ANALYSIS OF ATTAINABLE ENERGY CONSUMPTION REDUCTION IN ICT BY USING DATA CENTER COMPREHENSIVE LOAD MANAGEMENT (Extended Abstract)

Daniel Schlitt, Marko Hoyer,
OFFIS – Institute for Information Technology,
{schlitt,hoyer}@offis.de

Kiril Schröder, Wolfgang Nebel
C.v.O. University of Oldenburg,
{schroeder,nebel}@informatik.uni-oldenburg.de

ABSTRACT

Many technologies like thin clients or cloud computing possess a growing interest, which currently results in a movement towards increasing usage of growing data centers and an accompanying rising energy demand. To ensure an economical and ecological operation of data centers nevertheless, action to cut down on energy have to be taken by the operators. A good possibility for this purpose is the virtualization of services followed by a dynamic allocation to the servers. Thus unneeded server capacities can be shut down to save energy. In a survey we analyzed possible application areas as well as present challenges of load management both inside a data center and data center comprehensive. Furthermore we theoretically evaluated the energy saving potential: Inside a data center load management could reduce the energy consumption by 20% to over 40% and if it is extended to a network of data centers costs of 10% and more could be saved additionally. In another field of application the data center comprehensive load management is used to flatten the energy demand of the data center's neighborhood such that the energy supply can easier adapt the energy production to the requirement. Thus, 50% of regulation demand of an average settlement could be covered.

1. INTRODUCTION

The Information and Communication Technology (ICT) in Germany causes an energy consumption of 55.4 TWh [2] corresponding to 10.5% of the complete national electrical energy consumption. The consequently resulting carbon dioxide emission exceeds the emission of the whole German aviation. Although consumer electronics – first and foremost the entertainment electronics – take the biggest portion of the ICT, recent calculations [2] demonstrate that especially data centers and the internet will take an ever growing share of the total energy consumption within the next years, because of an exponentially increasing demand. The rising energy demand in combination with increasing energy prices provokes that in 2010 the energy costs will account for up to 50% of the total IT budget, referring to Gartner. For the operators of data centers this development directly results in the need for immediate action to reduce the energy consumption.

For small and medium-sized enterprises, outsourcing of their IT infrastructure is a good opportunity to lower the purchasing and particularly the operating costs. Due to the fact that outsourced IT services are typically operated in large-scale data centers among the services of further enterprises, IT outsourcing is strongly related to the cloud computing domain. This is a good example for the trend towards increasing usage of constantly growing data centers. However, the sustained movement into data centers not only increases the energy consumption but also yields new technological challenges: Besides the well-known problems of heat-removal, ensuring the energy supply is a new challenge, since nowadays large data centers can easily reach power consumptions in the megawatt range and thus are significant large consumers of energy.

A good way of increasing the efficiency and hence the profitability of data centers lies in achieving a higher utilization of the existing computation capacities. IT services usually still get statically bound onto dedicated servers due to maintenance, performance, and security reasons. This static assignment causes low server utilization most of the time, because the available hardware resources are determined by the expected peak load. This leads to a waste of energy, since servers still have 50% and more of their maximum power consumption, even when completely idle. A solution to address this problem exists in decoupling the services from the servers using virtualization. With this technique a service is operated as a virtual machine (VM) while a virtual machine monitor controls the hardware accesses as interface between server and VM. Now several services can be run on a single physical server and thus the servers can be consolidated to reduce their number, which is equivalent to an energy and cost reduction.

Until now much work has been invested in the virtualization domain and many extensions have been proposed. An exemplary and moreover effective improvement is found in dynamic load management, i.e. the dynamic redistribution of services at runtime according to their current resource demand to achieve an even higher server consolidation rate and therefore switch off further servers. Another enhancement expands the load management over several data centers to take advantage of specific (geographically) characteristics like climatic conditions, energy prices, or the energy efficiency of the data centers.

In a survey [1], we have analyzed the possible application areas, the existing challenges and the attainable energy saving potential of dynamic load management inside a data center and in a data center network. This extended abstract briefly describes the basic load management approach, subsequently summarizes our findings of the survey, and delivers insight into required future work.

2. DYNAMIC LOAD MANAGEMENT AND ITS APPLICATION

The basic prerequisite for realizing dynamic load management is the live migration technique. By means of this technique it is possible to move a service as VM between two servers without noticeable interruption of operation. That allows the dynamic allocation of services to the servers by analyzing and using the diverse services' resource demands to use as few servers as possible at anytime. Consequently, servers currently not in use can be shut down to save energy.

Nevertheless there is the problem that an unexpectedly increasing resource demand may not be provided as fast as necessary, because a previously shut down server has to be powered up again which takes a certain amount of time. During this boot up time the services' performance may suffer due to a shortage of server capacities. A solution for the described problem can be found in predictions of the services' future resource demand. If the demand is known early enough, the capacities can be allocated in time and no performance issues will occur assuming accurate forecasts. In order to achieve good predictions the services' resource demand has to possess a periodical behavior which can be projected into the future.

Another challenge is the consideration of other optimization criteria but the minimization of servers. There already exist several detached solutions with different optimization targets, for instance avoiding hot spots[1] in the data center or aiming for an evenly aging of hardware. But these solutions only consider a fraction of the available influencing factors such that their decisions are suboptimal in a holistic view. To achieve the best results in data center energy efficiency all these optimization criteria and its required influencing factors have to be regarded by a superordinate holistic system management. Furthermore for a holistic view the focus should not only concentrate on computing hardware but on the whole data center infrastructure such as climate control or power supply. Just like the adaption of the number of servers to the current resource demand the infrastructure should scale with the actual need so that the whole data center behaves energy-proportional.

If an enterprise maintains several geographically distributed data centers, a load management which comprises all these data centers is applicable. In this case the work load or rather the services will be moved beyond the boundaries of separate data centers. Thus the site-specific parameters of participating data centers can be used to obtain a more energy- and cost-efficient processing of the given work load. For instance, if there are two data centers in regions with considerably different energy prices, it would be smart to process as many services as possible at the lower-cost site to reduce overall costs.

There are several such parameters worth considering, which divide into different domains. The domain of the energy supply provides data about energy prices which will be valid in the nearer future – this corresponds to the aforementioned example. Furthermore, it also provides information about the regional energy requirements which can be used for energy regulation. In this field of application energy demand of the data center's neighborhood is regulated by migrating energy consuming work load from one data center site to another. The aim of this technique, for instance, is to use up an oversupply of energy by migrating work load to this site or to flatten the energy demand in a data center's neighborhood so that the energy supply can easier adapt the energy production to the requirement.

The next domain of site-specific parameters deals with the climatic conditions. Via weather forecasting the sites with the lowest environmental temperatures can be determined to minimize the cooling costs by using free cooling as much as possible. Moreover the forecasting can be used to ascertain the future renewable energy production (e.g. wind and solar energy) of data center sites to decide where to process work load in an ecologically friendly way. Another domain covers the data center's energy efficiency, i.e. the required energy to process the incoming work.

A big challenge is the combination of the site-specific information with the services' utilization profiles to get a holistic optimization of energy consumption, environmental acceptability, and cost reduction, particularly due to the required aforementioned holistic system management of a single data center. Further challenges exist in standardization of the provisioning process of services to different enterprises' data centers as well as the associated billing process, which should only charge the actually used hardware capacities. Moreover additional development for the data safety and security domain is needed so that enterprises have more confidence in saving important corporate data in external data centers. Foremost monitoring tools allowing the enterprise to oversee their corporate data should be developed to provide confidence.

The concept of load management in distributed data centers to reduce energy need and costs is a fairly new research topic. In [3] and [4] ideas for geo diversified server clusters were introduced, which take energy ecological aspects (e.g. the day rhythm of the sun's position) and further aspects concerning service level agreements (e.g. inquiry processing and failure safety) into account. However, a concrete application scenario, showing how it works and in what way certain aspects affect the management decisions, is missed so far. A more detailed description of our idea behind load management in a data center network can be found in [5].

3. POTENTIAL OF LOAD MANAGEMENT

We have performed detailed theoretical evaluations of the potential energy and cost savings when the described load management approaches are applied. Furthermore we determined the potential of data center comprehensive load management for the application as an energy regulation system.

Beginning with dynamic load management in a single data center, we used models representing the different components which can be found inside a data center. Primarily the models provide information about the energy consumption of server, storage, and network hardware, the uninterruptable power supply (UPS), and the climate control with different operating modes (e.g. free air cooling).

As input data for these models we used utilization profiles containing the future hardware resource demand of different IT services, such as ERP, CRM, or data base applications, which posses a certain fluctuation over the course of the day. These profiles base on measured values and include not only utilization data for working but also for weekend days. The remaining input data comprise different environmental temperatures and several data center hardware and infrastructure configurations, while the particular energy efficiency has been derived from the data center's power usage effectiveness (PUE)[2] value.

The savings in energy consumption of selected data centers by using dynamic load management compared to a static binding of services to servers have been analyzed in a theoretical evaluation. The results are presented in Figure 1. For utilization profiles with different magnitudes of fluctuation in resource demand and for

[1] Areas of elevated temperature at the inlet side of computer equipment typically attributed to either a lack of cooling capacity or inability to deliver the cooling where it is needed.

[2] A data center's PUE is defined by the quotient between total and computing hardware energy consumption, i.e. the smallest and best value is 1.0 and higher values denote worse energy efficiencies.

several data center PUE values the potential energy savings range from 20% to over 40%. Data centers with low PUE values and high resource demand variation save on a percentage basis the most energy, which is plausible, because these data centers have a low infrastructure overhead and a high potential for server consolidation.

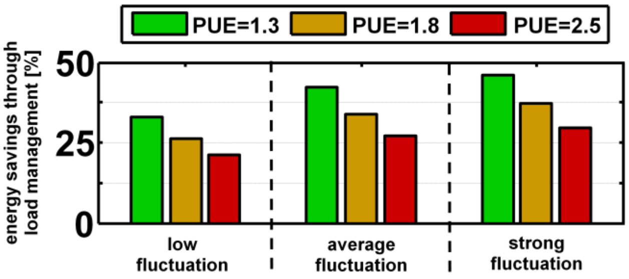

Figure 1. Energy savings in a data center when using dynamic load management compared to a static virtualization approach with preallocated resources.

For the simulation of distributed load management we used a data center network model which bases partially on the data center model for the first evaluations. The simulated data centers were distributed across 16 regions with differences in environmental temperature and energy price, which both change over time and base upon real values. The data centers differ in their size and their energy efficiency denoted again by their PUE values. Furthermore we used again the aforementioned service utilization profiles. Last but not least we included additional costs caused by the distributed load management into our model, which are network, migration, and replication costs as well as costs for extra storage.

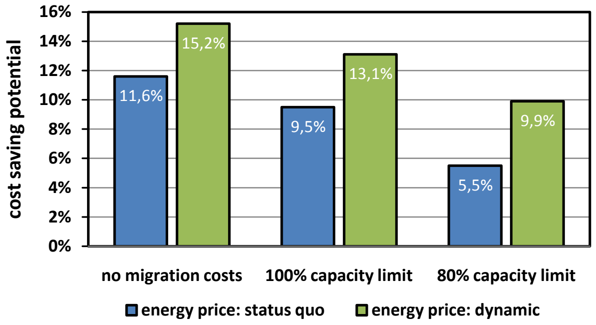

Figure 2. Cost saving potential by using data center comprehensive load management for current (status quo) and future (dynamic) energy price models.

The simulation of data center comprehensive load management begins with an initial position in which dynamic load management is already applied in every data center. So the obtainable savings, which are shown in Figure 2, are additional to the ones inside each data center. These savings have been ascertained for different settings, which affect the capacity limit[3] for servers in a data center, the energy price model, and whether migration costs are included or not. For a current energy price model (status quo) additional cost savings of about 10% are possible, even if migration costs are included. With higher energy price fluctuations, the cost savings will increase further, as seen by the assumption of a future dynamic energy price market with regional energy price differences of up to 30% at the same time in Figure 2 (dynamic). Another raise of cost and energy saving potential is expected, if the simulation accounts for a higher variation by increasing the number of data centers, which was limited for the presented evaluations due to a high complexity.

To get an intuition of the energy regulation potential of a data center comprehensive load management approach we analyzed the energy requirement of an average settlement of 100 households and two industrial enterprises, which is shown in Figure 3 (blue line). The required energy regulation demand is the variation in power consumption over time. By migrating work load to or from a nearby data center this variation of energy requirement can be lowered or even totally flattened so that the energy provider can easier fit the production to the actual requirement. But the energy regulation potential is directly coupled to available network bandwidth because each service migration requires at least the working memory transferred. In a first evaluation we analyzed this potential and figured out that by using data center comprehensive load management, 50% of the regulation demand in the mentioned settlement could be covered, if the sites are connected via a next generation network with a bandwidth of 10GBit/s.

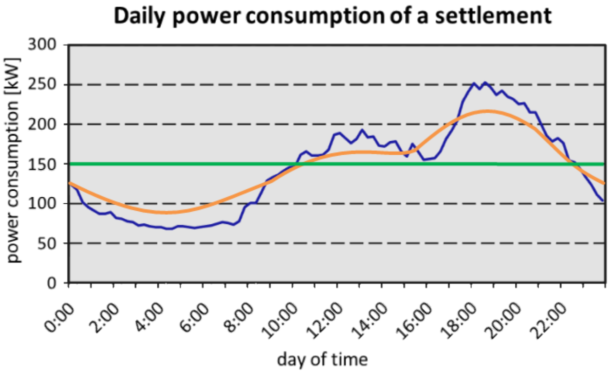

Figure 3. Energy requirement could be regulated to a lower dynamic or even be totally flattened. In turn power stations could be operated at their most efficient operating point.

A more detailed description of the performed evaluations can be found in the survey [1].

4. CONCLUSIONS

The evaluation results show that load management inside a data center as well as in a data center network is a good way to save energy in the fast-growing server market. The internal load management has an energy saving potential of 20% to over 40% and the distributed load management could save additionally 10% and more for German energy price structures. In a similar survey [6] the potential in the US has been analyzed with the result of cost savings of up to 45%, which we could approve by repeating our simulation with the energy prize models used in [6].

Further research work for realizing data center comprehensive load management is necessary, but it encounters some challenges, which have to be addressed. As a start a data center model which includes heterogeneous server environments and the topology for the cooling overhead has to be created. On the basis of this model a holistic system management for optimizing the energy efficien-

[3] A server's capacity limit represents the maximum fraction of resource capacity, which may be allocated to services. The residual fraction is used as backup, if services need more resources than scheduled.

cy of a data center can be developed. This system manager then should also include dependable utilization forecasts of services and power management of the cooling infrastructure. The data center comprehensive load management needs this holistic view as well, but in this case further site-specific influencing factors like energy prices, environmental climatic conditions, and network costs have to be considered additionally. And yet another critical task is the development of adaptive replication (placing) strategies, which are indispensible for distributing the service's permanent data and thus lay the foundations for migrating services between data centers.

5. REFERENCES

[1] W. Nebel, M. Hoyer, K. Schröder and D. Schlitt, *Untersuchung des Potentials von rechenzentrenübergreifendem Lastmanagement zur Reduzierung des Energieverbrauchs in der IKT*, Survey, OFFIS, 2009.

[2] L. Stobbe, N. Nissen, M. Proske et al., *Abschätzung des Energiebedarfs der weiteren Entwicklung der Informationsgesellschaft*, Survey, Fraunhofer IZM and ISI, 2009.

[3] K. Church, A. Greenberg and J. Hamilton, *On delivering embarrassingly distributed cloud services*, In HotNets, 2008.

[4] N. Haustein and R. Krause, *System and method for selecting data centres based on environmental conditions and energy efficiency*, White Paper, IBM Deutschland, 2008.

[5] K. Schröder, D. Schlitt, M. Hoyer and W. Nebel, *Power and cost aware distributed load management*, In eEnergy 2010, Passau, Germany, 2010.

[6] A. Qureshi, R. Weber, H. Balakrishnan et al., *Cutting the electric bill for internet-scale systems*, In ACM SIGCOMM, Barcelona, Spain, 2009