The Journal of Supercomputing (2019) 75:5324–5346
https://doi.org/10.1007/s11227-019-02799-5

Multidimensional hierarchical VM migration management for HPC cloud environments

Sonja Filiposka[1] · **Anastas Mishev**[1] · **Katja Gilly**[2]

Published online: 11 March 2019

Abstract
Efficient resource management is crucial for balancing performance and energy consumption in large-scale data centres. In the case of additional requirements such as guaranteed resources and low communication latency, it is of great importance to implement not only an efficient initial placement algorithm, but also maximise consolidation by migration techniques, making sure that network performance is not sacrificed. In this paper, we introduce a hierarchical approach to migrations based on a combination of efficient packing algorithms and network communities. Results analysis shows the benefits of using a two-level approach where the combination of localised consolidation and network awareness improves both performance and energy efficiency, while maintaining low network hop distance.

Keywords Migration management · Energy efficiency · Consolidation · Network awareness · Performance

1 Introduction

The usage of today's high-performance computing (HPC) systems residing on large computing platforms such as data centres has shifted dramatically with the addition of a virtualised layer on top of the available hardware infrastructure. Providing additional elasticity and flexibility, the abstracted virtualised access to resources is based on the anytime from anywhere access paradigm, while taking advantage of a large

✉ Katja Gilly
katya@umh.es

Sonja Filiposka
sonja.filiposka@finki.ukim.mk

Anastas Mishev
anastas.mishev@finki.ukim.mk

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj No. 16, 1000 Skopje, Republic of Macedonia

[2] Department of Computers Engineering, Miguel Hernandez University, Elche, Alicante, Spain

pool of physical servers with many cores and high amount of memory. The trend of the computing demand resulted with the creation of very large data centres that can provide the storage and processing power needed to run big data problems. Locating these data centres closer to the customers enables an increased usage of computing resources for the purposes of many highly demanding applications [2]. These HPC systems, which use the virtualisation technology for provisioning of computational resources in the form of virtual machines (VMs), are usually referred to as an HPC cloud [1]. Built on top of virtualised data centres [2], HPC clouds have become a very popular adoption [3].

However, this concentration of computing power in the form of an HPC cloud creates some problems that need to be addressed successfully in order to guarantee optimal usage of resources in accordance with the changes in user demands [4]. Powering these systems is very costly, since consumption rapidly increases with the scale of the system. Thus, it is of high importance to employ efficient resource allocation and scheduling techniques that can provide not only adequate quality of service level for the customers, but will also take into account additional constraints such as minimum energy consumption [5].

Traditional resource management approaches in the virtual environment have always focused on the optimal resource scheduling that will guarantee high performance [6]. Thus, the majority of resource allocation techniques available today are trying to optimise the usage of available resources so as to provide the maximum performance using a minimum number of physical servers. There are different approaches starting from the simple first-fit heuristics, to more sophisticated solutions that take into account the dynamics of the system as a whole, the network topology of the data centre, the location of the storage relative to the physical machines (PMs), etc.

With the rapidly increasing number of large-scale data centres, the issue of energy efficiency became an additional major concern having a direct impact not only on the total cost of ownership, especially the costs for installing and maintaining an adequate cooling system for the data centre, but also an increased negative environmental influence due to the increasing carbon footprint of today's large data centres [7]. This interest in energy efficiency has given rise to efforts for developing resource management techniques that aim to optimise the power consumption in the data centre as a whole. However, in these cases, the decisions on VM placement are done solely on the basis of energy optimisation, disregarding any other optimisation as long as the placement conditions are such that they meet the basic requirements for resources. Aiming to consolidate the used resources and, thus, to decrease the power consumption, energy-aware techniques are mostly based on migrations [8], that is, moving the running VMs from one physical host to another host in order to minimise the number of used hosts. For HPC applications, only live migration techniques can be considered so that there is no down time of VMs during the process of migration. But even in the case of live migrations, there are penalties that need to be paid such as increased demand for bandwidth.

When approaching the problem with a holistic perspective, in order to provide HPC resources to users in an efficient manner, a reconciliation of approaches is needed. Resource management allocation and migration techniques employed must

be such that they aim to satisfy both high performance and low power consumption, building on both efficient consolidation and smart allocation that will still provide high performance for the set of tightly coupled tasks.

Towards this goal, in this paper we propose an hierarchical VM migration approach that combines the three management concerns:

- efficient usage of the available physical resources,
- network-aware placement and
- energy-wise saving based on consolidation.

The presented results show how a two-phased approach can successfully provide similar performance and energy efficiency compared to the typical migration techniques, but additionally can maintain a low latency between communicating VMs. This is achieved by enhancing migration algorithms with a network-aware approach when deciding where to place the migrated VM, instead of the typical first-fit implementations. In our proposal, two possibilities are investigated:

- strict placement in the current network community and
- flexible placement that hierarchically expands the destination community when necessary.

The rest of the paper is organised as follows. Section 2 provides an overview of the requirements for efficient resource management solutions that can be deployed in a HPC cloud, and the related work is detailed in Sect. 3. The popular single-objective and network-aware techniques are discussed. Section 4 proposes a hierarchical approach for VM migrations management in the HPC cloud that is based on a two-level combination of network awareness, localisation and consolidation. The implementation Sect. 5 describes how the proposed approach is implemented in the popular CloudSim simulator, followed by a presentation of the developed simulation scenarios. Section 6 provides a comparison of the effectiveness of employing hierarchical VM migration against typical approaches. The detailed analysis of the effects includes different aspects of the system: CPU utilisation, energy consumption, number of migrations, but also energy consumption of network devices and ports in use. The final section concludes the paper.

2 Traditional resource management in the HPC cloud

In the case of an HPC cloud, the HPC applications that are run by users can be represented as cloud services (jobs), i.e. a set of tightly coupled VMs with some desired characteristics. Each of these VMs runs one or multiple tasks that are part of the cloud service problem. Thus, VMs need to synchronise and exchange data with different frequency depending on the nature of the problem. Once defined by the user, the HPC application that is described in the form of a cloud service has to be placed on one or more PMs in order to provide it with the needed resources

in correspondence with the VM characteristics. The main goal when choosing the target PM is to minimise the number of used PMs.

From the theoretical point of view, this resource allocation problem can be represented as a multidimensional bin packing problem [9]. VMs and PMs (aka items and bins) can be described using vector bin packing where each vector dimension represents a different characteristic of the resource (e.g. CPU, RAM, storage). Finding the optimal solution to the packing problem is a well-known NP-hard problem, even when only one dimension is used. Adding more dimensions only makes things more difficult. Thus, there have been many heuristics that attempt to find solutions within acceptable running times [10].

When implementing a variation of the multidimensional vector bin packing, only one objective is solved: efficient usage of the physical resources available when placing new VMs. However, in the case of HPC applications, care must be taken when choosing the resource allocation method because it has to be based on space sharing, as opposed to time sharing whose allocation method allows under-provisioning of resources by advertising more available virtual resources compared to the real physical resources of the machines. Therefore, employing time-sharing CPU and memory is not desirable for HPC applications because this would imply that the resources provided to the user are not always available resources and, hence, not high performing. In contrast, the space-sharing scenario ensures that once some resources are allocated for a VM, these resources belong only to this VM and are thus available for use at any time while providing always a high-performance scenario.

2.1 Network-aware techniques: performance and power consumption implications

The interconnectivity between the chosen PMs that will host the VMs belonging to one HPC application is of major importance for HPC applications due to the nature of the problems that are being computed. Namely, the complete job represented as a cloud service is divided into a number of tasks executed in parallel, thus needing to regularly exchange information. Therefore, the solution of the resource allocation problem should choose some very well-interconnected PMs, meaning that the resource allocation technique employed must be network aware. By incorporating knowledge about the network architecture of the data centre, the choice of potential PMs for a given cloud service is constrained not only by the VM characteristics, but also by the network-wise closeness of the PMs. Effectively, employing a network-aware space-sharing resource allocation method provides a solution that encompasses two of the discussed objectives: efficient resource usage combined with low latency guaranteed by network-wise close placement.

There are a number of approaches that combine information about the network when deciding on the VM placement, where the optimisation is done in order to minimise the latency or maximise the bandwidth between the communicating VMs. Most of the techniques are based on a pair-wise network optimisation focusing on pairs of VMs [11]. Since many parallel processes constituting an HPC application communicate frequently, time spent in communication forms a significant fraction

of total execution time. The impact of cluster topology has been widely studied by HPC researchers. In the context of HPC cloud, it is necessary to use VM placement algorithms which map the multiple VMs of an HPC application in a topology-aware manner to minimise inter-VM communication overhead. One of the most intuitive approaches in this case (relating to the typical cluster approach in HPC) is the community-based VM placement [12] that adopts a hierarchical approach dividing the data centre into a hierarchy of communities based on the network topology [13] and uses this knowledge in order to place the complete HPC job (cloud service) inside the smallest viable physical community so that the interconnection between VMs benefits from the short hop distance, thus resulting in lower latency. This methodology enables enforcement of packing VMs to nodes in the same rack compared to a random placement policy, which can potentially distribute them all over the data centre. While these benefits of employing communities have been analysed in depth in [12], it has only been done for the initial placement of the VMs without considering the effects on power consumption.

Network-aware resource allocation techniques not only balance consolidation and high network performance, but also provide other potential means for achieving decreased power consumption in the data centre. It has been shown that the power consumption of the network elements in the data centre represents 10–20% of the total power consumption [14] leaving possibilities for improvement by smart placement. Additionally, when trying to model the power consumption of network devices, the analysis in [15] has shown that the power consumption is not related to the amount of traffic that passes through the network port, but it can be modelled using an 'off'/'sleep'/'on' port behaviour, where in case of no traffic on the port, its power consumption can drop down to 50% of the 'on' state. Similar conclusions have been drawn when considering a network device as a whole (a switch or a router), where again the switch in 'on' state consumes a fair amount of power independently of the amount of traffic that actually passes through it. Taking this into account, it is of the interest in this paper to observe how network-aware VM placement can help to consolidate the use of ports and switches in the data centre network so that energy saving can be achieved in the network elements as well.

2.2 Energy efficiency via migrations

When discussing energy efficiency of physical hosts, it is a well-known fact that the most power-hungry element in data centres is the CPU (cooling facilities excluded) [14] being responsible for at least 30% up to more than 80% of the total consumed power of a physical server. It is interesting to note that due to this fact the power consumption of a physical host rises dramatically when comparing the 'on'/'sleep'/'off' states. The difference between a 20% CPU usage and 100% CPU usage in terms of power consumption is very small in the cases when there is no hardware technique available for enforcing energy efficiency. Thus, the CPU power optimisation is the first and the foremost option for implementing energy efficiency. Significant reductions in the power consumption of physical servers can be achieved by employing hardware-based methods like dynamic frequency and voltage scaling (DVFS) [16].

The idea of DVFS is based on the fact that the energy dissipation of the CPU quadratically depends on the voltage. By reducing the working frequency of the CPU, the voltage reduction will lead to a decreased power consumption. However, the technique itself is limited due to the decreasing power consumption gap between the idle and the fully utilised state in today's CPU architectures [17]. It must be noted that DVFS as a power-aware technique has only local significance focusing on reducing the power consumption on an individual CPU based on the nature of the running tasks, and is thus not able to provide any global power consumption optimisation [18]. Even so, DVFS is the first step towards an energy-efficient data centre and today's large-scale computing facilities are built using physical servers with an incorporated DVFS behaviour.

On a global power-wise optimisation scale, the initial resource allocation approach does not include the notion of continuous consolidation, which is needed in a dynamic system wherein the availability of the physical resources changes over time as new VMs are being created, while old VMs that host finished tasks are destroyed and the resources they held are again available for new allocation. In order to answer the problem of continuously changing environment, migration techniques need to be additionally introduced so that consolidation of the used resources can be regularly attempted [19]. In the case of HPC applications, the problem of migrations must be approached with great care due to performance hindrance that might be introduced when migrating VMs. Relying only on live migration so that zero downtime of the VM is guaranteed, when deciding on migrating a VM in order to minimise the number of used hosts, care must be taken to ensure that the new physical host for the VM remains in the network vicinity of the rest of the VMs that belong to the same cloud service. In this way, the HPC application will not suffer any increase in network delay and will continue to run with the maximum available performance.

Coming back to the HPC-based efficient resource management in data centres, this mainstream approach is somewhat problematic since, in this case, the workload of the allocated VMs is not highly dynamic on the one hand, and, even more importantly, the notion of overutilised host in a time-sharing environment is unacceptable due to the lack of high performance guarantees of the allocated resources. In other words, in a space-sharing environment that fosters high-performance resource allocation, the notion of overutilised hosts is not applicable. Since the virtualised physical resources cannot be allocated to more than one VM, the physical host cannot be overutilised. On the contrary, it is desirable that the utilisation of the physical hosts is as close to 100% as possible so that the consolidation is maximised and the available resources are used in the most energy-efficient way possible (in conjunction with DVFS). Thus, in the space-shared environment of HPC, for the purposes of decreasing power consumption through server consolidation, only the underutilised hosts should be considered and this is part of the approach described in this paper. By employing threshold-based techniques, we can identify the underutilised hosts. The attempt to enforce consolidation in this case is to try and migrate all VMs residing on an underutilised host, so that the host can be shut down or put to sleep state in order to save on power consumption. All VMs from the underutilised host need to be migrated to other hosts that are already in use, otherwise the whole process may end up with increase in power consumption instead. If the algorithm employed

cannot find suitable new hosts for all VMs that reside on the chosen underutilised host, the process is cancelled.

3 Related work

Thorough survey of the energy efficiency of data centres as large-scale distributed systems is presented by Orgerie et al. in [20]. The authors present the energy efficiency of each of the building blocks of large distributed systems, starting from the elements of individual nodes, networking equipment, up to the software optimisation models. Virtualisation and cloud computing energy efficiency considerations are also elaborated in great depth.

There are several works describing the energy efficiency of the VM placement in the cloud data centres. Sotiriadis et al. in [21] consider the problem of scheduling general-purpose VMs in the cloud data centre, optimising the performance of the VMs and the energy efficiency of the cloud data centre. They offer quite comprehensive review of different approaches towards the cloud VM scheduling and energy efficiency.

The main body of research on the topic of VM migration techniques as a means for physical server consolidation revolves around the groundbreaking work in energy and performance efficiency of data centres as presented in [22]. The main idea of these approaches is based on dynamic VM migration in order to improve the power footprint of the data centre by consolidating the usage of both overutilised and underutilised hosts. Several different techniques for the identification of the hosts of interest to consolidation can be combined with different selection policies to select the most suitable VMs from those hosts and migrate them to a more efficient alternative. The concepts of VM migration employed in this case [22] are mainly based on the assumption of a highly dynamic workload on VMs in a time-sharing environment for the physical resources. Thus, the methods for identifying overutilised hosts are based on whether the upper limit of a host utilisation has been reached. This limit is defined as a threshold which can be a static absolute value, or dynamically derived and updated using local regression, interquartile range, or median absolute deviation (MAD), for an example. Kansal and Chana in [23] propose the usage of the firefly algorithm for performance and energy efficiency of the data centres through VM migration.

4 Hierarchical VM migrations approach

The main goal of this paper is to introduce a hierarchical VM migrations approach that will combine all the beneficial aspects discussed in the previous section: efficient packing, localisation, network awareness and power saving via consolidation. The discussion of the various approaches has pointed out the strong and weak points of each separate technique; however, the research on how different techniques can be combined in order to achieve complex multiobjective goals (such as high-performance computing with low network latency and considerable

energy savings) is still underway. Aiming to fill in the gap and analyse the possible trade-offs between performance and energy efficiency for HPC cloud environments, in this paper we present the implementation of a combined network-aware balancing approach for the initial placement and the consolidation attempts using migration.

In other words, we are considering an HPC cloud environment where the placement of jobs (cloud services) is done using a space-shared approach to ensure high-performance use of the available resources. The initial placement methodology is chosen to be community-based: the data centre is divided into hierarchical sets of physical nodes based on the interconnecting networking topology. Considering this division, the first step is a matchmaking process where a logical community (job consisting of a group of VMs) needs to be mapped to the smallest viable physical community (group of well-interconnected PMs). The second step is actual VM placement within the chosen physical community, where a load balancing or server consolidation approach [24] can be used in order to achieve the maximum performance, i.e. minimum number of servers used. Both the load balancing (LB) and the server consolidation (SC) methods are based on variable size bin packing vectorisation techniques with an objective to minimise the physical hosts used to place the VMs. The combination with the community matchmaking process ensures that the chosen hosts belong to the same community and are thus well interconnected providing high bandwidth and low latency.

To address the objective of energy efficiency, we propose a variation of the community placement approach that provides dynamic consolidation over time employing the technique of discovering underutilised hosts and attempting to migrate their VMs. The hierarchical approach avoids sacrificing network performance due to migration. To achieve this, every new destination host for every migrating VM must belong to the same community as the original host. In order to achieve higher energy efficiency, we consider that DVFS is used in conjunction with the resource management algorithms. The penalty for the live migration processes is induced by increased CPU usage due to migration (CPU is active on both the origin and destination of the migration) and by using up all of the available bandwidth on the origin and destination host for the transfer of the VM processes which puts additional stress to the network [25].

Let us put together the main features of the hierarchical approach that is proposed in this paper:

- space sharing of resources
- network-aware community-based hierarchical approach for dividing the data centre into physical communities
- load balancing or server consolidation VM placement algorithm that works locally on a given community of PMs
- detection of underutilised hosts using MAD around a specified threshold according to the current workload
- migration of all VMs from underutilised hosts limited to other, already active, hosts located in the same community

It is of interest to note that the chosen approach benefits the resilience characteristics of the data centre's ability to host HPC applications. Namely, by enforcing network awareness throughout the complete HPC job lifetime, all of its components running in separate VMs are going to be hosted on PMs with the minimum hop distance available. This means that a failure of the switches that connect other data centre parts will have no effect on the running HPC-based cloud service. Only in the case of a targeted failure in one of a very small number of switches, an HPC application error can occur. On the other hand, by employing the communities hierarchy, a relatively small number of cloud services will need interconnection through the edge-level and aggregate-level switches when compared to a traditional random best fit scenario. In this way, there are very few long-distance communications and, thus, the data centre is able to function with very high efficiency even in the case of failure of the root-level switches.

5 Implementation

For the purposes of the implementation and effectiveness analysis of the proposed hierarchical VM migrations approach, we have chosen a very popular tool for simulation of cloud data centres that offers virtualised infrastructure as a service, which is the basis for the HPC cloud. This tool, which has already been shown to be adaptable for HPC applications by being extensible and configurable in a parametrised fashion [26], is the CloudSim simulator. One of the main reasons for the popularity of CloudSim is the ability to model the power consumption of the PMs by using a *power network data centre* extension of the regular *data centre* model in the simulator. CloudSim also supports creating simulations that will analyse the usage of the data centre network using a separate *network data centre* branch that also supports entities like *switches*, *ports* and *links*.

Since in this case, the goal is to observe the performance and energy efficiency obtained on the basis of network topology awareness, for the implementation of the community-based placement and hierarchical VM migration techniques, a new branch of a *network power data centre* has been created so that both features are captured. In this way, the data centre model allows defining a power consumption model for PMs and network devices, in addition to the definition of the network topology and availability of link bandwidth.

Following the most typical network topology for today's data centres, the *network power data centre* model has been extended to support the creation of a fat tree network topology [27] with two or three layers including edge, aggregate and root switches with a given number of ports. The results presented in this paper are based on two large-scale data centres consisting of 5400 or 10,800 physical hosts interconnected in a three-layer fat tree topology with 36-port switches.

Upon defining the network topology and PM characteristics using the network power data centre model, the VM placement and running is handled by data centre brokers. In the case of HPC cloud, as it is already discussed, the job to be executed is defined as a cloud service which consists of a number of VMs with specific desired characteristics on which the processes that are executed in parallel are started. Some

of the processes on different VMs may finish earlier than others depending on the nature of the job. The job is considered finished when the whole cloud service is finished, i.e. all VMs that belong to the cloud service have finished and are destroyed. This behaviour can be modelled in CloudSim by instantiating one data centre broker for each cloud service. Every data centre broker is tasked to make sure that the VMs that belong to the given job are placed on a PM and that the processes are started. It then monitors the execution of processes and cleans up VMs with finished processes. Once all processes have finished, the broker will complete the job. The processes that run on VMs are referred to as cloudlets in CloudSim, so in the HPC cloud scenario, each VM has one cloudlet assigned to it and it consumes all of the available resources of the provided VM. In Table 1, we detail some of CloudSim already implemented classes and the new network-aware community-based classes we have coded to illustrate previous description.

In order to analyse how different amounts of the overall workload in the data centre impact the placement and migration events, a number of simulation scenarios have been created wherein the number of instantiated cloud services is varied from 1000 to 2000 with a different maximum number of VMs per cloud service: 10 or 20. The actual number of VMs is chosen randomly following a uniform distribution that aims to create a very diversified demanding workload on the data centre which is more difficult to pack compared to a Gaussian-oriented workload generation.

PMs and VMs can have multiple characteristics that define the number of dimensions needed for the vector representation of the variable bin packing problem. CloudSim supports features such as number of CPU cores, amount of RAM, bandwidth, operating system (OS) family and storage size. For the purposes of the analysis presented here, the first three characteristics have been used while assuming that the complete data centre supports one OS family (which is common for HPC applications) and that there is enough storage space for all running jobs (in this case, the storage can be represented as a shared central storage which is also common in practical implementations). The shared storage space indicates that during the process of live migration, only the running memory of VM needs to be migrated.

The realistic power model for physical hosts implemented in the simulator, namely the IBM x3550 Xeon X5675-based host with 6 cores, 12 GB RAM with

Table 1 CloudSim original and new network-aware community-based classes

Original classes	New classes
Cloudlet	Network_PowerDatacenter
PowerDatacenter	Network_PowerHost
DatacenterBroker	Network_PowerVmMigrationPolicyCommunityBased
PowerHost	OptimVMAllocationPolicy_CommunityBased*
PowerVm	
PowerVmAllocationPolicyMigration*	Network_PowerVmAllocationPolicyMigration*
PowerVmAllocationPolicy	Network_PowerVmAllocationPolicy*
PowerVmSelectionPolicy	Network_PowerVmSelectionPolicy

inherent implementation of DVFS and 1 GB available bandwidth, has been used to represent the data centre infrastructure. The sizes of the VMs that are generated as part of the cloud services vary from 1 or 2 cores and 2 or 4 GB RAM, each consuming 100 MB bandwidth and all available MIPS of the host. The cloudlet length has been defined to be randomly chosen from the 1 to 5 GB interval with a step of 1 GB. The running time of each process in a VM is defined as the cloudlet length divided by the allocated MIPS.

The space-sharing resource allocation technique is natively supported by the simulator which, out of the box, provides only the basic first-fit packing heuristics for the initial VM placement. Thus, the simulator has been extended to support the full community-based VM placement framework as described in [12] with the ability to support both the load balancing (LB) and the server consolidation (SC) techniques for VM to PM placement and the choice of matching logical communities (cloud services) to physical communities (sets of PMs). In this way, four different combinations are available depending on the choice for the higher- and lower-level placement, namely LB-LB, LB-SC, SC-LB and SC-SC, in addition to the native first-fit which is used for comparison. While logical communities are defined for one cloud service, physical communities are described as a hierarchical dendrogram, where: the smallest community is each individual physical host, the next level is represented with communities comprised of hosts connected to the same edge switch, the upper next level are communities where hosts share aggregate switch, and the root of the dendrogram is one community encompassing all hosts in the data centre.

Once the simulation is started and the brokers have distributed their VMs and cloudlets according to the decisions for the initial VM placement, a regular check is done for possible consolidation via migration. There are several implementations of different migration policies available in CloudSim, all based on the overutilised and underutilised hosts discussed in Sect. 2. When defining scenarios based on these available implementations, the simulator automatically combines them with the first-fit algorithm for initial placement. Therefore, the code has been changed so that the VM placement policy is separated from the VM migration policy in order to allow the scenario creator to freely choose between any available VM placement algorithm and VM migration algorithm. This enables the already implemented migration policies to be combined with the community-based placement approaches. The original migration implementations are such that no matter the chosen VM migration and selection policies, once the decision has been made on which VM to migrate, the destination host is chosen using exclusively the first-fit approach.

In order to implement our hierarchical VM migration approach, a new VM migration policy has been defined that provides choices for the techniques used to determine the potential VMs to be migrated and the way the destination of the migration will be decided. This extension has allowed the creation of simulation scenarios wherein upon deciding which VM is going to be migrated, the destination host can be chosen based on the rules defined according to the community-based principles. Two approaches are defined in this case:

– *strict rules*—the chosen destination host must not only have available resources but also belong to the same community together with the rest of the PMs that host VMs from the same cloud service;
– *flexible rules*—if no suitable host is found in the same community, the algorithm widens the search to hosts that belong to the larger hierarchical community.

Having in mind the HPC application nature with a lack of highly dynamic cloudlets and overutilised hosts, all available migration policies actually fall back to the median average deviation (MAD) around a static threshold of 30% total CPU utilisation. Thus, this policy has been used for comparison purposes.

Another enhancement that has been implemented in CloudSim is the migration delay. In the original version, the migration delay is introduced as the time needed to transfer the working memory of the VM using the full available bandwidth of the host. In our implementation, the conditions are made more realistic taking into account the fact that after an underutilised host has been identified, the available bandwidth for migration of each of its VMs is calculated as the minimum of the available bandwidth of the new destination host and the original source host divided by the number of migrating VMs.

The network power consumption is modelled using the available information provided in [15], according to which there is a fixed amount of power consumed by the switch chassis plus additional power consumed represented as 1% of the chassis power by each port that is currently active (i.e. it has traffic flow). From a network energy efficiency perspective, it is assumed that root switches are always on and fully powered up. Also, there are two edge switch ports active per each active host (down and up link) and an edge switch is off only if all ports are off. Similar reasoning is implemented for the aggregation layer with two ports per each edge switch.

6 Results and discussion

The aim of this section is to discuss the performance of the proposed hierarchical VM migrations approach by comparing it with other already implemented approaches in CloudSim, combining it with different types of initial placement, and analysing the relationship between the energy consumption savings in the data centre and its effect on the performance of running HPC applications. For these purposes, the energy consumption in the data centre has been measured for a number of scenarios where the different discussed techniques for initial VM placement and VM migrations have been defined. The number of hosts and the mean host utilisation are analysed in order to visualise the packing performance. Also, the number of migrations and their influence on the changes in the hop distances between VMs that belong to the same cloud service are discussed. The last subsection is devoted to the energy consumption by the network data centre and how it is related to the hosts power consumption and the overall performance of the system.

For convenience and consistency, all figure legends are based on the following coding:

```
{initial-placement . migration-technique}
```

There are five options in total for initial placement: `first-fit` and four combinations for the community-based approach based on the algorithm used on the community level and the one used on the host level (`lb-lb, lb-sc, sc-lb, sc-sc`). Whenever the difference in results obtained for the four options is negligible, in order to increase the readability of the figure only `lb-lb` results are presented. Six options for migrations are compared:

- `dvfs`—no migration policy is in effect, only DVFS is used on the hosts.
- `mad`—the original CloudSim median average deviation (MAD) technique for underutilised hosts discovery with first-fit for new destination hosts selection.
- `lb-mad`—hierarchical VM migration based on the original CloudSim median average deviation technique for underutilised hosts discovery, strict rules for community choosing and load balancing algorithm for new destination hosts selection.
- `sc-mad`—hierarchical VM migration based on the original CloudSim MAD technique for underutilised hosts discovery, strict rules for community choosing and server consolidation algorithm for new destination hosts selection.
- `flb-mad`—hierarchical VM migration based on the original CloudSim MAD technique for underutilised hosts discovery, flexible rules for community choosing and load balancing algorithm for new destination hosts selection.
- `fsc-mad`—hierarchical VM migration based on the original CloudSim MAD technique for underutilised hosts discovery, flexible rules for community choosing and server consolidation algorithm for new destination hosts selection.

6.1 Energy efficiency

We represent in Fig. 1 the energy consumed by the hosts during the simulation time for one example environment. When using techniques for initial placement that are not combined with migration techniques (first two items in the legend) and just DVFS is used as means for continuous adaptive energy saving, it is evident that the overall energy consumption in the hosts of the data centre is much higher as there are no consolidation attempts. In the case of combining network-aware placement techniques with migration techniques, a noticeable drop in energy consumption can be achieved. The reason is because when cloudlet processes are started after the initial placement, as different cloudlets (according to their length) finish earlier than others during the simulation time, the energy consumption starts to drop as expected. Therefore, migrations are taking place and consolidation of servers is reflected in a reduction in the energy consumed. This is most evident in the time frame of 400–500 s when the shortest cloudlets have already finished and the first set of underutilised hosts has been detected. These triggered migrations manage to produce considerable energy savings. The more subtle differences between the different migration possibilities are visible after 800 s with the non-network-aware migration (MAD) enabling the best energy savings, while the strict and flexible versions of hierarchical network-aware migrations being only slightly worse. This small increase in power consumption is due

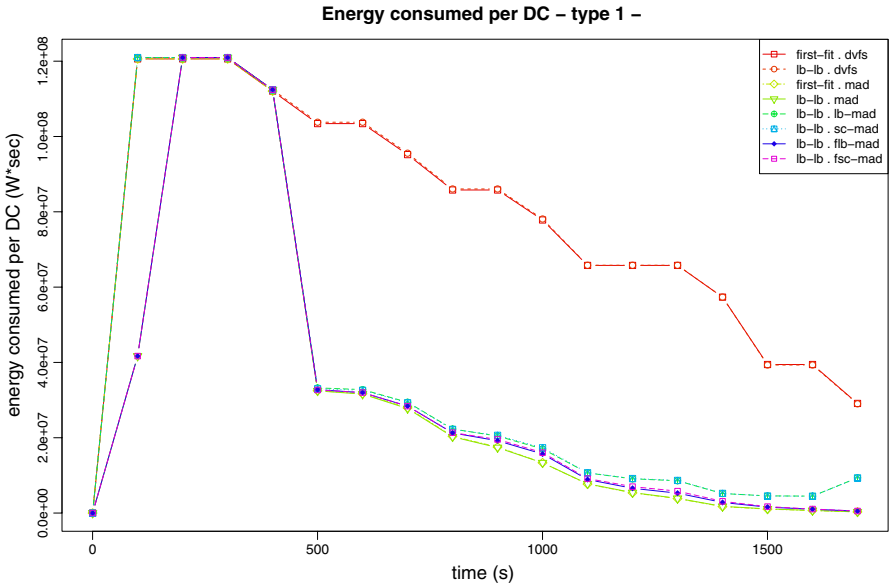


Fig. 1 Evolution in time of host energy consumed in the data centre for an example data centre with 10,800 hosts and a workload of 2000 cloud services each with max 20 VMs

to the more strict policies in choosing the physical hosts for migrating VMs. The network awareness that keeps the cloud service VMs close together results in less efficient packing per host as it is presented in the results in the next section.

The total energy consumed by data centre's hosts is presented in Fig. 2 together with the total number of migrations (right-hand y axis) averaged over different simulations that have been run using different workloads in terms of number of cloud services and maximum number of VMs per service. Results show that the energy-wise price to be paid in the case of using strict community-based migrations opposed to the flexible version or the non-network-aware MAD-based migration policy is around 10–13% increase in total energy consumption. The use of load balancing or server consolidation approaches in the initial placement does not seem to have any significant impact on the energy consumption or the number of VM migrations. Yet, it must be noted that the number of VM migrations in the case of strict network-aware migrations, that is, when the destination host for migration must belong to the same community with the rest of the cloud service, is more than 50% lower, which has a twofold benefit: (1) it manages to achieve considerable consolidation with a small energy sacrifice, and (2) the lower number of migrations induces a more stable environment for guaranteed performance while creating far less additional network traffic. In the cases when the sacrifice in energy is not acceptable, flexible rules for hierarchical VM migrations can be used and results show that this network-aware approach incurs just a slight increase in energy consumption (less than 4%) compared to the random first-fit algorithm.

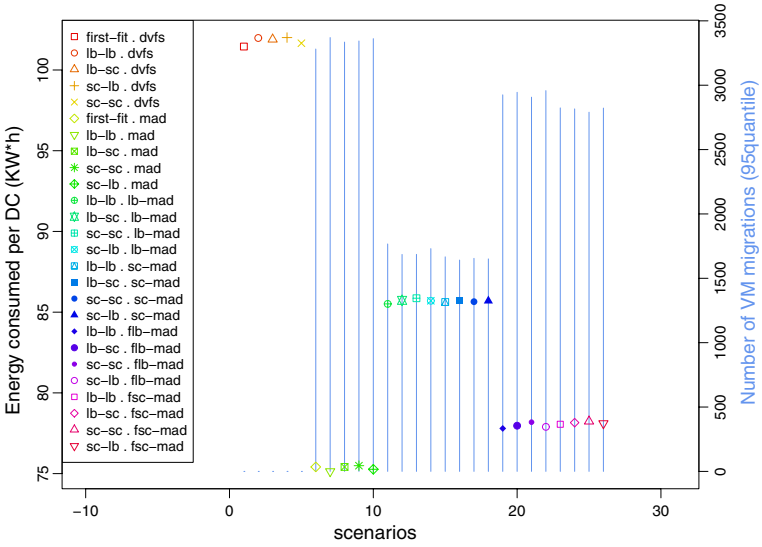
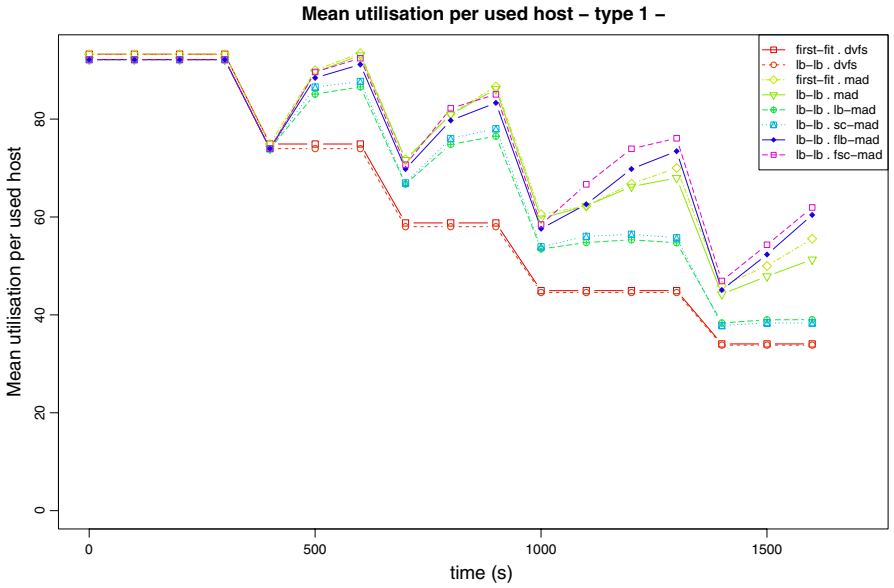


Fig. 2 Energy consumed by hosts (points) and the number of migrations (lines) in the data centre for an example data centre with 5400 hosts with varied workloads

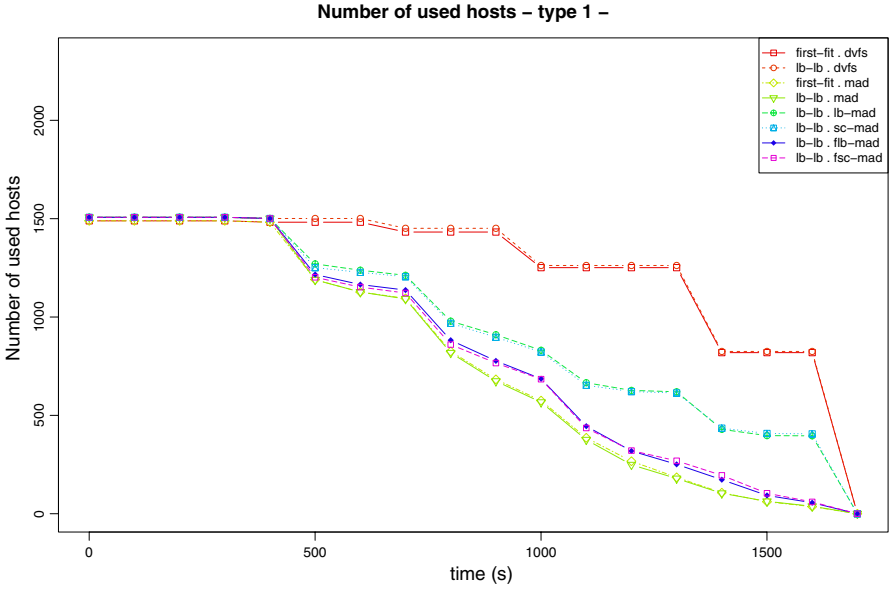
6.2 Host utilisation

The host power consumption is the main component of the overall data centre energy consumption, and thus, it is of great interest that initial VM packing algorithms together with migration policies manage to produce as efficient packing as possible in terms of obtaining the maximum utilisation of active hosts and the minimum number of used hosts. These metrics and their change over the simulation time are presented in Fig. 3. All considered initial placement algorithms show comparable efficiency providing close to 90% utilisation of the active hosts (about 1500 in total) at the beginning of the simulation. As part of the jobs are finished, the hosts utilisation starts to drop since the initial placement is no longer very efficient. In the case of the scenarios that do not use migration techniques, this leads to big drops in host utilisation, with a much smaller decrease in the number of hosts used, so that in the last time interval of the simulation around 900 hosts are used (Fig. 3b) with a mean utilisation of only 40% (Fig. 3a). Employing hierarchical VM migration techniques, the situation can be very much improved, so that the consolidation of the VMs induces larger drops in the number of used hosts, and a compensation in the utilisation of the used hosts which is achieved over longer periods of time (200–300 s).

The MAD-based migration policies are able to effectively consolidate the number of used hosts over time, whereas the network-aware migration policies based on flexible rules are just slightly worse in their consolidation efforts. However, as the simulation time progresses it is increasingly difficult for the strict rules version of the community-aware migrations to follow the pattern of the previous two, leading to about 45% mean host utilisation compared to the others' 60%. Thus, the



(a) Mean utilisation per used host.



(b) Number of used hosts.

Fig. 3 Host utilisation in the data centre for an example data centre with 10,800 hosts and a workload of 1000 cloud services each with max 10 VMs

performance energy trade-offs paid when employing the strict migration policies is accompanied with a lower host utilisation over time. However, in a more realistic scenario, there would be new cloud services placed over time so this should not impose a big problem.

6.3 Migrations

Figure 4 depicts the number of started VM migrations over the simulation time for an example scenario. The four peaks in the figure are occurring in the time frames when a number of VMs are finished with processing and are being destroyed (as defined with the cloudlet length), and consequently, they correspond to the drops in hosts utilisation as shown in Fig. 3a. This change of the current status of the hosts is a trigger detected via the identification of underutilised hosts, which results in an increasing number of decisions for migrations. The MAD migration policy has a very fast reactive approach triggering migrations as early as possible during the simulation even if this will result with a severe negative impact on the network distance between the VMs that belong to the same cloud service. The hierarchical VM migrations-based approaches (strict and flexible) are not able to find viable destinations for all potential migrating VMs, and thus, they postpone some of the migrations to occur later, leading to a larger number of migrations in 1000 and 1400 s when a larger number of potential destination hosts are available.

More details on the effects of the migrations can be concluded when comparing Fig. 5a, b, which present the mean hop distance between the source and destination hosts during migrations and the number of migrations for each simulation scenario.

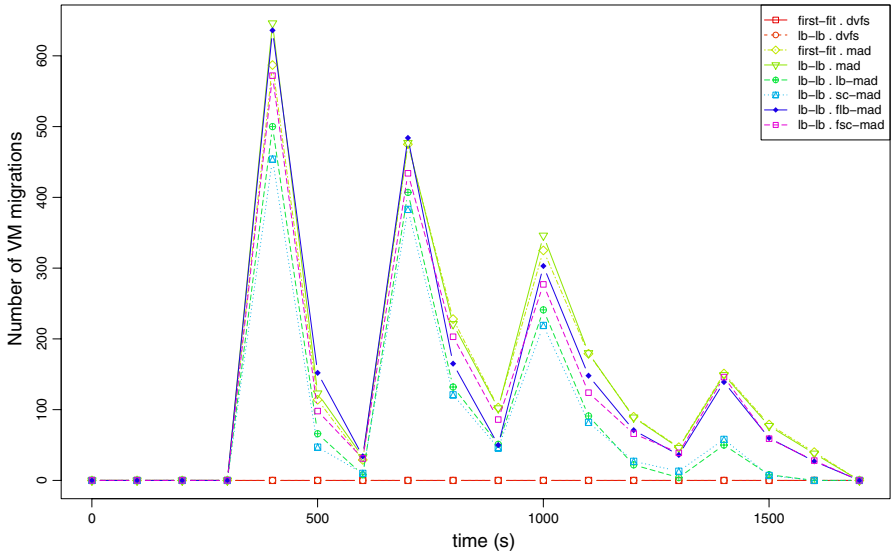
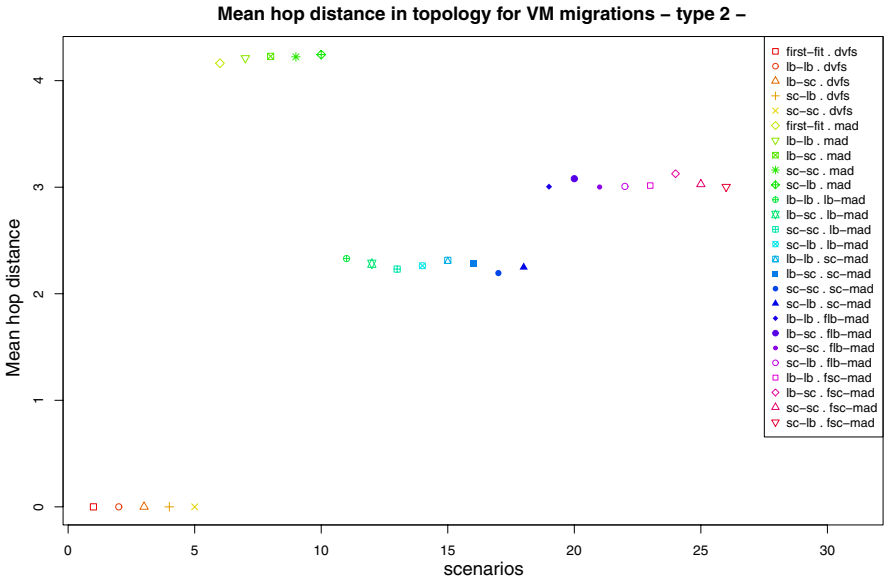
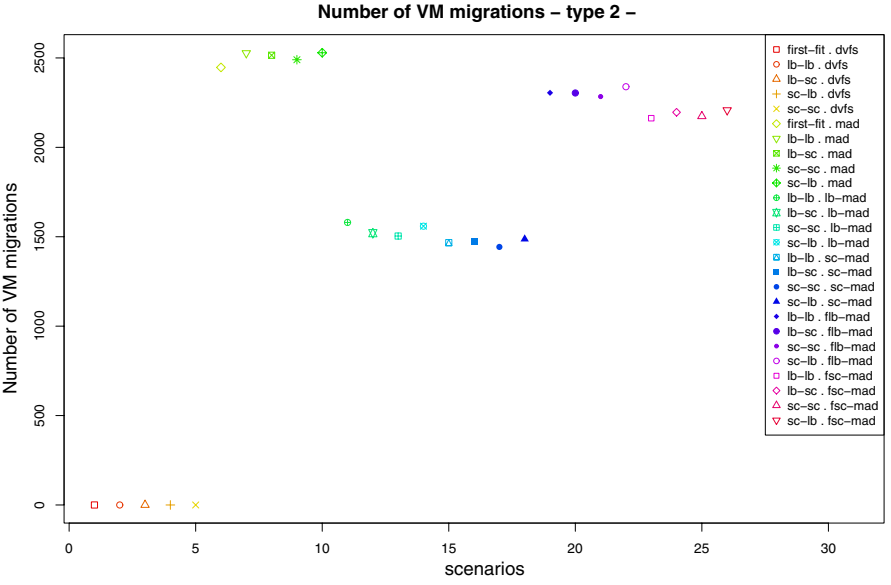


Fig. 4 Comparison of the evolution of a selected number of migrations during the simulation time for an example data centre with 5400 hosts and a workload of 1000 cloud services each with max 10 VMs



(a) Mean hop distance between source and destination hosts when migrating VMs.



(b) Number of migrations.

Fig. 5 Migrations and changes in mean hop distance for an example data centre with 5400 hosts and a workload of 1000 cloud services each with max 10 VMs

The scenarios that do not support migrations are presented with 0 in both figures, while again there is no major discernible effect on the choice of either load balancing or server consolidation approach when using community-based techniques for initial placement. The choice of following strict or flexible rules for hierarchical VM migrations is a different topic though.

First of all, the trade-off of using MAD as non-network-aware migration policy, which incurs the maximum number of VM migrations and thus results with maximum energy savings, is presented also with the maximum average hop distance (that is 4.2 hops). In the case of a fat tree architecture, this hop distance means that jobs allocated on hosts are communicating via all three layers of switches (edge, aggregate and root). On the other hand, in the network-aware hierarchical VM migrations case, the communication is localised up to the aggregate (3 hops) or up to the edge (2 hops) level for flexible and strict rules, respectively. In particular, Fig. 5a is the main highlight of the performance energy trade-offs, since the hop distance is directly related to the communication delay between the parallel processes in the HPC application.

6.4 Network energy consumption analysis

The results presented in this subsection are focusing on the energy consumption in the data centre network and how it is affected by the choice of resource management. All chosen example results provided in figures are for high workloads in order to clearly show the relatively small differences in effectiveness of the compared combinations, which are diminished for smaller workloads in terms of the total number of cloud services or the total number of VMs. As previously discussed, the results presented are based on the assumption that all active switches consume a given fixed amount of power for the chassis, while every active port adds to this amount an additional 1% [15]. Thus, the total relative power consumed by the data centre's network is calculated as the sum of the active switches plus 1% of the total number of active ports. To get the real energy consumed in Ws, one must multiply the presented numbers by the power of the switch chassis (50 W for an example).

The main obvious conclusion that can be drawn from Fig. 6 is that, compared to a basic first-fit initial placement, when using a network-aware placement approach (any type of the community-based versions) the energy consumption in the network is lower mainly due to the fact that a smaller number of switches need to be activated because of good consolidation of the cloud services on the target hosts. Figure 7a, b provides a detailed breakdown of the total energy consumption. The communication between the VMs is mostly local and, as the processes are finishing, the number of active ports decreases causing additional savings in power, especially in the cases when the complete switch can also be turned off. Another observation that can be made from this figure concerns the best performing group of scenarios based on network-aware placement combined with traditional MAD migrations. The energy consumption is the lowest in this case due to the best packing provided which results in the minimal number of active switches and active ports. However, the price to be paid in this case is the relative

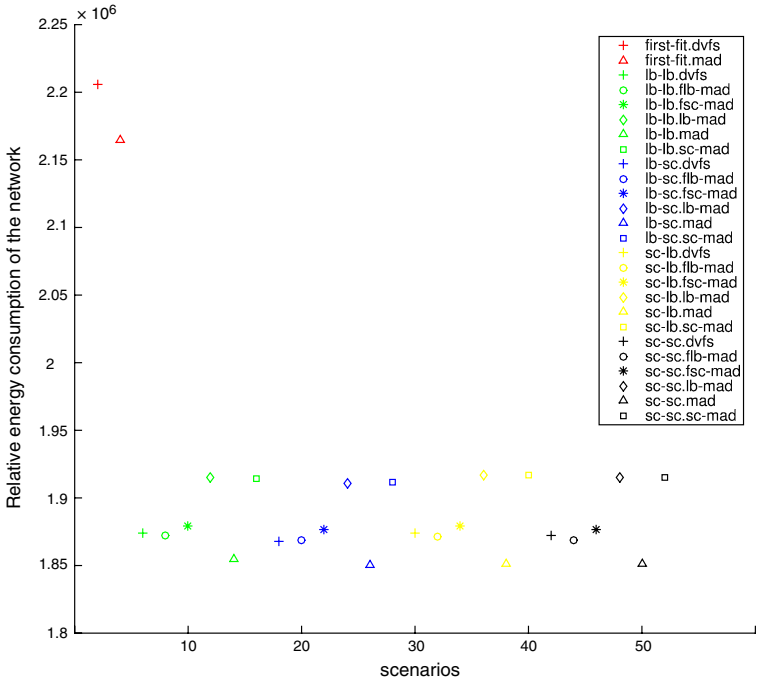
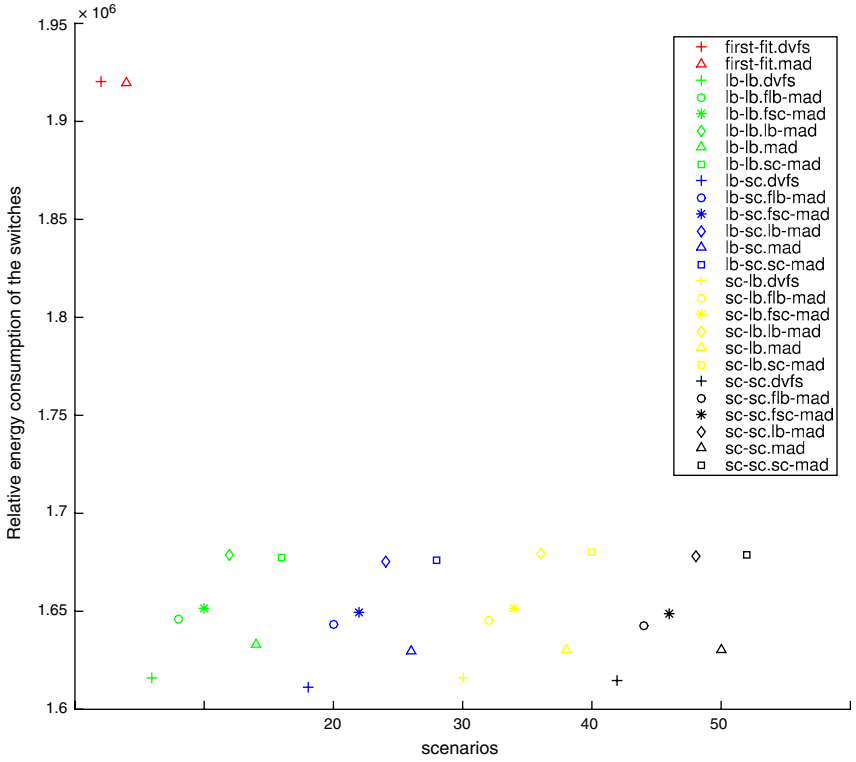


Fig. 6 Relative total energy consumption in the data centre network for an example data centre with 10,800 hosts and a workload of 2000 cloud services each with max 20 VMs

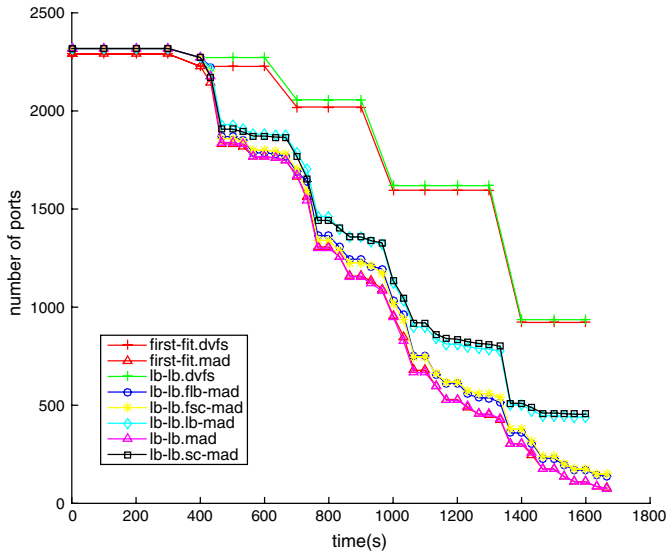
increase in hop distance between the VMs of the same cloud service, which will negatively impact the performance. It is clear that paying this price is not reasonable since the energy-wise performance of the scenarios where the migrations are done using the strict or flexible hierarchical VM migration approaches is very much comparable.

While the total network energy consumption is mostly dependent on the employment of a network-aware VM placement algorithm, it is important to emphasise that when using migrations to consolidate the usage of hosts, the number of ports is also consolidated. However, the price to be paid is putting more strain on the aggregate- and root-level switches. The main problem that arises is that additional switches sometimes need to be turned on in order to complete the process of migrations. This consequently reflects in the overall increase in network energy consumption taking into account that the main consumption in the network is made by the switch chassis.

It must not be forgotten, however, that the amount of power consumption that incurs in the network is by far smaller compared to the host power consumption (only 10–20%), so the overall power consumption in the data centre will be lowered due to consolidation via migrations. In other words, in order to improve the energy efficiency of the overall data centre using migrations, the employed migration techniques will somewhat increase the network energy consumption due to their attempt to consolidate the number of hosts used. In addition to this, it must be noted that the



(a) Relative energy consumption made by switches chassis.



(b) Number of active ports.

Fig. 7 Network energy consumption breakdown for an example data centre with 10,800 hosts and a workload of 2000 cloud services each with max 20 VMs

performance from a network point of view will be also lowered because of the extra traffic that will use up the bandwidth in the network during the migration processes.

7 Conclusions

In order to maximise the usage of available resources in large-scale data centres that support the HPC cloud, it is of utter importance to choose a very good resource management suite of algorithms that will provide high performance with good energy efficiency. Our proposal includes efficient packing for high resource utilisation, network awareness for low latency communication, and good consolidation via migrations, considering hierarchical VM migrations as an extension to a community-based placement approach. We analysed the effects on these three goals by comparing scenarios with and without network awareness or hierarchical migrations in large-scale data centre simulation scenarios. Our results show that the initial placement is extremely important for high performance of the system, but the continuous energy efficiency can only be achieved with regular consolidation via migrations. However, migrations should be avoided if the tightly coupled communicating VMs that compose a cloud service need to go through an increased number of hops along the network architecture, having then a negative impact on the communication delay. It has been shown that by following the proposed hierarchical VM migrations approach, very good consolidation and energy efficiency can be achieved while maintaining high communication performance. Network-aware consolidation efforts are efficient solutions to improve the performance of the network while reducing the carbon footprint of the HPC data centre.

References

1. Garg SK, Yeo CS, Anandasivam A, Buyya R (2009) Energy-efficient scheduling of HPC applications in cloud computing environments. CoRR arXiv:0909.1146
2. Sotomayor B (2010) Provisioning computational resources using virtual machines and leases. Ph.D. thesis, University of Chicago
3. Mauch V, Kunze M, Hillenbrand M (2013) High performance cloud computing. Future Gener Comput Syst 29(6):1408–1416
4. Zakarya M, Lee G (2017) Energy efficient computing, clusters, grids and clouds: a taxonomy and survey. Sustain Comput Inform Syst 14:13–33
5. Chaabouni T, Khemakhem M (2018) Energy management strategy in cloud computing: a perspective study. J Supercomput 74(12):6569–6597
6. Vinothina V, Sridaran R (2012) A survey on resource allocation strategies in cloud computing. Int J Adv Comput Sci Appl 1(3):97–104
7. Thakur S, Chaurasia A (2016) Towards green cloud computing: impact of carbon footprint on environment. In: 2016 6th International Conference Cloud System and Big Data Engineering (Confluence). IEEE, pp 209–213
8. Beloglazov A, Jemal A, Rajkumar B (2012) Energy-aware resource allocation heuristics for efficient management of data centres for cloud computing. Future Gener Comput Syst 28(5):755–768
9. Chekuri C, Sanjeev K (1999) On multi-dimensional packing problems. In: Proceedings of the 1999 10th annual ACM-SIAM symposium on discrete algorithms. Baltimore, MD, USA, pp 185–194
10. Panigrahy R, Panigrahy R, Talwar K, Uyeda L, Wieder U (2011) Heuristics for vector bin packing. research.microsoft.com

11. Hamdi K, Kefi M (2016) Network-aware virtual machine placement in cloud data centers: an overview. In: 2016 International Conference on Industrial Informatics and Computer Systems (CIICS), Sharjah, pp 1–6
12. Filiposka S, Mishev A, Juiz C (2015) Community-based VM placement framework. J Supercomput 71(12):4504–4528
13. Filiposka S, Juiz C (2015) Community-based complex cloud data center. Physica A 419:356–372
14. Dayarathna M, Wen Y, Fan R (2016) Data center energy consumption modeling: a survey. IEEE Commun Surv Tutor 18(1):732–794
15. Mahadevan P, Banerjee S, Sharma P (2010) Energy proportionality of an enterprise network. In: Proceedings of the First ACM SIGCOMM Workshop on Green Networking, pp 53–60
16. Choi K, Ramakrishna S, Massoud P (2005) Fine-grained dynamic voltage and frequency scaling for precise energy and performance tradeoff based on the ratio of off-chip access to on-chip computation times. IEEE Trans Comput Aided Des Integr Circuits Syst 24(1):18–28
17. Le Sueur E, Heiser G (2010) Dynamic voltage and frequency scaling: The laws of diminishing returns. In: Proceedings of the 2010 International Conference on Power Aware Computing and Systems
18. Von Laszewski G, Wang L, Younge AJ, He X (2009) Power-aware scheduling of virtual machines in DVFS-enabled clusters. In: IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE
19. Carli T, Henriot S, Cohen J, Tomasik J (2016) A packing problem approach to energy-aware load distribution in Clouds. Sustain Comput Inform Syst 9:20–32
20. Orgerie AC, Assuncao MDD, Lefevre L (2014) A survey on techniques for improving the energy efficiency of large-scale distributed systems. ACM Comput Surv (CSUR) 46(4):47
21. Sotiriadis S, Bessis N, Buyya R (2018) Self managed virtual machine scheduling in cloud systems. Inf Sci 433:381–400
22. Beloglazov A, Rajkumar B (2012) Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centres. Concurr Comput Pract Exp 24(13):1397–1420
23. Kansal NJ, Chana I (2016) Energy-aware virtual machine migration for cloud computing—a firefly optimization approach. J Grid Comput 14(2):327–345
24. Mishra M, Sahoo A (2011) On theory of VM placement: anomalies in existing methodologies and their mitigation using a novel vector based approach. In: 2011 IEEE 4th International Conference on Cloud Computing, USA, pp 275–282
25. Liu H, Xu CZ, Huazhong HJ, Gong J, Liao X (2013) Performance and energy modeling for live migration of virtual machines. Cluster Comput 16(2):249–264
26. Gupta A, Kalé LV, Milojicic D, Faraboschi P, Balle SM (2013) HPC-aware VM placement in infrastructure clouds. In: 2013 IEEE International Conference on Cloud Engineering (IC2E), Redwood City, CA, pp 11–20
27. Prisacari B, Rodriguez G, Minkenberg C, Hoefler T (2013) Bandwidth-optimal all-to-all exchanges in fat tree networks. In: Proceedings of the 27th International ACM Conference on International Conference on Supercomputing, pp 139–148