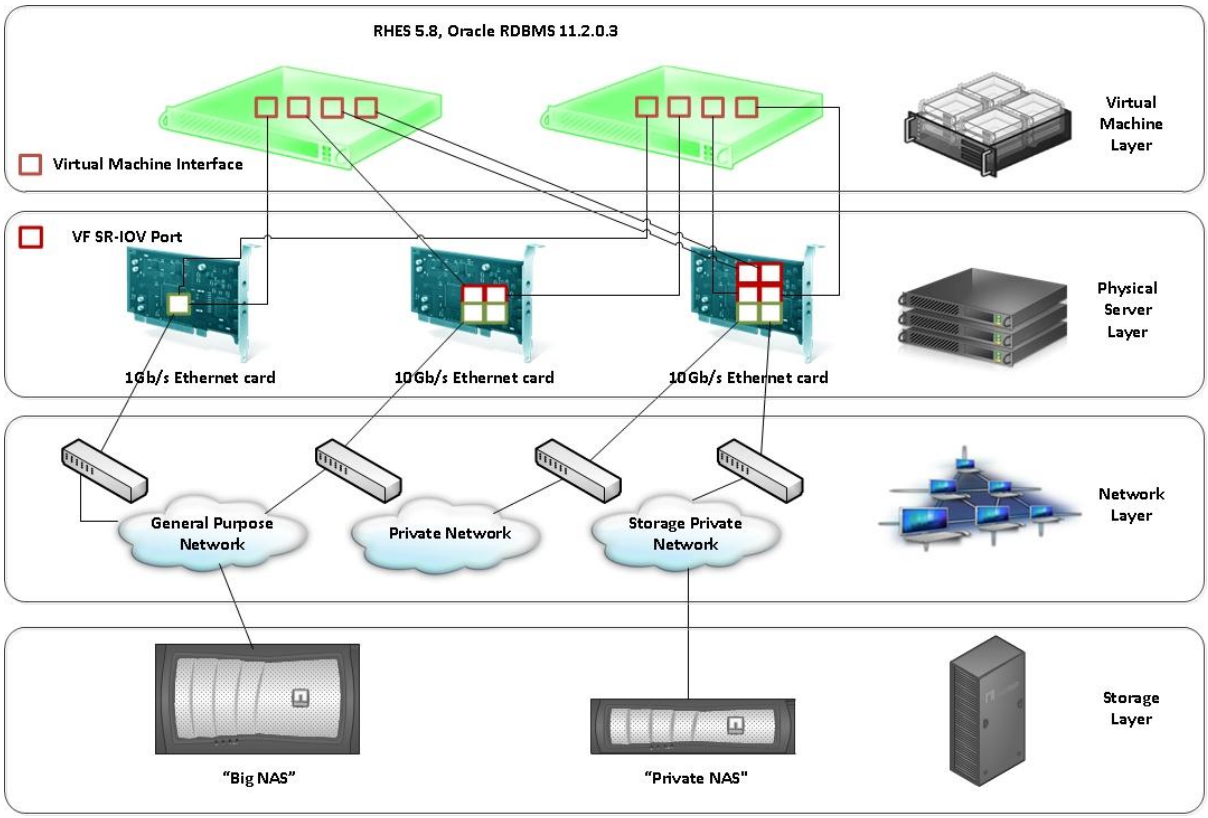

Figure 5. CERN SR-IOV test environment.

As seen at Figure 5, two physical servers were used for the tests. Each server has four Gigabit Ethernet ports and two 10 Gb/s Ethernet ports with SR-IOV feature. Each 10Gb/s card has two ports and is cabled with SFI/SFP+ 10Gb/s network connections. 1Gb/s ports are cabled with standard twisted pair gigabit cables.

Connection scheme is as follows: CERN General Public Network (GPN) is connected via one 1Gb/s port and one 10Gb/s port on first 10Gb/s cards. Servers are reachable through SSH via GPN, which means SSH daemon listens on both 1GB/s and 10Gb/s cards. A second 10Gb/s Ethernet card is used for the private networking. One port is connected to the private NFS-based storage network (Storage), the other is connected to the private interconnect network (Interconnect). Two Network Attached Storage [14] filers (NAS) are accessible from the physical servers. First filer, further called "Private NAS" is attached directly via switch on the storage private network. Second filer, further called "Big NAS" is accessible on the GPN Network. BigNAS filer is reached via a routed network, in 3 hops, which implies higher latency.

Software specifications of the test environment can be found in Appendix A.

Since this paper focus is on SR-IOV performance tests, for general SR-IOV information and setup information you can refer to SR-IOV primer [15]

3.2. SR-IOV tests overview

The aim of following tests is to compare network performance of a virtual machine with SR-IOV enabled vs. SR-IOV disabled. Performance evaluation of virtual machines is always difficult, since many different layers are involved; therefore several iterations of the tests were done in order to eliminate possible side effects of virtualisation architecture as well as caching at various levels.

Two types of tests were setup: simple network transfer speed test and database performance tests.

First test was done by transferring big amount of data to and from NAS filers. Second test was focused on measurement of database performance with I/O intensive workload. Both tests were run 10 times in order to collect trustworthy statistics.

Enabling and disabling SR-IOV was done by switching xen-paravirtualized 10Gbit card with xen pci-attached virtual function card provided by the ixgbe driver, and rebooting the machine in order to changes take effect. This way data passes through the same physical devices and cables, with and without making profit of SR-IOV. Results on virtual machines were validated with the results achieved with the same tests on physical servers.

3.3. SR-IOV data transfer speed test

Many tools can be used to evaluate network performance; however the choice was done for the simplest, most reliable and available in any Linux distributions – the dd command. Since 10Gb Ethernet device can potentially transfer 1.25GB/s, a sufficiently big data source should be found in order to get reliable results. According to following graph, the best dd performance is observed with *4096KB block size.*

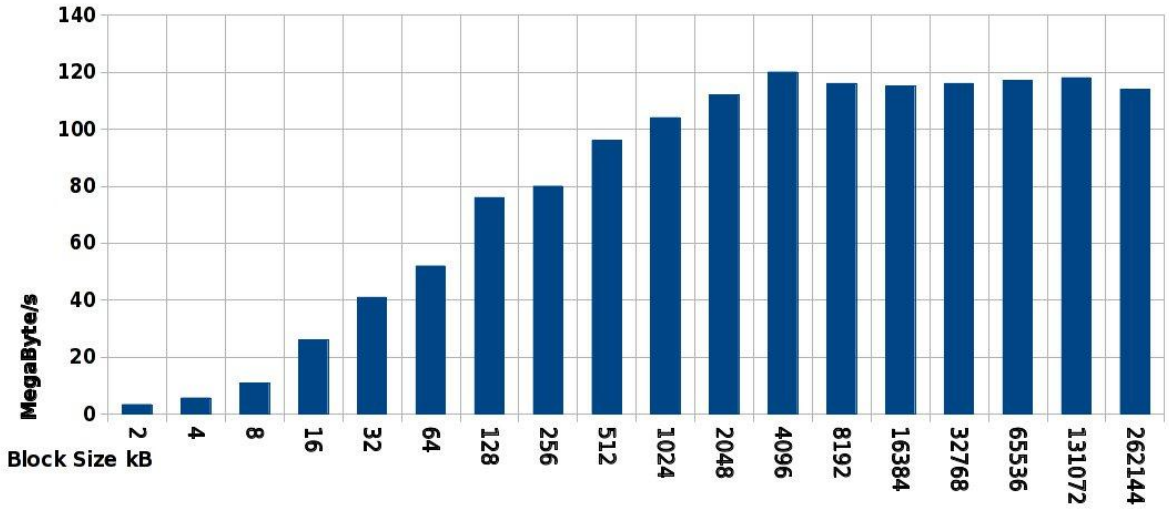

Figure 6. Block size influence on transfer speed.

Any consecutive read will be affected by caching effect; therefore read tests are done every time with a fresh file. The size of the file was chosen big enough in order to avoid any possible cache effect on the NAS filers, and the file is always overwritten before being read again.

3.4. Data Transfer speed test results

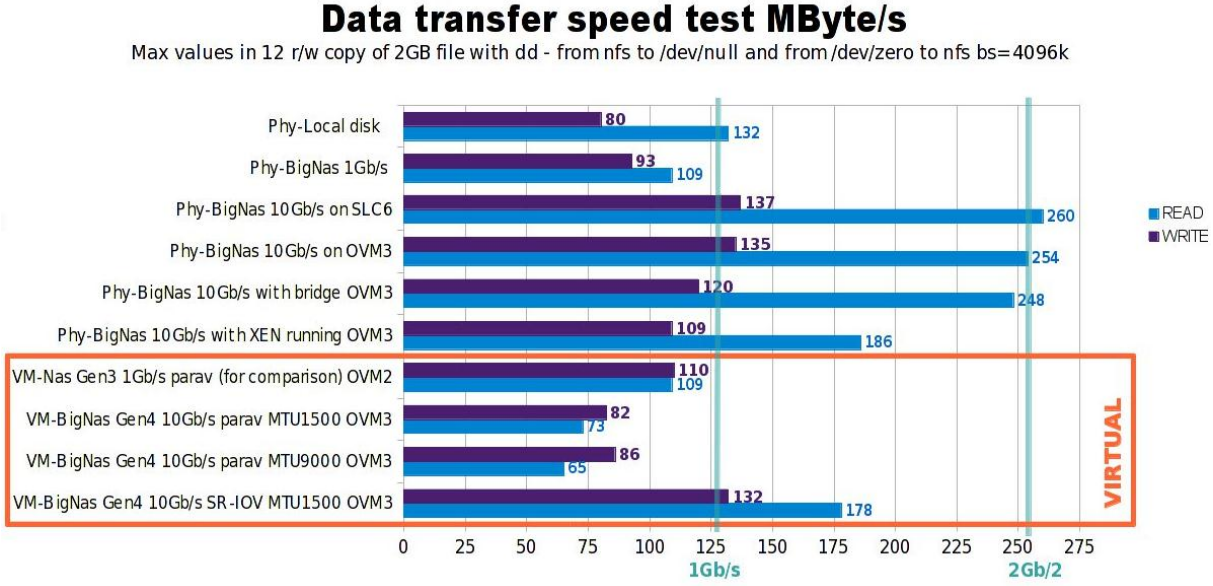

Figure 7. Data Transfer speed comparison.

Figure 7 illustrates the results of the tests done. Since the goal of the tests is to identify the best possible performance of different networking setups, only the best result out of 10 consecutive runs was chosen for the report. The first part of the figure represents measurements for physical servers. Read speed is shown in blue, write speed in violet.

First of all, the speedup from 1Gb/s network to 10Gb/s network on the GPN Big NAS is immediately visible. Big gap is observed in read speed, however same magnitude speedup is not observed in write speed. Tests done on OVM 3.0 and Scientific Linux CERN 6 (SLC6) [16] showed similar results.

The tests done on Private NAS showed instability of the read speed due to production traffic served by these servers; therefore it was decided to not include these results in the paper. Overall, the write speed tests showed speed benefits due to usage of Jumbo Frames which is currently the case in all our production environments.

Since Xen internally uses bridging, it was interesting to identify if bridging is influencing networking performance. Tests done on a physical machine with bridging showed no performance penalty, however running Xen server alone proved to influence the performance.

In order to compensate this negative effect, it was decided for the tests reference a physical machine running Xen to be taken.

Speed test on1Gb/s interface with Xen paravirtualized drivers was set to be the reference point for the virtual machine tests. The 10Gb/s card with Xen paravirtualized drivers performed worse than the 1Gb/s card, showing that that is not a good combination. Once the virtual PCI SR-IOV card was attached to the setup, the performance improved dramatically and passed 1Gb/s barrier on OVM 3.0 based virtual machine.

Comparing this value with the physical setup reference values showed that virtual machine with SR-IOV has a better performance in write test then physical machine.

3.5. SR-IOV database performance tests

Measuring database performance is always a challenging task. Two possibilities exist: replay of a real workload, captured on production database with Real Application Testing (RAT) [17] feature of Oracle RDBMS or executing synthetic load by the means of scripts, measuring peak database performance metrics.

Since CERN databases have different types of workloads, getting a representative workload capture is not possible, therefore a synthetic workload method with scripts was chosen for the tests.

Majority of CERN databases use Network Attached Storage (NAS) and therefore boosting network speed by using SR-IOV should make database I/Os (read and write) perform better. Physical reads can be categorised as single block reads (e.g. data access via index) and multi-block read (e.g. full table scan). Our tests address both cases separately.

For the tests, the latest version of Oracle RDBMS was installed on top of virtual machine. Database performance was measured by executing SQL scripts locally on the database. The script is forcing Oracle database to do full table scan of 250 GB table stored on Big NAS, thus utilizing on maximum network connectivity to the storage.

As seen from the graph below, produced by CERN LEMON Monitoring System [18], network performance on Xen paravirtualized network was peaking at 110MB/s and completed in 42 minutes, while with SR-IOV enabled network performance was significantly higher and reached 390 MB/s, completing the script in 11 minutes. Database software gradually increases network use; therefore having big table gives the possibility to better find peak network speed.

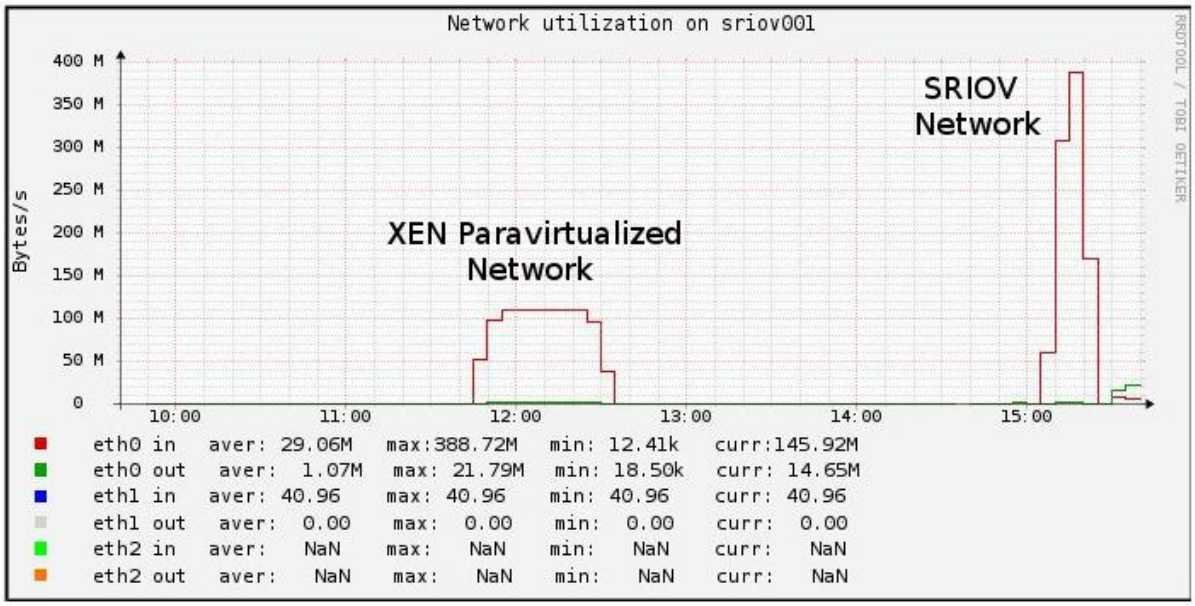

Figure 8. Network speed during full table scan test.

Real database workloads will rarely try doing full table scan of such big tables; therefore an additional test with full table scan of 2GB table was performed (Fig. 9 below), showing factor of 2.5 speedup for SR-IOV enabled system. That means, with NAS and proper network configuration we can exceed the speed of a local drives, even with virtual machine, which can be a game changer for people interested in pooling resources.

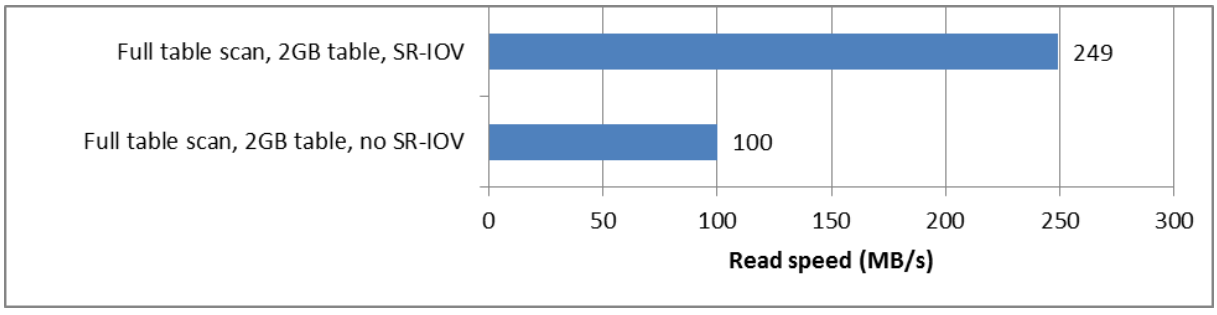

Figure 9. Network speed during 2GB table full table scan.

For the single block read scenario several test cases were done. The tests measured number of single block reads per second (SBR/s) achieved by single non-concurrent user, average SBR/s of single user for a system with 100 and 500 concurrent users and total SBR/s for the system, with and without SR-IOV.

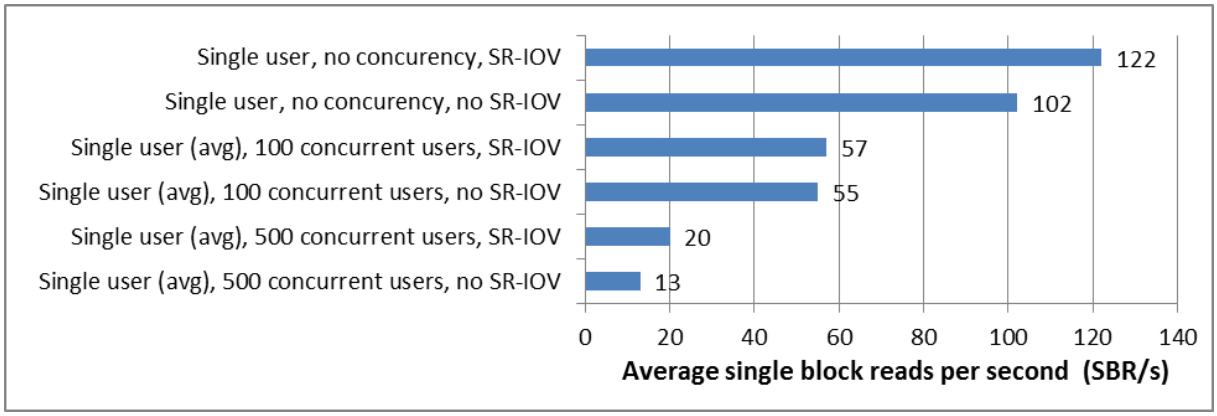

Figure 10. Average random single block read performance.

As seen on the Fig. 10 above and Fig. 11 below, number of single block reads per second improves when SR-IOV is enabled. In its best case, single user with no concurrency, 19.6 % improvement is achieved.

On the total system performance, as seen on the graph below, enabling SR-IOV gives a boost of 46%. From these 2 results we can see SR-IOV brings a real benefit when the system is looking for high performance, either for a single user or on the other hand when many users are using the virtual machine. With a regular charge, the benefits are not so visible.

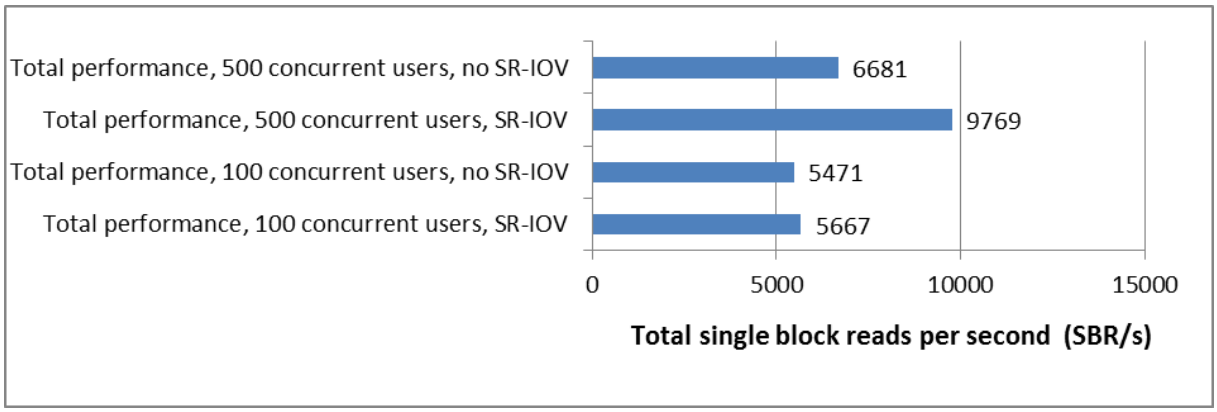

Figure 11. Total random single block read performance.

4. Conclusions

As seen throughout the paper, virtualization proves to be beneficial technology for solving current computing infrastructure problems for HEP institutes. While standard Xen and Oracle VM give a good base for less I/O intensive use-cases, the experiments databases, which are very I/O intensive, were not good candidates to run on virtualized machines. However, the boost of networking performance provided by the latest SR-IOV technology, with gains of up to 60% on writes and up to 150% on reads, as described in this paper, gives a good ground for the future intensive virtualization of all databases.

References

[1] **Oracle, Oracle VM Manager online documentation**
http://www.oracle.com/technetwork/server-storage/vm/overview/ovm-211034.html
[2] **Oracle, Oracle Enterprise Manager online documentation**
http://www.oracle.com/oms/enterprisemanager11g/index.html
[3] **PCI-SIG, Single Root I/O Virtualization online documentation**
http://www.pcisig.com/specifications/iov/single_root/
[4] **Oracle, Oracle RDBMS database online documentation**
http://www.oracle.com/us/products/database/overview/index.html
[5] **Xen.org, Xen hypervisor online documentation**
http://www.xen.org/products/xenhyp.html
[6] **Red Hat, Enterprise Linux pages**
http://www.redhat.com/products/enterprise-linux/
[7] **Oracle, Oracle VM online documentation**
http://www.oracle.com/us/technologies/virtualization/oraclevm/overview/index.html
[8] **Oracle, Oracle Real Application Clusters online documentation**
http://www.oracle.com/us/products/database/options/real-application-clusters/overview/index.html
[9] **Dominic Giles, Swingbench Load Generator, project web page,**
http://dominicgiles.com/swingbench.html
[10] **Oracle, Oracle Data Guard online documentation,**
http://www.oracle.com/technetwork/database/features/availability/dataguardoverview-083155.html
[11] **Xen.org, xm command man pages,**
http://xenbits.xen.org/docs/unstable/man/xm.1.html
[12] **Oracle, Oracle Enterprise Manager Cloud Control 12c online documentation,**
http://www.oracle.com/technetwork/oem/grid-control/overview/index.html
[13] **Oracle, " Infrastructure as a Service" online documentation,**
http://docs.oracle.com/cd/E24628_01/doc.121/e28814/cloud_part2.htm#sthref122
[14] **Wikipedia, Network Attached Storage page,**
http://en.wikipedia.org/wiki/Network-attached_storage
[15] **Intel, "PCI-SIG SR-IOV Primer: An Introduction to SR-IOV Technology",**
http://www.intel.com/content/www/us/en/pci-express/pci-sig-sr-iov-primer-sr-iov-technology-paper.html
[16] **CERN**, **Scientific Linux web page**,
http://linux.web.cern.ch/linux/scientific6/
[17] **Oracle, Real Application Testing online documentation,**
http://www.oracle.com/us/products/database/options/real-application-testing/overview/index.html
[18] **CERN, LEMON - LHC Era Monitoring**,
http://lemon.web.cern.ch/lemon/index.shtml

Appendix A. SR-OIV test environment specification

Latest version of Oracle virtualization software - Oracle VM 3.0 (OVM 3.0) was installed on top of the Dell 810 servers. Based on Linux 2.6.32.21-41xen kernel and runs Xen 4, OVM 3.0 was fit with Intel(R) 10 Gb/s PCI Express Network ixgbe Driver version 3.2.10-NAPI.

Guest virtual machines were installed with the help of Pxeboot. Different operating systems were tested, but for simplicity, this paper will rely on the tests done on Red Hat Enterprise 5.8 with kernel 2.6.18-308.1.1.el5 x86_64, which is currently the standard for CERN production database services.

Appendix B. Transfer speed measurements

Local disk I/O speed was measured using dd with 4096k bs size and -direct flag in order to avoid buffering and be close to database type workload:

 # dd if=/dev/zero of=/tmp/bigfile bs=4096k count=500 oflag=direct

Write transfer of 1.7 GB was done in 22.3327, resulting in 74.4 MB/s or 595Mb/s writing speed. Read transfer, avoiding cache effect, was done with the following command:

 # dd if=/tmp/bigfile of=/dev/null bs=4096k

and showed 1.7 GB copied in 13.2451 seconds, resulting in 125 MB/s or 1Gb/s.

These simple tests clearly indicate that local disk is not suitable as a source or a destination of the transfers evaluating a 10 Gb/s Ethernet card due to its low performance, leaving the choice of memory as the only possible choice:

 # dd if=/dev/zero of=/dev/null bs=4096k

showed 28 GB copied in 4.32346 seconds, resulting in 6.5 GB/s or 52Gbit/s.

This test proves the choice of memory to be suitable both for source (/dev/zero) and destination (/dev/null).

Therefore,

*dd if=/dev/zero of=/NFS/bigfile bs=4096k count=500 oflag=*direct and

dd if=/NAS/bigfile of=/dev/null bs=4096k were used for write and read tests respectively.

Appendix C. Database performance measurements

Database measurements are based on performance views available in Oracle.

V$MYSTAT was used to get the throughput data and *V$FILE_HISTOGRAM* was used for calculating single block read access times.

Simulation of concurrent users was done with the means of RDBMS_SCHEDULER jobs. Single block access is based on block access by rowid from a 1TB partitioned table with 131 million blocks of 8Kb each. For the single user, 100 000 blocks are accessed, for the concurrent users - 10 000 blocks per user. Database caching is minimal in the test case, since more than 99.7% of the reads were physical and the *V$FILE_HISTOGRAM* view reports only them.

Appendix D. Equipment used

Dell PowerEdge R810 physical servers

Broadcom Corporation NetXtreme II BCM5709 1Gb/s Ethernet cards

Intel's 82599EB 10Gb/s cards with SR-IOV support

NetApp FAS 3240 (“Private NAS”)

NetApp FAS 6240 (“Big NAS”)