Energy-Efficient Heterogeneous Cluster and Migration

Jukka Kommeri and Tapio Niemi
Helsinki Institute of Physics
CERN, Geneva, Switzerland
Email: jukka.kommeri@cern.ch

***Abstract*—A recent trend of server consolidation using virtualization has allowed datacenters to improve their server utilization rate and total energy efficiency. Virtualization technologies are used to share physical hardware between multiple services. This has led to another challenge: how to place the virtual servers into the physical servers? Especially, this is important in cases in when the workload of virtual servers is not constant.**

In this paper, we study the overhead of virtual server migration in physics computing on energy efficiency with an emphasis on quality of service. Our method is based on the standard migration technique that allows us to move virtual machines between physical machines without significant interference on the service running on the virtual machine.

Our results indicate that by utilizing dynamic resource sharing among the virtual servers and load balancing between heterogeneous physical machines, it is possible to improve energy efficiency of online cloud services.

Keywords-energy-efficiency, virtualization, migration, heterogeneous hardware, physics computing

I. Introduction

The energy consumption of the information and communications technology (ICT) sector has been rapidly growing for the past decade. This has received a lot of attention and concerns [1], [2]. Much of this is due to over provisioning of hardware to serve the peak loads and provide high service availability. According to many studies, the average utilization rate of a server is around 15% of maximum but it depends much on the service and it can be even as low as 5% [3], [4]. These values are much lower than in the other fields of industry, although peak loads do not only exist in the IT sector.

New technologies, both software and hardware related, have been studied and adopted to face this increase in energy consumption. The two main approaches in this field are 1) energy-proportional hardware, and 2) service consolidation through virtualization. Developing energy-proportional hardware has appeared to be very challenging, especially developing such a technology for the memory has not so far progressed much. On the other hand, just manufacturing computers takes both money and energy, thus keeping servers idle waiting for peak loads is not a profitable approach even with possible future energy-proportional hardware [5]. Instead, server consolidation applying virtualization could solve the problem in some cases. This is already visible, for example Koomey [6] indicates that server virtualization has already decreased the growth in the server market.

Though virtualization is already used to improve the energy efficiency of physical servers, there is still room for improvement. Datacenters need to serve their customers and provide capacity for the peak loads. As our random sample containing a subset of servers in CERN, Conseil Europen pour la Recherche Nuclaire, datacenter log in Figure 1 indicates, the peak loads are surrounded by less busy periods. The cluster needs to adapt for these changes. This can be achieved by dynamically moving virtual machines between physical hosts depending on the load of the virtual and physical machines. There are also big differences among different physical hardware solutions since some of them are better or more suitable for heavy computation while others work more efficiently in light-weight computing.

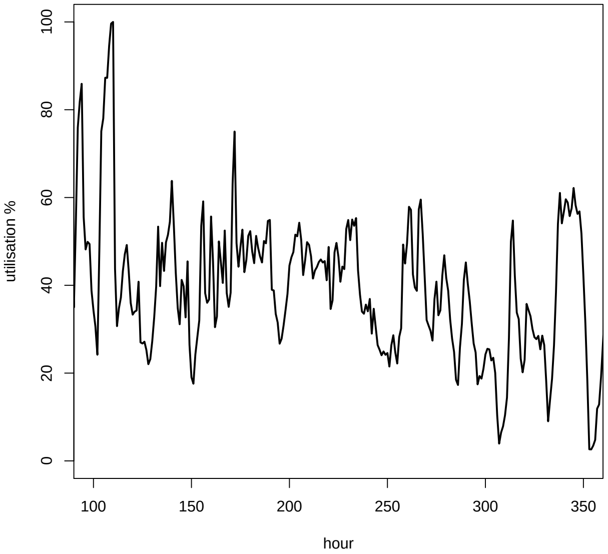

Figure 1. Cluster utilization rate.

In this study, we focus on energy efficiency of migration and how it affects the quality of service of a virtualized service. Our hypothesis is that the energy consumption of the system improves when dynamically placing virtual servers to different hardware based on their current load. We validate our hypothesis by studying the energy consumption of virtualized servers with realistic physics analysis software while migrating the virtual servers between physical servers. Our test set up consists of two types of hardware: one low-performance energy-efficient server to host mostly idle or even dormant virtual machines and one powerful server to host heavily

ISBN: 978-1-61208-332-2

loaded virtual machines. We used an open source virtualization platform Kernel-based Virtual Machine (KVM) and the Compact Muon Solenoid Software framework (CMSSW) [7] as a realistic test case. The same software is used at CERN to analyze collision data produced by the CMS experiment. Our results show that heterogeneous hardware can be used to improve both the energy efficiency of the cluster and optimize the use of floor space.

The paper is organized in the following way. In Section II, we have related work, which is followed by Section III that gives more detailed description on migration. Then, we have our test methodology in Section IV and describe our test environment in Section V. These are followed by test results in Section VI and conclusion in Section VII.

II. Related work

Virtualization itself is not an optimal solution since logical sharing of physical resources with virtualization presents some overhead that depends on both the placement of virtual servers and their workload [8]. This makes the sharing of physical resources between virtual machines challenging, especially, when the load of virtual servers is not constant. A lot of research has been put into virtual machine scheduling, the way how virtual machines are placed in a cluster of physical machines. Many different variations of bin-packing algorithms have been proposed [9], [10], [11]. Some of these algorithms also do load balancing dynamically by moving virtual machines between physical resources. This technique is called migration. Migration gives more freedom in choosing the physical hardware as the virtual machine is not bound to single hardware. It also gives flexibility when managing physical resources.

Some studies already point that exploiting heterogeneous hardware in a virtualized cluster can be beneficial. Hirofuchi et al. [12] use dedicated servers for virtual machines with less load and other dedicated servers for running heavily loaded virtual machines. Virtual machines are moved between these two types of physical servers in function of their processing needs. In their work, the only difference between virtual machine pool server and execution server is the amount of memory. Profiting from energy-efficient hardware was also studied by Verma et al. [9], but in their study the implementation was left at the level of power models where different hardware had different power models. Also, the preference of allocating virtual machines to more energy-efficient hardware has been studied [10], [13].

We extend this thinking by using an energy-efficient server for the virtual machine pool server and a power-efficient server as an executing node. As we have found in our earlier work [8], virtual machines do not consume physical resources when they are idle. This makes it possible to store as many idle virtual machines on a pool server as one can fit in its memory. As the idle servers do not need much computing power, one can choose a more energy-efficient hardware for that purpose. On the execution server one does not need that much memory, but more computing power per virtual machine.

III. Migration

Migration is a technique used to move virtual machines from one physical server to another. In practice, this means that the hard disk, memory contents and processor state is moved from one physical server to another physical server. There are several ways to perform a migration. Choosing the proper one depends much on the way computing environment is set up and what the current use case is. Migration can be done either online or offline. In offline migration, the virtual machine is shutdown for the duration of the state and memory transfer as in online migration this is minimized and the virtual machine does not experience a notable down time. Online migration is also known as live migration. For live migration to work properly, one has to set up a shared virtual machine disk image or a root partition in both physical servers. This can be achieved with various network-attached storage (NAS) services; NFS, iScsi, AoE, Ceph, etc. Having a NAS generally speeds up the migration process. Without a dedicated NAS one would also have to copy the contents of the virtual machine disk image from the original host to the destination host. This takes much more time than just moving the contents of the memory of the virtual machine to a new location.

When making a migration decision, it is good to evaluate the possible parameters that affect migration and also how migration affects other virtual machines or applications running on the servers. Liu et al. [14] have studied how different parameters affect the migration time and energy consumption. They have come up with a model for predicting migration time, downtime and energy consumption. This model is defined by Equations (1,2,3). Most important parameters are memory dirtying rate (D), rate of transmission (R) and virtual machines memory size (V_{mem}). If rate of transmission, i.e., network bandwidth is smaller than memory dirtying rate, live migration is not possible. From these three it is possible to calculate migration time (T$_{mig}$).

$$\lambda = D/R \tag{1}$$

$$n = log_{\lambda} \frac{V_{thd}}{V_{mem}} \tag{2}$$

$$T_{mig} = \sum_{i=0}^{n} T_i = (V_{mem}/R) \times \frac{1-\lambda^{(n+1)}}{1-\lambda} \tag{3}$$

These equations are mainly indicative as many parameters vary during the migration process and due to the choice of the migration type. Usually both V_{mem} and V_{thd} are configured statically. For the live-migration to succeed λ needs to be bigger than than zero. Equation (3) is valid for both online and offline migration where T_0 is the time it takes to migrate the initial memory set. Equation (3) shows the iterative nature of migration process where n is the number of iterations. Configuring V_{thd} can have a big impact on T$_{mig}$ as then the dirtying rate has a bigger impact [15].

As mentioned earlier, virtualization makes it possible to control the load of physical resources by migration of virtual machines. Cluster management systems are based on heuristics that use data gathered from both physical machines and virtual machines. In the simplest form, they make their decision based on how physical CPUs are loaded while the complex ones take into consideration also other issues such as memory usage, network traffic, service level agreements (SLA), server energy consumption, server thermal state, virtual machine intercommunication, etc. [11]. Even the simplest algorithm improves energy efficiency of a cluster that has a varying workload.

Development of the most optimal algorithms have gotten a lot of attention and several different solutions have been proposed. In this section, we explore a few of them. These algorithms vary much in complexity and as there exist no standard way for testing, their comparison is difficult. Also, the fact that most algorithms have been tested with different simulators does not make the comparison any easier. As Srikantaiah et al. [16] write, development of an algorithm is a compromise between time and performance. The algorithms that produce the most optimal results tend to take longer to calculate, e.g., the constraint programming algorithm by Hermier et al. [17]. But even the simplest algorithms such as the one by Shrikantaiah [16], that picks the best server by computing a simple Euclidean distance of the addition of new workload, can improve cluster energy-efficiency as then the system is reacting to the change of resource requirements. However, this Euclidean distance algorithm like many other algorithms, assumes that the requirements of the new task are known. The basic idea of all these management systems is to minimize the use of physical machines and this is done by packing already loaded servers more efficiently.

IV. Methodology

In many services, the workload fluctuates a lot as a function of time. For example, the workload can be near zero during nights and weekends. This is very much the case in our data as Figure 1 illustrates. During off-peak periods, the request rate is just around 1/5 of the peak periods, thus also less computing power is needed during off-peak periods. Based on this observation our hypothesis is:

> Energy efficiency can be improved without decreasing the quality of service by moving virtual servers to an energy-efficient low power server during off-peak periods.

Naturally, we assume that the energy-efficient server is not powerful enough to host virtual servers during the peak hours. Additionally, to simplify our theoretical model, we assume that all virtual servers are always hosted by the same server. Now, we can form our model as follows: Let the idle power of the energy-efficient server A be A_{idle} and full power A_{peak} and for the power server B, B_{idle} and B_{peak}, respectively. Now the upper limit for the energy consumption of running the system x hours will be:

$$E = A_{peak} \times x \times p_A + A_{idle} \times x \times (1 - p_A) \\ + B_{peak} \times x \times (1 - p_A) + B_{idle} \times x \times p_A + n \times E_m$$

Where p_A is the portion of time that the server is hosted on the server A, n is the number of migrations, and E_m is the extra cost of one migration operation.

In our test setting (see Section VI), the values for the parameters are as follows: $A_{idle} = 25W$, $A_{peak} = 78W$, $B_{idle} = 101W$, $B_{peak} = 246W$, and $E_m = 1Wh$.

We compare two cases: 1) the virtual service is hosted all the time on Server B and we do not have Server A, and 2) during off-peak periods, the service is migrated to Server A from Server B. The energy consumption can be computed as follows:

$$\begin{array}{ll} \text{Case 1:} & E_1 = B_{average} \times x \\ \text{Case 2:} & E_2 = A_{peak} \times x \times p_A + B_{average} \times \\ & x \times (1 - p_A) + n \times E_m \end{array}$$

We assume that the load level is high enough to keep Server A almost in its peak power all the time when the service is run on it. Further, we assume that even off-peak load is still high enough to make Server B to consume 75% of its peak power. These assumptions are reasonable as illustrated in Figure 7. It is quite straightforward to see that Case 2 is more energy-efficient if the workload is not very high. Upper and lower bound power values of the servers A and B are measured from real hardware. An estimate of the cost of migration is determined by collecting a large sample of migration results while having realistic workload on the virtual machine.

V. Test environment

Our tests aimed at measuring the energy consumption and overhead of virtualization and migration. As a workload in virtual machines we had real physics analysis applications, that are used at CERN, and data that is produced at CERN particle accelerator, Large Hadron Collider (LHC). We measured how performance was affected by migration.

In our tests, we had separate hardware for basic migration test and hardware comparison tests. In the migration tests, we wanted to have identical servers so that the effect of migration could be measured. In the migration test, we used Dell 210 single processor servers with Intel X3430 processors and 12GB of memory as hypervisors. In the hardware comparison tests, we had a Dell 210 II server with energy-efficient Intel E31260L 2.4 GHz processor and a Dell 415 server with a high efficiency AMD Opteron 4276HE 2.6 GHz processor and 24GB of memory. Servers were connected with a gigabyte network. Power usage data of the servers was collected with a Watts up? PRO meter via a USB cable as it is illustrated in Figure 2. Power usage values were recorded every second.

In migration tests, the hypervisors were connected with a secondary network, which was dedicated to migration traffic. Also, the hypervisors were configured to use all the bandwidth of the migration network. In both test cases the virtual machine

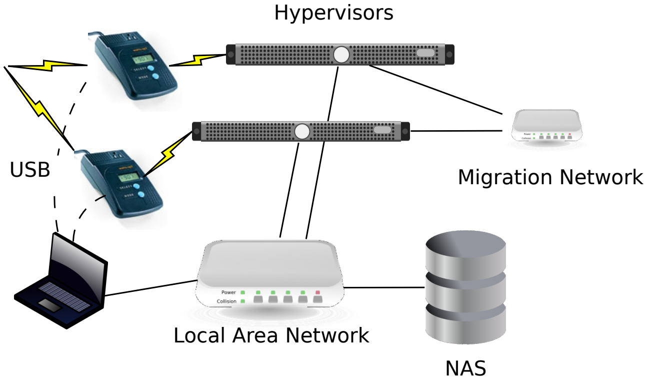

Figure 2. Illustration of test setup.

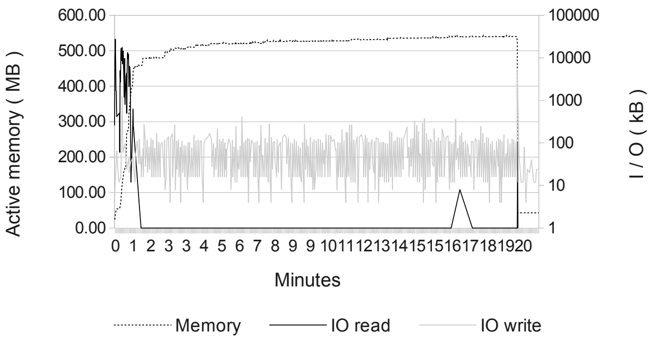

Figure 3. CMMSW task memory and IO curves.

images, hard drives, are on a powerful remote server HP Proliant DL585 G7. A separate network storage server, Dell PowerEdge T710, was used to serve physics analysis software and data. Both network storage servers were connected to the local area network, which is separate of the migration network.

A. Test Applications

The operating system used in all machines, virtual or real, was a standard installation of 64 bit Linux, Ubuntu. Physical servers were installed with Ubuntu version 13.04 as the virtual machines had a version 10.04 of Ubuntu. KVM was used as virtualization technology. Virtual machine images were shared over iSCSI from a network attached storage.

Virtual machines were installed with a separate root that contained Scientific Linux at Cern 5 (SLC5) installation and CMSSW version 4.2.4. From this separate root, CMSSW was run using chroot system call. The separate root was chosen as the SLC5 was not as performant Ubuntu as a virtual machine operating system. For the CMSSW tests real data files produced by the CMS experiment were used. These data files were shared to the virtual machines over a network file system, NFSv4. Execution of the analysis was sped up by limiting the number of analyzed events to 300. Depending on the hardware the execution of single analysis take from 10 to 19 minutes.

Behavior of our workload is shown in Figure 3. These statistics were collected with a Linux tool called Vmstat and measured inside the virtual machine, which was running a CMSSW analysis. The CPU curve was left out as it was almost a straight line from the beginning.

B. Test cases

We divide our tests into two categories. In the first part, we conducted a parallelism test on two different type of servers and in the second part we conducted migration overhead tests with varying workload. In both test cases, the workload used was the same as in previous section. The hardware for the migration test was different as for this test a homogeneous environment is more suitable. Migration overhead was tested by doing the migration at different points of the execution of the analysis software. As the execution on hardware chosen for migration test took about 19 minutes, the migration points 1,5,9,13 and 17 minutes were chosen. Only one migration per one analysis task was run and the test was repeated ten times for all combinations. Also, the same tests were repeated with different background loads. Higher background load was achieved by starting three extra virtual machines on both hypervisors and running two CMS analysis on each one. Making the total load per hypervisor six jobs in addition to the virtual machine that is being migrated.

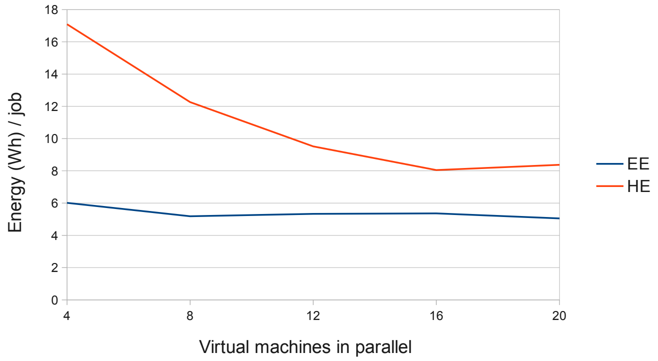

Figure 4. Energy consumption per physics analysis job when running them in parallel and in dedicated virtual machines.

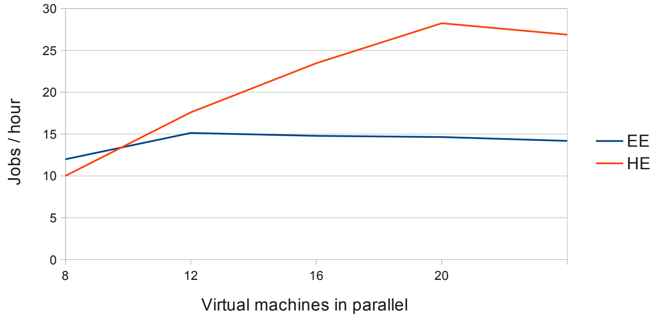

Figure 5. Server throughput when physics jobs in parallel and in dedicated virtual machines.

In the hardware comparison tests, we run the physics analysis task on a different number of virtual machines each running a single analysis task. The same tests with varying load were performed on both types of hardware, the energy-efficient server (EE) and the high efficiency server (HE). Here the energy efficiency means that the server can do more work with less energy as the high efficiency means that the server is able to more work in less time. Test servers were initialized with virtual machines and then the workload was started remotely on each virtual machine at the same time and test would capture the duration from the termination of the last workload. As with all our tests, between every test the servers were rebooted to reset the test environment.

VI. Test Results

At first we measured how different our energy-efficient server is from the high efficient one. Figure 4 illustrates how energy consumption per job changes when parallelism is increased. The two CPU Opteron server never reaches the energy-efficiency of the single CPU server, but its energy efficiency improves when running on higher load. In Figure 5, we have the total throughput of the physical servers with different loads. It shows how the 2 CPU server can handle much more load than the energy-efficient server.

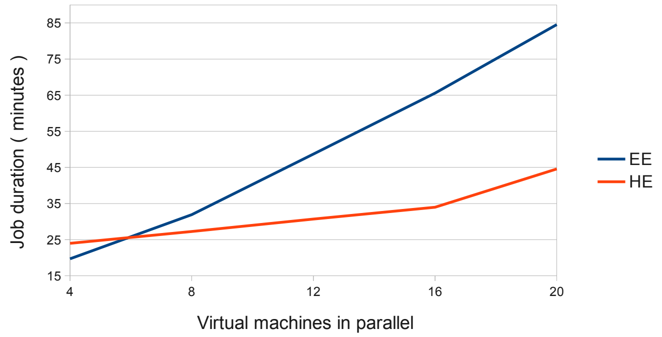

Figure 6. Job duration in minutes with different number of virtual machines running single job each.

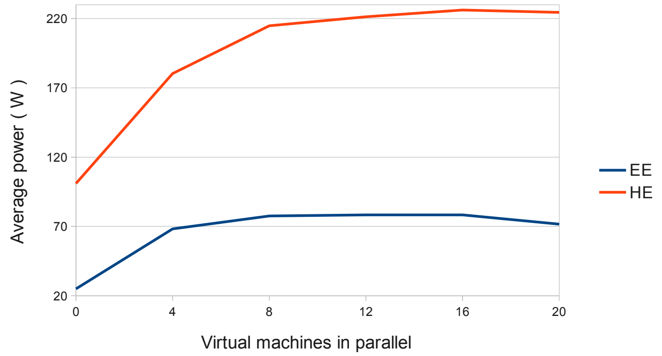

Figure 7. Average power in Watts with different number of virtual machines running single job each.

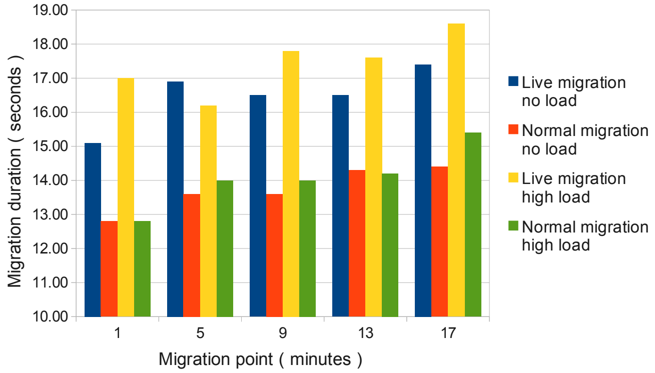

Figure 8. Duration of one migration on different parts of the CMS analysis task.

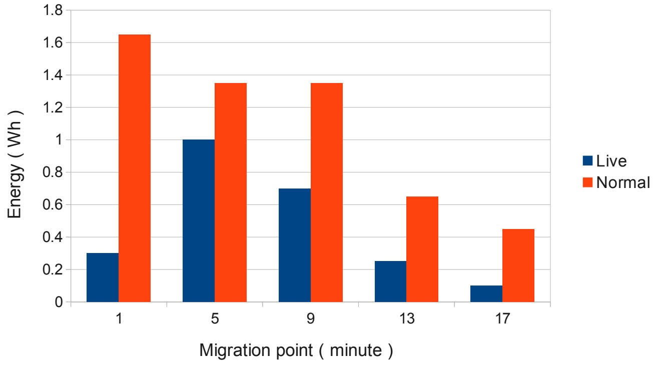

Figure 9. Energy overhead of one migration.

Another way of measuring performance is quality of service. Figure 6 illustrates better how the quality of service deteriorate when more parallelism is added. The cores of the energy-efficient server are more efficient. They are better both in efficiency as in energy efficiency. As the load increases the higher core count of the high efficient server provides better performance.

There are big differences in energy consumption of different hardware types. Figure 6 shows that the energy consumption of an idle single CPU energy-efficient server can be almost three time that of the high efficient two CPU server. Also, the power range of the energy-efficient server is narrower. The two processor server requires some load to become more energy-efficient. In our case more than eight parallel jobs.

Finally, we measured how big an impact one migration has on the CMS analysis task. Migration time was measured at different parts of the CMS analysis task. In Figure 8, we have the results of these migration measurements. As described earlier, the migration time depends much on the variation of load in the virtual machine. The duration of a normal migration rises steadily as the virtual machine expands its memory space. In all tests, the virtual machine runs one CMS job and then it shuts down. Virtual machine accumulates memory contents and as it has plenty of memory, it is never released.

Although the migration times of live migrations are longer, the energy consumption can still be lower. Figure 9 illustrates the difference of energy consumption overhead of live and normal migration. Although the difference between the two types is not significant, there are still benefits in choosing the correct one. The benefit of live migration is that it can continue the execution even though its been transferred from host to another. Thus, even though the migration lasts longer its energy overhead due to the continuous execution is less. On the other hand, if we double the load of the virtual machine, live migration will not succeed. Normal migration will succeed regardless of the load. We can also see that the effect of the background load was minimal.

VII. Conclusions and Future Work

We measured the energy consumption of high-energy physics analysis workload on two types of hardware. Results indicate that energy efficiency of a cluster can be improved by using heterogeneous hardware: it lowers the idle consumption, but does not reduce peak performance. Energy-efficient servers are good for lower load and storing, or parking, idle virtual machines, while high-efficient servers are needed to serve the peak loads and they can even be turned off when there is less load. Yet, high-efficient servers cannot be completely replaced by energy-efficient servers as the increased number of servers would consume more floor space. Migration is the key to make the management of virtual machines possible. Since the use of migration does not impose too much overhead, it is a suitable tool to manage virtual machines in large clusters and optimize the load between heterogeneous hardware.

In our tests, the choice of hardware was not optimal since the test workload was optimized for Xeon hardware. Having an up to date two processor Xeon server might have given better performance than our AMD processor. Moreover, the energy-efficient server proved to be almost too efficient for hosting idle virtual servers. In the case of the energy-efficient server, the ability to host the total memory space of the virtual machines is more important than the computing power. The memory requirement could even be optimized automatically by using the virtual machine memory size management that allows to over commit physical memory dynamically. One conclusion we made, is that when choosing hardware for a computing cluster, one could optimize both investment and running cost by purchasing task specific hardware, since there is a significant price difference between energy-efficient and high-efficiency hardware.

Our future work will focus on building an automatic system for optimizing the placement of virtual servers. Our aim is to use, e.g., CPU load values of the virtual server as an indicator whether do the migration or not in the following way: if the virtual server is running on the energy-efficient server and its short time load value goes over the threshold, it will be moved to the powerful server. Since migration is still a relatively expensive operation, it should not be performed too frequently. Thus, the virtual server is only moved back to the energy efficient server, if its long time average load value goes back under the given threshold.

REFERENCES

[1] B. Schäppi, F. Bellosa, B. Przywara, T. Bogner, S. Weeren, and A. Anglade, “Energy efficient servers in europe,” Austrian Energy Agency, Tech. Rep. October, 2007.

[2] E. STAR, “Report to congress on server and data center energy efficiency,” U.S. Environmental Protection Agency ENERGY STAR Program, Tech. Rep., 2007.

[3] L. A. Barroso and U. Hölzle, “The case for energy-proportional computing,” Computer, vol. 40, 2007, pp. 33–37.

[4] W. Vogels, “Beyond server consolidation,” Queue, vol. 6, January 2008, pp. 20–26.

[5] H. Chung-Hsing and S. Poole, “Revisiting server energy proportionality,” in Parallel Processing (ICPP), 2013 42nd International Conference on, Oct 2013, pp. 834–840.

[6] J. Koomey, “Growth in data center electricity use 2005 to 2010,” Oakland, CA: Analytics Press, 2011.

[7] F. Fabozzi, C. Jones, B. Hegner, and L. Lista, “Physics analysis tools for the cms experiment at lhc,” Nuclear Science, IEEE Transactions on, vol. 55, 2008, pp. 3539–3543.

[8] J. Kommeri, T. Niemi, and O. Helin, “Energy efficiency of server virtualization,” International Journal On Advances in Intelligent Systems, v 5 n 3&4, 2012.

[9] A. Verma, P. Ahuja, and A. Neogi, “Power-aware dynamic placement of hpc applications,” in Proceedings of the 22nd annual international conference on Supercomputing, ser. ICS ’08. New York, NY, USA: ACM, 2008, pp. 175–184.

[10] B. Li, J. Li, J. Huai, T. Wo, Q. Li, and L. Zhong, “Enacloud: An energy-saving application live placement approach for cloud computing environments,” in Cloud Computing, 2009. CLOUD ’09. IEEE International Conference on, 2009, pp. 17–24.

[11] A. Beloglazov and R. Buyya, “Energy efficient allocation of virtual machines in cloud data centers,” in Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on, 2010, pp. 577–578.

[12] T. Hirofuchi, H. Nakada, S. Itoh, and S. Sekiguchi, “Reactive cloud: Consolidating virtual machines with postcopy live migration,” Information and Media Technologies, vol. 7, no. 2, 2012, pp. 614–626.

[13] R. Nathuji and K. Schwan, “Virtualpower: coordinated power management in virtualized enterprise systems,” SIGOPS Oper. Syst. Rev., vol. 41, no. 6, Oct. 2007, pp. 265–278. [Online]. Available: http://doi.acm.org/10.1145/1323293.1294287

[14] H. Liu, C.-Z. Xu, H. Jin, J. Gong, and X. Liao, “Performance and energy modeling for live migration of virtual machines,” in Proceedings of the 20th International Symposium on High Performance Distributed Computing, ser. HPDC ’11. New York, NY, USA: ACM, 2011, pp. 171–182. [Online]. Available: http://doi.acm.org/10.1145/1996130.1996154

[15] C. Isci, J. Liu, B. Abali, J. O. Kephart, and J. Kouloheris, “Improving server utilization using fast virtual machine migration,” IBM J. Res. Dev., vol. 55, no. 6, Nov. 2011, pp. 365–376.

[16] S. Srikantaiah, A. Kansal, and F. Zhao, “Energy aware consolidation for cloud computing,” in Proceedings of the 2008 conference on Power aware computing and systems, ser. HotPower’08. Berkeley, CA, USA: USENIX Association, 2008, pp. 10–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1855610.1855620

[17] F. Hermenier, X. Lorca, J.-M. Menaud, G. Muller, and J. Lawall, “Entropy: a consolidation manager for clusters,” in Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments, ser. VEE ’09. New York, NY, USA: ACM, 2009, pp. 41–50. [Online]. Available: http://doi.acm.org/10.1145/1508293.1508300