Incentives to Apply Green Cloud Computing

Tommi Makela[1] and Sakari Luukkainen[2]

Aalto University, Department of Computer Science and Engineering, Espoo, Finland,
[1] tommi.makela@aalto.fi, [2] sakari.luukkainen@aalto.fi

Abstract

In recent years, there have been two major trends in the ICT industry: green computing and cloud computing. Green computing implies that the ICT industry has become a significant energy consumer and consequently, a major source of CO2 emissions. Cloud computing makes it possible to purchase IT resources as a service without upfront costs. In this paper, the combination of these two trends, green cloud computing, will first be evaluated based on existing research findings, which indicate that private clouds are the most green option to offer services. Hosting of private clouds can be outsourced, which allows companies to focus on their core competences. Furthermore, three case studies of state-of-the-art companies offering green hosting services are presented and incentives affecting their energy-efficiency development are analyzed. The results reveal that currently there is no demand in the market for green hosting services, because the only incentive for companies is low costs. Service providers should illustrate their greenness with transparent efficiency metrics, draw up green service level agreements and compete with greenness. Then it is up to end users to require more green web services and create derived demand for green cloud services and green hosting services.

Keywords: Private cloud, Hosting service, Derived demand, Efficiency metric, Green service level agreement

1   Introduction

The information and communication technology (ICT) industry utilizes a variety of equipment and infrastructures to offer services. Large-scale infrastructures that host equipment are referred to as data centers. Companies can either own a data center or they can rent floor space from data centers maintained by commercial actors. Even though the ICT industry can be beneficial for environment, such as by reducing the need for travelling, it also causes plenty of emissions when the equipment used by the ICT industry is designed, manufactured, operated, and disposed. The research group Gartner estimates that the ICT industry produced up to 2% of the worldwide carbon dioxide (CO2) emissions in 2007. The figure included most of the equipment and infrastructures used by enterprises and governments, but also computers and mobile phones purchased by consumers. Gartner forecasts that financial and legislative issues would eventually force the ICT industry to become greener in future. Furthermore, Gartner predicts that customers will start to demand more environmentally friendly ICT services. (Site 1)

In addition to environmental awareness, cloud computing is an another trend which has had a major impact on the ICT industry. Companies no longer need to possess IT equipment or plan resource provision to provide services, because they can purchase all of the required resources from cloud providers, which in theory have an unlimited amount of resources. Proper resource provision has typically been problematic, because under-provision causes revenue losses while over-provision wastes capital and electricity. Cloud computing allows companies to eliminate their up-front costs and shift them to operational costs. Furthermore, companies only need to pay for resources that they actually use, which improves the companies' ability to adapt to market changes and avoid potential losses when demand suddenly decreases. [2] However, these problems become cloud providers' problems and it is up to them to optimize computation and minimize latencies in order to gain some profits [8].

Cloud computing should strive for more green computation, because cloud resources are presumably highly utilized, when compared to the traditional client-server model [9]. However, what are the most green cloud configurations in the market? Furthermore, what are the incentives for companies to apply more green computation, especially when cloud services have characteristic threads, including data compromises, insecure interfaces, and vulnerable shared infrastructures [4]? To study these questions, this paper is divided into three main sections: a literature review with the evaluation of green computing, case studies of novel green hosting services for private clouds, and a discussion about the incentives of green cloud computing.

2   Literature Review

The main literature findings related to green cloud computing are presented in this section.

2.1   Green Computation

Microsoft has studied if its cloud services are more green than its on-premise applications for business customers. Three business applications, Exchange, Dynamics CRM, and SharePoint, were selected for comparison. The environmental impacts of these applications were defined by using small, medium, and large user groups. The comparison indicates that cloud services, which provided applications for a large group of 10 000 users, had 30% smaller carbon footprint than on-premise solutions. When a small group of 1 000 users utilized these products, an even larger gap existed, because cloud services released 90% less CO2 emissions than on-premise solutions. [1]

Microsoft has also listed the key factors of cloud services that enable low electricity consumption per end user. One of the factors is server utilization, which is high when compared to similar on-premise solutions. The power draw of servers only slightly increases when utilization rate rises. For example, a four times higher utilization rate causes two times higher electricity consumption. Better use of resources reduces cloud service providers' need to purchase new servers in the short term. High server utilization is achieved with multi-tenancy, which centralizes computation and offers uniform infrastructure to multiple companies and their web services. Because of multi-tenancy, demand peaks of individual companies can be combined to a more predictable and steady demand rate which then helps to allocate resources. Furthermore, the gap between peak and average demand in cloud services is small in comparison to conventional servers and as a result cloud services require less extra resources during rush hours. [1], [9]

Cloud services can be scaled based on workloads [2]. For example, during demand peaks web services scale out to acquire more instances that serve extra end users. On the other hand, when web services do not need all their instances, they shut down their spare resources and reduce operational expenses. If several hosted services scale in their resources at the same time, cloud providers should put their underutilized servers into sleep in order to centralize workloads to remaining servers and to reduce electricity consumption [17]. This feature is also known as dynamic provisioning, and it is based on advanced IT equipment which monitor network traffic and predict upcoming demand. It strives to prevent the over-provision of servers and to reduce capital and operational expenses. [1]

Some measurements, however, indicate that cloud services still have underutilized servers, even if they supposedly have advanced mechanisms to centralize computation. It seems that servers rarely have multiple instances at the

Tommi Makela
Sakari Luukkainen

same time or instances are mostly idle, and as a consequence the average utilization rate in cloud services can be below 10%. [12] Furthermore, there are also some performance issues related to provisioning and booting instances, because latencies vary from one to ten minutes. If instances are activated too slowly, web service providers will permanently lose some of their end users that are too impatient to revisit on their services. Notable latencies imply that cloud software stacks still need to be improved so that cloud service providers could use sleep modes on their servers without compromising the quality of their services. [9], [16] Scheduling algorithms that enable sleep modes on servers and instances need to perform so efficiently that the end user will not notice their existence. [15]

Jayant Baliga and his associates have also examined the greenness of cloud services. The study revealed that energy savings can be gained by using cloud services, when compared to thick clients. Nevertheless, the increased amount of network traffic seems to have a negative impact on the overall power draw and, therefore, the most green of cloud services is not unambiguous. Conducted measurements and simulations indicate that private cloud as a deployment model is the most energy efficient option. The main reason for this result is short distances between data centers and end users. Especially if private clouds are placed in the same premises where the majority of the staff is working, only a few network devices are required to transfer network traffic between end points. Because of the large amount of network traffic, frequently needed data should be stored locally, while rarely used data, such as backup files, could be stored in cloud services. Furthermore, large files, including high-quality videos, should be stored close to the end users, even if files were fetched rarely. [5]

To maximize the greenness of cloud services, highly utilized servers need to be hosted within highly energy-efficient data centers. The energy efficiency of a data center infrastructure can be measured with power usage effectiveness (PUE) metric, which is calculated by dividing total electricity consumption with IT electricity consumption [1], [20]. The average PUE value is around 2.0, which indicates that half of the electricity is used by the infrastructure, including cooling systems, power transformers, and lighting [1], [19]. Due to the magnitude of cloud services, cooling costs in data centers can be substantial. For example, in 2010, data centers used globally 130 billion kilowatt-hours to cool servers [6]. Traditionally data centers are cooled using Computer Room Air Conditioning (CRAC) units, which blow chilled air though raised floor in front of the server racks and suck hot air near the ceiling. Units utilize cool liquid to exchange heat with air and the liquid is cooled using chillers, which are significant energy consumers. [6], [17]

According to Microsoft, state-of-the-art data centers for cloud services strive for extremely low 1.1-1.2 PUE values, which can be achieved by adopting new energy-efficient methods to cool servers, including modular design, outside air and water evaporation based cooling [1]. Furthermore, novel cooling systems should be dynamic in order to adapt different workloads. For example, a cooling system should notice when servers are powered down and modify its air flow patterns. It would require numerous sensors to collect information about servers' cooling needs. [15] There are also more energy-efficient methods, including liquid cooling, to shift heat in data centers than CRAC units. Liquid as a medium is more capable to obtain heat energy than air, and therefore liquid should be brought as close as possible to server racks. [19] Another strategy to gain energy savings is more hot data centers, which require less cooling. Recommend operating temperatures for servers are 20-25°C [8], [17].

2.2   Towards Green Private Clouds

In addition to environmental aspects, the overall expenses of different cloud solutions need to be carefully evaluated before adopting them. Conventional facilities and especially green data centers tend to have substantial investment costs, whereas public clouds have no upfront costs. However, public clouds could be more expensive than acquiring own infrastructure and servers in the long term. First of all, cloud service providers need to have a marginal profit in order to renew their hardware from time to time and to develop their processes. As a consequence, instances are not as affordable as they could be. Secondly, transferring data can be a greater item of expenditure than instances for some customers. [21] It seems that data transfer costs can be substantial for those companies which have many end users and often need to shift a significant amount of data to a cloud service. In the long term, a private cloud would be more affordable as well as more environmentally friendly when it is hosted within a green data center.

Perhaps the most significant drawbacks of cloud computing are related to data privacy. First of all, customers need to trust that cloud providers are able to protect the confidentiality of data on shared platforms [4]. Furthermore, they also need to trust that cloud providers will not exploit their sensitive data for commercial purposes. Especially major actors could have a great interest to sell customers' data for advertisers. In this sense, private clouds in companies' own premises are safer options to store sensitive data than public clouds. Even though companies trusted public cloud providers, existing legislation could be a major problem. Some data protection laws might require that certain information must be kept inside specific regions, such as the European Union [23]. Many of these restrictions are justified, because some other countries might have loose laws to protect privacy and some governments possibly have excessive rights to explore stored data on cloud services in their regions. As a result, private clouds could be the only option for those companies who want to apply cloud services and store sensitive data.

The literature implies that private clouds are currently the most environmentally friendly solutions in the market [5]. Furthermore, private clouds overcome many of the cloud-related issues, including the previously listed privacy issues. Because of the small number of end users, individual households and small businesses cannot benefit from private clouds, which would lack of economies of scale. Furthermore, centralized resources in a small scale would not offer significant energy savings compared to servers and workstations [1]. Companies with less than a thousand users on

their servers should consider using public clouds or virtualizing their current servers to gain energy savings. Then it is up to small companies themselves to determine which of the public clouds would be the most desirable for them; for example, the most affordable instances near the end users. Large companies should, however, consider adopting private clouds, mainly because data transmission costs and distances to cloud data centers are so substantial.

For those companies that decide to put instances into operation, there are at least three sub-deployment models of private clouds from which to choose (figure 1). A private cloud is commonly referred as a cloud which is operated and hosted by the company itself. IBM specifies that there are also private clouds that are both operated and hosted by a third party. [18] This option resembles a public cloud deployment model, because IT equipment and data center infrastructure are fully controlled by a service provider. This option, however, separates customers more firmly than public clouds, because each company has its own resources that are not shared with others. As a consequence, this option provides better privacy than public clouds. Then there are private clouds that are operated by a third party and hosted by the company itself [18]. In this option, the IT maintenance is outsourced to service providers with more IT expertise, such as Hewlett-Packard, Dell, or IBM. Furthermore, there are also private clouds that are operated by the company and hosted by a third party [18]. This option allows companies to focus on their core competences, because they do not need to take care of the data center infrastructure. Because the infrastructure is fully outsourced, it is up to the companies to select a service provider which would host servers within a novel facility with minimal electricity consumption and CO_2 emissions. As a matter of fact, the combination of green hosting services and self-operated private cloud services could be the most energy efficient option in the market.

Figure 1: Different types of private clouds

Based on the literature, it could be argued that cloud computing is greener than the traditional client-server model [1], [5], [21]. However, are there any incentives to apply green cloud computing in practice? The next section introduces three companies with novel facilities and cloud services. Their services resemble the first and the third private cloud types (figure 1), where the third parties host and operate private clouds or only host private clouds.

3   Case Analysis

The selected research methodology and applied framework for this study are introduced in this section.

3.1   Methodology

The literature review forms a basis for a framework (table 1), which will used be in the empirical section of this paper. It helps to collect relevant information about green cloud computing and limit the scope of the research. Based on the literature, there are four factors, which define green cloud computing. The first category focuses on the methods to centralize computation in order to increase server utilization. The second category examines cooling systems which would use significantly less energy than traditional CRAC units to cool data centers. The third category concentrates on the possible issues between energy efficiency and other requirements of IT services. Finally the fourth category emphasizes the importance of efficiency metrics to illustrate the greenness of services to customers and end users.

Table 1: Research framework

Dimension	Meaning
Server utilization [1], [5], [21]	Approaches to centralize computation in companies
Cooling systems [6], [19]-[20]	Solutions to chill facilities more energy efficiently
Quality of Service [15], [17], [22]	The relationship between greenness and quality
Efficiency metrics [3], [7], [15], [20]	Methods to illustrate the greenness of services

The goal of this empirical section is to identify features of state-of-the-art facilities that would provide new knowledge about the incentives of green cloud computing and green hosting. This study was conducted by using the multiple-case-study research methodology, where each case should be treated as a single experiment. If similar results are obtained in several cases, a replication has taken place. By using the case study method, it is possible to observe the phenomena specific to the cloud computing industry in a detailed level. While the secondary sources mainly describe what has been done, the explanatory case studies provide more detailed insight into the interrelationship

Tommi Makela
Sakari Luukkainen

between the energy efficiency development and related incentives. Case studies also facilitate concentration on the new factors that arose during the research and the reasons for the differences in the case companies' behavior. [24]

The information for this study was collected by interviews. According to the literature, interview studies as a research method have several benefits. They allow interviewees to express themselves as freely as possible. Secondly, if the field of research is uncharted or unknown, it is difficult to expect certain answers beforehand in which case interview studies are a very flexible method to apply. Moreover, they are also suitable when the collected data needs to be used in a wide context. In some cases it is predictable that the topic of research will result in complex answers, which require clarification. Furthermore, supplementary questions can be asked during the interviews, if the original questions were imperfect or the answers were so insufficient that they needed clarification. [11]

In this study, all the listed benefits were identified. It is useful that an interviewer can ask for clarification if he needs more justification for some of the responses. Moreover, it seems that questions for interviews do not need to be as accurate and explicit as questions for questionnaires, because an interviewer can ask follow-up questions, which are especially useful, if interviewees offer unexpected new information. For example, in this study additional questions clarified what kind of cloud services participating companies offered and how their business customers evaluated environmental aspects. Moreover, the research topic was somewhat uncharted, because the green data centers were either under construction or just put into operation. Furthermore, the main goal was to utilize the interviews in a holistic view to green cloud computing incentives and as a result the interview case study was selected as a method to collect empirical evidence to support the conclusions of this study.

Naturally, interview studies also have some drawbacks when compared to other methods to collect information. To maximize the amount of useful data collected from social interactions, an interviewer needs to have experience. As a result, those who want to perform surveys should train themselves to become skilled interviewers. Furthermore, performing interviews is time-consuming, because suitable interviewees first need to be located and then interviews need to be arranged separately. Transcribing informal material after interviews also requires plenty of time, because there are no ready-made models that would simplify the process. Moreover, results typically suffer from reliability issues, because people could behave abnormally in social interactions; for example, an interviewee wants to please an interviewer and therefore he provides answers that are expected by the interviewer. [11]

In this study, some of the proposed drawbacks were recognized. Definitely more experience about conducting inter-views would have been helpful, but all the interviewees were so competent and cooperative that they offered plenty of useful information, which was not directly related to the questions. In other words, it is very important to locate the correct people for interviews, who are able to answer questions and provide additional information. In this study it is unlikely that interviewees would have provided any false information, which would have impaired the reliability of the interview study, because some of the data was already public. The interviews, however, offered numerous technical details and provided a better overview of facilities. Furthermore, the fact that the companies remain anonymous in this study increases the reliability of their answers. However, if the questions were unintentionally related to their trade secrets, they declined to answer, which is completely understandable. It is rather difficult to determine if some of the questions are too sensitive; for instance, information related to the exact locations of facilities.

The goal was to locate domestic companies, who have promoted their services or data centers as green, and then agree appointments. Up to ten companies were asked to participate in the interviews. In the end, three companies agreed to participate, but luckily those companies had the most promising green data centers in the market. The participating companies had rather different business models, so the main common factor of the companies was brand-new facilities and different types of cloud services. All the participating companies could be considered as major actors in their field in Northern Europe. The main goal was to discuss new facilities and offered services with personnel, who have great knowledge of the topic. Luckily all of them who were initially asked for the interviews were able to participate, which then assured high quality answers. Suitable personnel were located from the companies' web sites, which often contained news sections, where the new data centers were announced and their project managers were listed. The personnel were mainly contacted via email. All the interviews were conducted in the premises of the companies. Two of the interviewees were IT managers and one of the interviewees was the CEO of a company, which then co-owns the actual company with new facilities.

The questions for the interviews were posed based on previous experience from data center infrastructures and basic knowledge of cloud computing. The questions involved three different themes: infrastructures, cloud services, and transparent efficiency metrics. Two of the interviews took an hour to perform, which was the estimated time for all interviews, but one of interviews took over one and half hours, because the company had so much say about their products and services. Nevertheless, the resulting wide range of answers was just a positive outcome for this study. Moreover, two of the companies had one extra person with technical knowledge to offer more details of the facilities. To ensure that all the available data could be fully exploited, all the interviews were taped with a voice recorder. Furthermore, notes, which were written during the interviews, were also utilized. The summary of the interviews is presented in the following sections and its content is approved by the companies.

3.2   Company A

Nowadays, company A offers numerous IT services, including equipment hosting, communication solutions, support, and voice-related solutions for business customers. Company A has its own data center facilities, core networks, and system experts. The first listed service, equipment hosting, allows companies to rent an infrastructure for their IT equipment, particularly for rack servers. Company A also offers some management services for hosted servers, but customers typically purchase and configure their servers by themselves. Equipment hosting has similar benefits as public cloud services, because companies can focus on their main businesses, thus reducing capital expenses.

All the servers are hosted in brand-new data centers in Helsinki. The first facility was officially introduced in the fall of 2010 and only one year later the second facility was in operation. The first one is compact, because it is built in an old bomb shelter with 500 m² of floor space. It could be considered as a proof-of-concept data center, because the second data center has as much as 5 000 m² of floor space. Nowadays, company A claims that their facilities are one of the most energy efficient in the market, because they have 1.25 PUE and they reuse plenty of waste heat.

To achieve substantially low PUE, the cooling system of a data center needs to be extremely energy efficient. Both of company A's data centers are quite unique, because they use district cooling to chill hosted servers, and the same pipe system is used to transfer waste heat to the city's district heating system. A local energy company mainly uses sea water to chill circulating refrigerant within the district cooling system. During the summer, however, the energy company utilizes waste heat generated by its power plant to create pressure to the refrigerant, which evaporates, releases its own heat, and finally cools down. Typically, the circulating refrigerant has too low a temperature when it returns from the data centers, and therefore, heat pumps are required to provide additional heat before waste heat can be shifted from the district cooling system to the district heating system.

At the time of the interview, company A did not benefit from the waste heat, which was shifted to the city's district heating system. This sounds a little odd, as company A could be considered as a small scale energy company, because the first data center generates so much heat that up to 500 detached houses can be maintained warm all year round. Moreover, when the second data center is full of server equipment, the amount of released waste heat equals the heating requirements of 4 500 households with 80 m² of floor space. These figures indicate that company A provides notable quantities of heat for the energy company and therefore it would be reasonable that data centers would at least have low prices for district cooling and electricity.

To achieve substantially high energy efficiency and to maximize the benefits of district cooling, the company utilizes liquid-cooled racks to chill hosted equipment. There is one slight difference between company A's two data centers. Due to the limited space in the first facility, all the IT equipment, including servers, network switches, and storage devices are cooled by liquid cooling. The second facility, however, uses liquid cooled racks only for IT equipment with high heat density and the rest of the devices are cooled with CRAC units. Based on the interview, liquid cooling for all IT equipment does not guarantee energy savings while it has rather high upfront costs. Company A considered applying economizers to replace CRAC units during wintertime, but there have been permission related issues.

The energy efficiency of hosted servers is customers' concern, meaning that the customers decide what kind of server models they want to purchase. For some reason, many customers, including telecommunication companies and major enterprises, want to host rack servers instead of blade servers. The majority of company A's business is related to hosting equipment. However, it also offers virtual machines, which operate in fault-tolerant clusters. Customers can upgrade purchased virtual machines afterwards, if they need more resources. Due to some missing cloud characteristics, company A wisely does not promote its virtual machines as instances.

At the time of the interview, it was difficult to determine how quickly the capacity of the data centers runs out. For example, if several large-scale companies decide to outsource the hosting of their private clouds, the data center capacity would then rapidly be exhausted. To minimize the costs of a partially empty facility, the second data center is built by using modular design, meaning that new sections will be put into operation when demand rises. When the facility was introduced in 2011, it already had a major customer, which offers numerous IT services and outsourcing solutions, including a cloud infrastructure for companies. It is interesting that the customer promotes its new product as green cloud computing. In other words, the combination of the customer's highly utilized cloud resources and the company A's highly efficient infrastructure can be considered as green cloud. From the business point of view, this is exactly what companies should do when they host their IT equipment in a green data center.

Company A has other unique features as well. For instance, the electricity consumption of every hosted server is separately measured, and the billing is based on these measurements instead of theoretical maximum electricity consumptions and fixed prices. This arrangement encourages companies to purchase energy-efficient server models. The company also offers water and wind power for companies which want to provide even greener services. The conducted interview revealed that none of the customers had purchased renewable energy. The price tag of supplied electricity was more important than the amount of released CO_2 emissions.

Tommi Makela
Sakari Luukkainen

3.3   Company B

Company B provides IT support and computational resources for companies, research institutes, and universities. It is administrated by the Finnish Ministry of Education, which naturally means that the company does not try to make any profit, but tries to maximize the value of taxpayers' money. In Finland, company B is one of the actors with the expertise, equipment, and all the necessary software tools to help scientists conduct their research.

Company B has two known data centers and a few secret facilities. To provide plenty of computational resources for scientists, company B has four supercomputers for performing computationally heavy tasks. When compared to conventional servers, supercomputers have substantially more processor units and main memory modules. The conducted interview indicates that the company has a challenging task to maintain its supercomputers on the top 500 most powerful machines listing, but at the same time they should be on the 500 most energy-efficient machines listing as well. Therefore, company B will acquire a new supercomputer for its new data center.

In the early stage of the project, the general confusion in the public was that company B was going to build a data center in the premises of an old paper mill in Kajaani. Actually, company B is a customer of a data center park, which is being developed mainly by a local energy company, a telecommunication company, and a paper manufacturer. In the first phase, 1 000 m² floor space will be allocated for a new supercomputer and other IT equipment, which will be shifted from existing data centers. In total, around 4 000 m² of floor space is reserved for the company. Because the premises are still under construction, it is difficult to determine the final PUE value for the data center. However, the company expects that the PUE would be around 1.15, which means that the infrastructure would drain significantly less electricity than hosted servers, which are rather power intensive due to their heavy tasks. When discussing the potential risks of the building project, the only major issue for company B would be unrealized energy efficiencies for the infrastructure. Therefore, it is important that the local energy company and the telecommunication company measure the electricity consumption of individual customers as accurately as possible.

Low PUE values require an especially energy-efficient cooling system. The company has decided to focus on airside economizers, because outside air is chilly the whole year around in Kajaani. As a matter of fact, the region is so cold during the winter that airside economizers will be modified to handle outside air below -45°C, before air is shifted directly indoors. Because the data center will be built in the paper mill, there are no residential areas nearby, which would make it difficult to utilize waste heat. The offices and warehouses inside the data center park, however, can be warmed by using hot exhaust air from the servers. Obviously, it does not make sense to build a separate district heating system if the distances are too long. To gain more energy savings, the company will follow the American Society of Heating, Refrigerating, and Air-conditioning Engineers' (ASHRAE) latest recommendations, which state that the maximum temperature for supply air is 27°C. Because certain computer components are regularly upgraded and replaced, higher operational temperatures should not significantly affect supercomputers' reliability.

Company B has several blade systems providing an infrastructure for cloud services, which are available for partner companies and universities. The company uses open source software stacks to manage cloud resources. The cloud is small in comparison with commercial products, which means that resources will not scale endlessly. This private cloud is an alternative for scientists who need to process sensitive data. In addition to a lack of support for software tools, commercial high-performance cloud clusters become notably expensive in the long term.

To be green as possible, company B will buy hydro-electric power for its IT equipment. As a matter of fact, there are hydro-electric power plants providing clean electricity near the data center park. These power plants, however, are rather small and therefore they are better suited for backup power. Due to the magnitude of the power draw, hydro-electric power will be purchased from Nordic electricity markets. There are up to five major transmission lines supplying electricity from the main grid and it is very unlikely that all of them would break down at the same time. As a result, the company will not install any uninterruptible power supplies (UPS) into its data center.

3.4   Company C

Company C offers several IT services and product development services for business customers in Northern Europe. Typically, all the services are tailored for each customer, which means that the content of services must be negotiated separately. Company C wants to provide extra value for its customers and as a result it will always have some control over hosted IT equipment and infrastructure. In this sense, company C does not offer similar hosting services as company A, where only the infrastructure is rent for rack servers. Company C typically purchases or allocates existing IT equipment for its customers, who then pay monthly fees.

In 2011, company C put its novel data center into operation. This particular facility is located at an industrial area in Espoo, which is one of the major cities in Finland. Furthermore, the facility has up to 6 000 m² of floor space to host customers' services. It is quite large in comparison with company C's five other facilities located in Finland and Sweden, because they have around 7 600 m² of floor space in total. In the first phase, only 1 000 m² of the new facility is fully equipped with a cooling system, redundant power distribution, and backup energy. The conducted interview indicates that it is more likely that electricity would run out before available floor space, especially because

80

company C tends to host high-density IT equipment. Therefore, the sufficiency of local power distribution alone is a major reason to favor energy-efficient solutions. Based on early measurements and future estimates, the PUE value of the data center will vary between 1.2-1.3. PUE value is the only metric which is constantly measured in company C's facilities. The interview indicates that accurate measurements are problematic. A data center infrastructure needs to possess equipment that measure every rack full of servers and storage devices as well as the components of cooling systems and backup systems separately. Furthermore, company C has additional challenges, because in the same data centers it hosts both customers' servers and equipment that are dedicated only for its own internal processes. Measurements are even more difficult to conduct if the customers and the company itself utilize the same IT resources, such as a shared cloud platform. In other words, there are many complex equipment-level measurement issues. For instance, how to collect relevant information from individual servers, instances, and software and then compose useful analyses without interfering with hosted services?

Similar to company A in Helsinki, company C also utilizes district cooling to chill its devices. In practice, a local energy company uses heat pumps to cool down the refrigerant that returns to the data center and to warm up the refrigerant that comes from the data center before waste heat is shifted to the city's district heating system. In Espoo, the district cooling does not involve any free cooling. Even though only 1 000 m² of floor space is currently equipped with data center infrastructure, as much as 1 500 detached houses can be warmed the whole year around. When the full capacity of the new data center is in operation, up to 90 000 households can be heated merely by using waste heat and heat pumps. It has been estimated that the data center makes it possible to reduce CO_2 emissions by roughly 10 000 tons, when the data center replaces other heat production that utilizes coal or oil in its processes. When the interview was carried out, company C's negotiations with the local energy company about the contracts were still underway. Nevertheless, it was clear that district cooling would be more affordable, because company C provides plenty of waste heat to the local energy company in return.

Company C uses liquid-cooled racks to chill equipment energy efficiently. There are products that isolate servers and keep the air loops short inside the cabinets. Company C has tested different types of liquid cooling solutions and it has determined that fully closed and tailored solutions are not necessarily more green or energy efficient. According to company C, inexpensive alternatives, which isolate commodity racks with extra doors and walls, are capable enough to prevent cold and hot air mixing with each other, and it is not completely necessary that the liquid cooling units are attached to the racks themselves. These kinds of solutions provide flexibility, because they can be applied in existing data centers. As previously stated, liquid cooling has major upfront costs compared to CRAC units. Due to the very competitive markets, keeping investment costs as low as possible is important for the company, which is why it favors flexible and modular solutions with both low initial and operational costs.

Company C offers two types of cloud services: individual private clouds for business customers as well as a hybrid cloud. For major customers, it also provides additional security services for private clouds. A customer can select if its data is stored in genuine private storage or in shared storage. Those customers who want to outsource even more can purchase software as a service with shared resources on the equipment level. The hybrid cloud is an alternative for customers who do not need massive amounts of instances and for whom privacy is not an issue. The billing of these cloud services is rather fixed compared to commercial public cloud services.

3.5   Discussion

The results of the multiple-case-study are summarized in the research framework

3.5.1 Server Utilization

In order to minimize the power draw of the hosted servers and to reduce the need for future server investments, all three companies use server virtualization to centralize their computation. Two of the companies exploit commercial products, including Citrix XenServer and VMware vSphere, to offer virtual machines to business customers. These kinds of products guarantee a certain level of quality of service and provide extensive IT support, which are often required by business customers. Company B, however, uses open source software stacks, such as OpenStack, to offer cloud services to selected partners. As a consequence, company B has a great role to ensure the sufficient functionality and reliability of the services. Both commercial products and open source solutions allow the companies to monitor their servers and to determine when extra resources need to be put into operation. When the servers are fully reserved for numerous instances, then it is up to customers to maximize the utilization of their instances.

In addition to better use of resources, cloud software stacks improve the quality of hosted services, because running instances can be moved to other devices when server malfunction occurs. Reliable software allows service providers to use unreliable commodity hardware to reduce their capital expenses and to offer affordable services to end users [8]. All three companies, however, utilize conventional blade and rack servers together with software stacks and therefore their services should be fault-tolerant enough to fulfill the high demands of business customers.

3.5.2   Cooling Systems

Based on the interviews, state-of-the-art data centers have similar features (table 2). Due to the substantial scale of the facilities, all three companies have applied modular design to postpone some of the capital expenses and to

minimize the risks related to major facility investments. Furthermore, modular design guarantees sufficient workloads for cooling systems, which then operate more energy efficiently. Two of the data centers use district cooling systems to chill hosted devices and one of the facilities utilizes a fully independent cooling system, which is based on airside economizers that exploit the cool of outside air. Energy-efficient liquid cooling seems to be the perfect combination with district cooling, especially when the waste heat is collected and then utilized. Furthermore, heat pumps are an essential part of district cooling and heating, because they offer both additional cool and heat. The ability to collect heat seems to be one of the features which modern data centers should support nowadays. The energy efficiency of the facilities is around 1.2, which is the same as the estimated 1.1-1.2 PUE for state-of-the-art data centers [1].

Table 2: Cloud and hosting services

Features	Company A	Company B	Company C
Facility design	modular	modular	modular
- Floor space	5 500 m²	4 000 m²	6 000 m²
Cooling system	district cooling	air cooling	district cooling
- Free cooling	heat pumps, sea cooling	airside economizers	heat pumps
- Used medium	liquid + air	outside air	liquid + air
- Heat reuse	district heating	warehouses	district heating
- Uniqueness	accurate power measurements	hydroelectric power, no UPS	tailored services for enterprises
Metrics	PUE	PUE	PUE
- Values	1.25	1.15	1.2-1.3

3.5.3   Quality of Service

The interviews indicate that even the most green companies in the market are not willing to sacrifice the quality of their services. Both hardware and software solutions need to be highly reliable. Naturally the supply of electricity to data centers is very important for all service providers. For example, company A uses a large diesel generator to produce electricity for its facility, if nearby major transmission lines are down. The company also owns UPS devices to handle short-term power outages. Both of these measures of precaution reduce the greenness of the facility, due to the occasional CO_2 emissions of the diesel generator and the power losses of the UPS devices. Companies A and C state that these kinds of measures are mandatory for cloud and hosting service providers, because business customers require high availability and reliability in all circumstances. In order to avoid confusion between different actors, the quality of services need to be determined in service level agreements (SLA).

The expectations towards quality are the reason why the companies have not applied any experimental products which would provide energy savings, but would possibly have a negative impact on the reliability of the services. For example, there are some network solutions, including ElasticTree, which utilize algorithms and alternative network topologies to minimize the power draw of network equipment, but all three companies have decided to use traditional network switches and topologies [10]. Furthermore, they have not experimented network virtualization, even though it could have similar benefits as server virtualization. For instance, OpenFlow allows centralizing network traffic flows and turning off network switches with no traffic [13]. Based on the interviews, cloud and hosting service providers need to carefully evaluate their software and hardware investments and try to find the balance between different requirements, because extremely green services are useless, if end users are unable to use them.

3.5.4   Efficiency Metrics

Even if all these facilities consist of innovative solutions, all three companies still mainly use traditional PUE values to illustrate their greenness. However, there are several other efficiency metrics which would illustrate the relationship between electricity consumption and emissions and present the ratio of utilized waste heat [3], [7], [20]. As a matter of fact, these metrics are not required by the business customers, such as web service providers, because they only emphasize low costs and therefore there is no demand for green cloud services or green hosting services. Currently, the only incentive for companies to enhance their greenness is to lower their own operational costs and possibly improve their competitiveness by reducing their prices. Companies A and C are major actors in their fields, which is why these findings are significant. If business customers themselves do not require green cloud services or green hosting services, it is up to end users to create derived demand (figure 2) for those services. The end users of web services and their environmental awareness are the incentives which are currently missing.
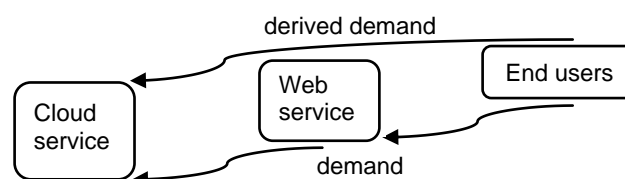

Figure 2: Derived demand for cloud services

Tommi Makela
Sakari Luukkainen

If end users start to increasingly appreciate and demand more green web services, then service providers need to illustrate their greenness. Web service providers and hosting service providers should negotiate and draw up SLAs, which would determine the quality of services and acceptable ranges for different values. Green SLAs underline the importance of transparent and uniform metrics that illustrate energy efficiencies and CO2 emissions [15]. A widely adopted PUE metric alone does not offer adequate information about the energy efficiency of services. To gain more comprehensive overview on the greenness of services, PUE needs to be used in conjunction with several other efficiency metrics. Energy Reuse Effectiveness (ERE); for instance, illustrates service providers' ability to utilize waste heat. It is calculated by subtracting total energy consumption with reuse energy and then divided by IT energy. [3] Another useful metric, Technology Carbon Efficiency (TCE), highlights the relation between energy consumption and CO2 emissions. In practice, TCE values are formulated by multiplying existing PUE values with carbon emission rates, which depend on energy sources, such as coal, nuclear, and water. Carbon emission rates vary between 0.1 and 2.3 pounds of carbon dioxide per kilowatt-hour. TCE, however, ignores other environmental impacts caused by services, such as the disposal of equipment [7]. Compute Power Efficiency (CPE) metric measures the computation efficiency of data centers. Its value is defined by dividing IT energy consumption with total energy consumption and then multiplied with average IT equipment utilization. It can also be calculated by dividing the average IT utilization with a previously measured PUE [20]. Transparent and uniform metrics enforce derived demand, because they enable end users to compare different web services and select more green alternatives. Efficiency metrics (table 3) and other certificates should be clarified, illustrated, and promoted for end users in order to have desirable impact on web service providers' and hosting service providers' sales and reputations.

Table 3: Efficiency metrics for green SLAs

Full name	Abbr.	Description	Formula
Power Usage Effectiveness	PUE	Illustrates the energy efficiency of data center infrastructure	$PUE = \dfrac{\sum P}{P_{IT}}$
Energy Reuse Effectiveness	ERE	Presents the utilizations of waste heat within data centers	$ERE = \dfrac{\sum P - P_{RE}}{P_{IT}}$
Technology Carbon Efficiency	TCE	Indicates the link between used energy and emissions	$TCE = \dfrac{\sum P * CO_2}{P_{IT}}$
Compute Power Efficiency	CPE	Emphasizes the computation efficiency inside data centers	$CPE = \dfrac{P_{IT} * IT}{\sum P}$

Because of outdoor conditions, the electricity consumption of cooling systems constantly varies, which is why the values of efficiency metrics are not static. Furthermore, different server workloads also affect the overall electricity consumption of web services. Due to the changing conditions, green SLAs need to have specific ranges for metrics instead of single figures. The proposed metrics have a few identical variables and therefore all of their results vary when outdoor conditions and workloads change. In addition to acceptable ranges for metrics with relative results, green SLAs should also guarantee absolute limits for certain important variables. For example, service providers should ensure that a specific minimum quantity of waste heat will be utilized in all circumstances. Moreover, service providers should state that their CO2 emissions will never exceed a certain maximum quantity. Some providers might even ensure that their IT resources will have high utilization rates, meaning that they do not waste electricity and acquire unnecessary devices. Properly utilized resources were not self-evident even in 2012 [14].

Measuring the accurate amount of energy, waste heat, and IT utilization in data centers could be challenging in practice. To calculate exact results, service providers need to measure their electricity consumption in various places, including near cooling systems, network switches, and rack servers. They also need to know where their electricity is produced and how much CO2 emissions it causes. Furthermore, in many countries kilograms need to be converted into pounds to use globally uniform TCE values. However, if service providers are unable to locate the correct power plants or exact CO2 rates, pre-defined rates for different energy sources can be used instead [7]. Determining the average utilization of all IT equipment could be difficult and therefore typically only the utilization of processor units is measured. Moreover, high utilization rates do not necessarily mean that services are energy efficient. Poorly coded programs; for instance, could drain all the available resources, but perform very little useful work. Determining the accurate IT asset utilization is one of the major challenges of modern energy efficiency metrics [14].

3.5.5 Incentives

The case studies reveal that currently, the only incentives for service providers to enhance their greenness are lower operational costs and improved competitiveness. However, if Gartner's predictions become a reality (Site 1), there will be demand for more green IT services and additional incentives for service providers (table 4). It requires that cloud and hosting service providers will draw up green SLAs and promote their greenness with transparent metrics and certificates, which offer credibility for services. Then they need to compete with greenness and sign contracts with web service providers who then need to illustrate their greenness to current and upcoming end users and try to differentiate themselves from other web service providers. Furthermore, regular end users need to comprehend that every action in the Internet requires electricity. For example, a single Google search uses 0.3 watts of electricity and Google's daily power draw accounts for 13 million watts (Site 3). Perhaps Google should illustrate this information on search results pages. Hopefully in the future, some end users would even be willing to pay more for services that are proven to be green by favoring; for example, web stores with slightly higher prices, but significantly lower emissions.

The case studies indicate that the relationship between service providers and local energy companies needs to be further developed. Nowadays, service providers do not benefit enough from the waste heat they provide for energy companies. It is of mutual interest that service providers would have more affordable electricity or district cooling in exchange for waste heat. It would also add another incentive for service providers to construct green data centers. After all, the waste heat of data centers replaces other energy sources, which would pollute the climate. At least in the European Union, it is advantageous for energy companies to minimize their CO_2 emissions (Site 2). Even though local energy companies certainly have a monopoly over district cooling and heating, unsatisfied service providers can still purchase their electricity elsewhere. Green IT services could be especially beneficial for energy companies, if service providers start to acquire more expensive electricity, which is produced from renewable sources.

Table 4: Incentives for different actors

Actors	Cloud/Hosting →	Web services →	End users
Lower opera-tional costs	Improved profit-ability or better competitiveness	No effect or lower list prices for cloud services	No effect or more affordable or free web services
Actors	**End users →**	**Web services →**	**Cloud/Hosting**
Greenness as a value	Interest to use more environ-mentally friendly web services	Need to improve competitiveness by emphasizing green features	Need to further improve energy efficiency and to reduce emissions
Actors	**Cloud/Hosting →**	**Web services →**	**End users**
Achieved greenness	Need to prove and guarantee green-ness (green SLA)	Need to promote and illustrate greenness (ads)	Values must be transformed into actions (use them)
Actors	**End users →**	**Web services →**	**Cloud/Hosting**
Excessive demand	Notable need for green services, long-term trend	Opportunity to increase prices & market shares	Opportunity to increase prices & market shares
Actors	**End users →**	**Cloud/Hosting →**	**Energy CO**
Green energy	Interest towards web services with zero emissions	Need to purchase green power for hosted services	Demand for re-newable energy, increased profits
Actors	**Energy CO →**	**Cloud/Hosting →**	**Web services**
Mutual interest	Need to ensure that both elect-ricity and district cooling is bought from their plants	Need to gain extra benefits when offering waste heat from the data centers	Need for informa-tion, the quantity of waste heat, the amount of CO_2 reduction

4   Conclusions

In the best case scenario, well-optimized web services are hosted on highly utilized servers that are located in green data centers. Due to their nature, cloud services should be able to fully exploit their resources, because they support numerous simultaneous end users and multi-tenancy, which then provide economies of scale [1]. Major public cloud services definitely have notable scale and they are a great option for many small and medium companies, who want to host their services on fully outsourced infrastructures. Public cloud services, however, do not suit all companies because of privacy and security issues [4]. They could also be expensive in the long term, if significant amounts of data need to be transferred between services and end users [21]. Furthermore, the amount of data and the distances between different parties mainly constrain the greenness of cloud services [5]. As a consequence, large companies should consider utilizing private cloud services near their employees and most of their business customers, even if private clouds have significant upfront costs. In the simplest arrangement a private cloud is managed and hosted by the company itself [2]. However, if large companies want to reduce some of their upfront costs, they can rely on third parties, which then host private clouds in their novel data center infrastructures [18].

This paper presented a multiple-case-study that was used to collect information about state-of-the-art data centers, and consequently to provide new knowledge about the incentives of green cloud computing and green hosting [24]. The research framework was drawn up based on the relevant literature. Even though the research had its limitations, it provided a few conclusions. Server utilization in cloud services can be high, because several instances are hosted on the same hardware. Then it is up to business customers to fully utilize their instances and further increase the utilization. Cooling systems in cloud services can be significant energy consumers, and therefore service providers should apply energy-efficient methods, such as liquid cooling, to shift heat in their facilities and use energy-saving solutions, such as free cooling, to cool the facilities. Nowadays data centers should be able to utilize their waste heat in order to enhance the greenness of cloud services. Even though greenness is an important feature for a service, it should not compromise the quality of a service. Service providers should be very careful, when they try new energy-

saving solutions; for example, software that automatically powers down servers or network switches. To guarantee greenness and other features in practice, cloud and hosting service providers need to draw up green SLAs, which determine all the relevant efficiency metrics and acceptable ranges for them.

The literature implies and the research confirms that the main incentive to improve energy efficiency and to reduce electricity consumption is lower operational costs [1], [5]-[6], [8]. However, there could be more incentives, if the end users of web services started to demand more green services. It would also create derived demand for green cloud services and green hosting services, and it would add extra value to them; for example, better reputations, improved marginal profits and possible larger market shares. Green services would be beneficial for energy companies as well, because they would have more market for renewable energy and energy-efficient district cooling systems.

Websites List

Site 1: Gartner press release, CO2 estimates
http://www.gartner.com/it/page.jsp?id=503867

Site 2: Emissions trading in European Union
http://www.energiamarkkinavirasto.fi/alasivu.asp?gid=199&languageid=826

Site 3: New York Times, Google Details, and Defends, Its Use of Electricity
http://www.nytimes.com/2011/09/09/technology/google-details-and-defends-its-use-of-electricity.html?_r=0

References

[1]   D. Albano, D. Abood, A. Armstrong, R. Murdoch and J. Whitney, (2010, November) Cloud computing and sustainability: The environmental benefits of moving to the cloud. Accenture. [Online]. Available: http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture_Sustainability_Cloud_Computing_TheEnvironmentalBenefitsofMovingtotheCloud.pdf.

[2]   M. Armbrust, A.Fox, R. Griffith, and A.D. Joseph, Above the clouds: A Berkeley view of cloud computing, EECS Department, University of California, Berkeley, CA, Technical Rep. UCB/EECS-2009-28, 2009.

[3]   Azevedo, M.Blackburn, J. Cooley and M. Patterson. (2011, March) Data Center Efficiency Metrics. The Green Grid. [Online]. Available: http://www.thegreengrid.org/~/media/TechForumPresentations2011/Data_Center_Efficiency_Metrics_2011.pdf.

[4]   Balding, J. Howie and D. Hurst. (2010, March) Cloud security alliance. Top Threats to Cloud Computing V1.0. [Online]. Available: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf.

[5]   J. Baliga, R.W.A. Ayre, K. Hinton and R.S. Tucker, Green cloud computing: Balancing energy in processing, storage, and transport, in Proceedings of the IEEE, Melbourne, 2011, pp. 149-167.

[6]   M.T. Chaudhry, T.C. Ling, and A. Manzoor, Considering thermal-aware proactive and reactive scheduling and cooling for green data-centers, in Proceedings 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT), Kuala Lumpur, 2012, pp. 87-91.

[7]   Cook. (2007, September) Technology carbon efficiency - A metric for measuring data center green impact. S3 Amazonas. [Online]. Available: http://s3.amazonaws.com/zanran_storage/www.datacenterdynamics.com/ContentPages/22802800.pdf.

[8]   Greenberg, J. Hamilton, D. Maltz and P. Patel, The cost of a cloud: Research problems in data center networks, ACM SIGCOMM CCR, vol. 39, no. 1, pp. 68-73, 2009.

[9]   R. Harms and M. Yamartino. (2010, November) The economics of the cloud. Microsoft Press Release. [Online]. Available: http://www.microsoft.com/en-us/news/presskits/cloud/docs/the-economics-of-the-cloud.pdf.

[10]   Heller, P. Mahadevan, and S. Seetharaman, ElasticTree: Saving energy in data center networks, NSDI'10 in Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation, San José, CA, 2010, pp. 249-264.

[11]   S. Hirsjarvi, Tutkimushaastattelu - Teemahaastattelun teoria ja kaytanto, Helsinki, Gaudeamus, 2010, pp. 34-40.

[12]   Huan, A Measurement study of server utilization in public clouds, in Proceedings IEEE the 9th International Conference, on Dependable, Autonomic and Secure Computing (DASC), Sidney, 2011, pp. 435-442.

[13]   Jarschel and R. Pries, An OpenFlow-based energy-efficient data center approach, in Proceedings of the ACM SIGCOMM Conference on the Applications, Technologies, Architectures, and Protocols for Computer Communication, Helsinki, 2012, pp. 87-88.

[14]   G.B. Kenneth. (2010, August) Revolutionizing Data Center Efficiency. Uptime Institute. [Online]. Available: http://uptimeinstitute.com/component/docman/doc_download/48-update-uptime-institute-revolutionizing-data-center-efficiency.

[15]   G.V. Laszewski and L. Wang, Green IT service level agreements, In Grids and Service-Oriented Architectures for Service Level Agreements (P. Wieder et al, Eds.). New York, NY: Springer, 2009, pp- 77-88.

[16]   Li, X. Yang, and M. Zhang, Cloud-Cmp: Comparing public cloud providers, in Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10), New York, 2010, pp. 1-14.

[17] J. Liu, F. Zhao, X. Liu and W. He, Challenges towards elastic power management in internet data centers, in Proceedings of the 29th IEEE International Conference, on Distributed Computing Systems Workshops, (ICDCS Workshops) Stevenson, 2009, pp.65-72.

[18] Peters. (2012, November) IBM cloud computing. Presentation of Cloud Computing Concepts. [Online]. Available: https://www.privacyassociation.org/media/presentations/12DPC/DPC12_Cloud_PPT3.pdf.

[19] Rubenstein, R. Zeighami, R, Lankston and E. Peterson, Hybrid cooled data center using above ambient liquid cooling, in Proceedings 12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), Las Vegas, 2010, pp. 2-5.

[20] J. Rylands. (2010, June) Energy Efficiency in the Datacentre. Questnet. [Online]. Available: https://www.questnet.edu.au/download/attachments/8519738/Jason_Rylands.pdf.

[21] Spellman and R. Gimarc, Leveraging the cloud for green IT: Predicting the energy, cost and performance of cloud computing, in Proceedings of CMG International Conference, Dallas, 2009, Paper 9138.

[22] Uchechukwu, K. Li, and Y. Shen, Improving cloud computing energy efficiency, in Proceedings of the Cloud Computing Congress (APCloudCC), IEEE Asia, 2012, pp. 53-58.

[23] A.T. Velte, T.J, Velte and R. Elsenpeter, Cloud Computing, a Practical Approach, New York: McGraw-Hill, 2010

[24] R. Yin, Case Study Research. Design and Methods. California: Sage Publications. Thousand Oaks, 2003.

Incentives to Apply Green Cloud Computing

Tommi Makela
Sakari Luukkainen