Case Studies for achieving a Return on Investment with a Hardware Refresh in Organizations with Small Data Centers

Joseph Doyle and Rabih Bashroush, *Member, IEEE,*

Abstract—Data centers have been highlighted as a major energy consumer and there has been an increasing trend towards the consolidation of smaller data centers into larger facilities. Yet, small data centers exist for a variety of reasons and account for a significant portion of the total number of servers. Frequent refreshes of IT hardware have emerged as a trend in hyper-scale data centers but little attention has been paid to how these savings can be achieved in smaller facilities. This work provides a comprehensive framework for the energy saving opportunities while determining when a return on investment can be achieved to enable small data center operators to create credible business cases for hardware refreshes. Various data center deployment scenarios are used as case studies (based on real-life datasets) to validate the proposed concepts. Our results show that a return on investment can be achieved for organizations with small data center in less than two years and that these organizations can save more than three times their annual electricity cost over five years.

Index Terms—Data centers, energy efficiency, hardware refresh rate, environmental impact.

1 Introduction

Data centers and the services they provide support a considerable portion of applications which are used by the world population every day. The digital economy is executed upon the infrastructure of data centers which contain servers as well as facilities for connectivity hubs, power distribution and physical security. The digital economy is entering a new age with an explosion in data, fed by the growth of paradigms such as the Internet of Things, Cloud, and Smart Cities, paired with improved connectivity (5G is beginning its rollout). This has resulted in an extraordinary increase in the demand for new applications as well as the infrastructure to support them.

New challenges have emerged as a result of these demands. Data centers have been highlighted as a significant consumer of electricity worldwide and the environmental impact of this also been considered [1]. The energy consumption of data centers in Western Europe was estimated to be 86 TWh in 2013 (3% of annual electricity consumption of Western Europe) and was projected to increase to 104 TWh by 2020 [2]. This trend has not been predicted in every part of the world [3].

Over the last decade, considerable work has been done to propose methods for reducing the energy consumption of data centers at the hyper-scale [4], [5], [6] but little attention has been paid to the energy consumption of smaller data centers. Small data centers are defined as data centers where servers are housed in small areas of less than 1,000 square foot and approximately 40% of servers in the US are housed in these data centers [7]. While there are other definitions of small data centers [8] they tend to be focused on supporting specific paradigms such as edge computing and their prevalence worldwide cannot be easily determined. Considerable energy savings as well as a corresponding reduction in the environmental impact of these facilities can be achieved. Over the last decade with the emergence of metrics such as the Power Usage Effectiveness (PUE - the ratio of the total facility load divided by the IT load) [9], tangible progress has been made in increasing the effectiveness of the data center cooling and power infrastructure, with PUE values of near 1.01 reported in some cases [10]. PUE in small data centers is typically high due to economies of scale and the level of investment required to achieve a lower PUE [11]. Lowering PUE, however, is not the only path to reducing the overall energy consumption of a data center. By examining the frequency of useful workload processing and using this to appropriately provision server resources during hardware refreshes energy savings can be achieved without large investments [12]. In this study, we scope hardware refreshes to the replacement of entire servers rather than upgrading components of the server or the related power provisioning system. Circular economy based analysis (remanufacturing, reconfiguration, component refresh, etc.) is the subject of a separate study. Utilization in data centers is typically low, with data centers reported as ranging from 6% to 12% [13] so the consolidation of workloads onto a smaller number of servers can also be used to reduce operational expenditure. Thus, the power provisioning system is likely to be sufficient for the needs of the new servers. In addition, the lifetime of power systems is typically five times that of servers [14]. Thus, its replacement is unlikely to offer a quick ROI and it is not considered in this work.

Part of the reason that the energy efficiency of small data centers lags behind hyper-scale data centers is that invest-

- *J. Doyle was with the School of Electrical Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom.*
  E-mail: j.doyle@qmul.ac.uk
- *R. Bashroush was with the School of Architecture, Computing, & Engineering,University of East London, London E16 2RD, United Kingdom.*

ment is limited. Thus, it is difficult to convince organization management teams to authorize hardware refreshes without a concrete business plan which indicates when an ROI will be achieved. This work provides a framework which organizations with small data centers can use to calculate when an ROI will be achieved if a hardware refresh is executed. In this work, three case studies based upon data from organizations currently operating small data centers are discussed. Data on the server population of the organizations as well as details on supporting equipment is presented and used to validate the calculations on how quickly an ROI can be achieved as well as the savings that can be achieved over five years. We make three contributions, which are summarized below:

- We present a framework for determining when a small organization will achieve an ROI after a hardware refresh including details such as the deployment cost and the downtime cost.
- We detail the hardware profiles of three organizations with small data centers as well as the utilization levels. We analyze these details to provide insights into how the hardware profiles of small data centers differ from hyper-scale data centers.
- We examine how quickly a hardware refresh will result in an ROI for these organizations. Our results show that an ROI can be achieved in less than two years and can save over three times their current annual electricity cost after five years.

The next section discusses related work. Section 3 describes the methodology used to calculate energy savings due to hardware refresh, factoring in the reduction of energy costs due to improved server efficiency, the age of the servers and the cost of electricity. In Section 4, additional factors to the cost of procurement are discussed as they can delay the ROI after a hardware refresh. The case studies for the three organizations with small data centers are discussed in section 5. Section 6 highlights the study limitations. Section 7 concludes the work.

2 Related Work

2.1 Server Refresh Cycles

Recently, there have been a number of works on hardware refreshes in data centers. Bashroush presents a comprehensive framework for examining how long it takes a hardware refresh to reduce the environmental impact of a data center as well as detailing the life cycle impact assessment of hardware refresh scenarios [15]. Wang *et al.* examine the failures rates of various hardware components in data centers and how this relates to the life cycle of these components [16]. They also examine the effectiveness of repairs and show that in a significant number of cases repeated failures occur. Alter *et al.* investigate the failure characteristics of 30,000 SSD failures in a Google data center over the course of six years and how this relates to the life cycle of the drives [17]. These works, however, focus on larger data centers while our work focuses on the savings which can be made at more modest installations.

2.2 Modular Data Centre Design

There have been numerous works on the design of modular data centers which are small data centers contained in shipping containers and small data centers which support edge computing. Vishwanath *et al.* present a model for the performance, reliability and cost of modular data center solutions [18]. Nikolaou *et al.* investigate the total cost of ownership (TCO) of micro-datacenters at the edge and demonstrate that the edge based solution can have a significantly lower TCO when compared to a cloud based solution [19]. Qouneh *et al.* examine the cooling costs of container-based data centers and show that considerable savings can be made when compared to raised floor data centers [20]. Khalid *et al.* examine steady-state energy and exergy destruction models for modular data centers to demonstrate that augmenting direct expansion cooling in modular data centers can result in significant energy savings in hot and arid climates [21]. These solutions, however, only consider specific scenarios such as edge computing or attempts to lower maintenance costs via modular data centers. In our work, we have analyzed more general small data centers in small organizations [7] to demonstrate that ROIs can be achieved in relatively short periods for this part of the data center market.

2.3 Energy Efficiency in Data Centers

There has been considerable research in reducing the energy consumption of data centers. This can be achieved through various means. Firstly, load balancing can be utilized to direct load to data centers which use more renewable energy. Laganà *et al.* propose using a hierarchical management architecture to effectively utilize renewable energy [22]. Guo *et al.* propose using an algorithm based upon Lyapunov optimization to reduce the energy consumption of colocation data centers while load balancing to minimize interference on active workloads [23].

Secondly, consolidation can be used to reduce the number of physical machines which host virtual machines and containers operating in a data center. Qiu *et al.* propose using a probabilistic demand allocation problem to minimize the number of physical servers required to service a workload without violating service level agreements [24]. Farahnakian *et al.* propose using a regression-based model to approximate future CPU and memory usage to improve the performance of consolidation using a vector bin-packing algorithm. Xu *et al.* propose selectively deactivating containers when a data center is overloaded to reduce the number of under utilized physical servers in data centers [25].

Finally, the energy efficiency of variations of the cloud computing paradigm such as fog computing has been examined. Xiao *et al.* propose a system where fog nodes cooperate to achieve the optimal performance of the trade-off between service response time and energy consumption [26]. Wang *et al.* propose an energy aware data collection algorithms to reduce the energy consumption in the Internet of Things platforms by reducing redundant data collection [27].

3 Background

In this section we discuss the methodology used to identify the point at which an ROI is achieved. This is similar to the

methodology used in [15]. This analysis, however, investigates the point at which monetary savings are achieved via a hardware refresh while the analysis in [15] examines the point at which energy savings are achieved. To identify the optimal time interval (if it exists) for hardware refresh rates, consider Figure 1 below. The diagram shows the cumulative cost of two scenarios. The first scenario (blue line) shows the energy cost of existing hardware. The second scenario (orange line) shows the cost of replacing the hardware and the cumulative cost of refreshed hardware at a point n in time.

In this scenario, the embodied cost (cost of purchasing and installing the hardware) of current hardware (in \$), introduced at time $t = 0$, is designated as E_e^c ; and the annual usage cost (in \$) as E_u^c. Similarly, for the refreshed hardware, introduced at time $t = n$, E_e^r and E_u^r represent the embodied and annual usage cost, respectively. Hardware disposal cost (end of life) is very small compared to embodied and usage cost (see section 6) and was not included in the model. Additionally, when the hardware is recycled or redeployed, disposal cost is considered a net saving. The point of intersection between the two scenarios (represented as $n + \mu$ in Figure 1) is calculated. The intersection point, if it exists, represents the point in time where the refreshed hardware and existing hardware would have the same cumulative cost. Beyond that point, savings will be accrued from a reduction in energy consumption (compared to no hardware refresh). To find the intersection point, the equations of the two scenario lines need to be calculated. For current hardware, the cumulative energy line (blue) passes through two points with coordinates $(0, E_e^c)$ and $(1, E_e^c + E_u^c)$ as shown in Figure 1. Accordingly, the equation of the line can be calculated as:

$$E^c(t) = E_u^c t + E_e^c \tag{1}$$

where $E^c(t)$ is the cumulative cost (in \$) for the current hardware at a point in time $t \geq 0$.

Similarly, we calculate the equation of the refreshed hardware cumulative energy line (orange), which passes through the two points $(n, E_e^c + E_e^r + nE_u^c)$ and $(n+1, E_e^c + E_e^r + nE_u^c + E_u^r)$, to be:

$$E^r(t) = E_u^r t + n(E_u^c - E_u^r) + E_e^c + E_(e)^r \tag{2}$$

where $E^r(t)$ is the cumulative energy consumption for the refreshed hardware (in \$) at a point in time $t \geq n$, and n is the refreshed hardware introduction time (in years). Given the above two equations (1) & (2), the intersection point of the two lines can be calculated by setting $E^c(t) = E^r(t)$, which gives:

$$E_u^r t + n(E_u^c - E_u^r) + E_e^c + E_e^r = E_u^c t + E_e^c \tag{3}$$

solving for t, the intersection payback time, τ, is found to be:

$$\tau = n + \frac{E_e^r}{E_u^c - E_u^r} \tag{4}$$

Expressing the intersection time τ in terms of n (refreshed hardware introduction time):

$$\tau = n + \mu \tag{5}$$

where μ is the time interval after which efficiency is achieved by the newly refreshed hardware (payback time). Then, replacing (5) in (4) produces:

$$\mu = \frac{E_e^r}{E_u^c - E_u^r} \tag{6}$$

Equation (6) shows the length of time needed to start reducing overall costs after a hardware refresh (factoring in embodied cost). The first observation in (6) is that it makes no reference to the embodied cost of existing hardware E_e^c. Thus, deciding on an optimal hardware refresh rate does not require knowledge of the cost of existing hardware as currently widely believed. Indeed, this is referred to as the sunk cost fallacy in behavioral economics [28]. Similarly, savings, ζ, due to hardware refresh can be calculated based on the difference between $E^c(t)$ and $E^r(t)$ over a time period $\sigma \geq \tau$, from equations (1) and (2), as:

$$\zeta_\sigma = E^r(t) - E^c(t) = (E_u^c - E_u^r)(\sigma - n) - E_e^r \tag{7}$$

As with equation (6), equation (7) shows that energy savings are also independent of the embodied energy of current hardware E_e^c.

One of the key variables in this equation E_r^c can be calculated using the cost of replacing the servers. This is depicted in equation (8) where i is an index for the replacement servers, s_i is the cost of the replacement servers and I is the number of servers being replaced.

$$E_e^r = \sum_{i=0}^{I} s_i \tag{8}$$

The difference in energy consumption $(E_u^c - E_u^r)$ can be estimated based upon the age and power rating of the server as well as the the cost of electricity.

$$(E_u^c - E_u^r) = \sum_{i=0}^{I} a_i r_i p_i l_i \tag{9}$$

This is depicted in equation (9) where a_i is the age of the server, p_i is the power rating of the server, l_i is the cost of supplying electricity to the server and r_i is the performance per watt improvement slope based upon chip improvements. This was found to be 0.03048 as shown in Section 5.1.

$$\zeta_\sigma = (\sigma - n) \sum_{i=0}^{I} a_i r_i p_i l_i - \sum_{i=0}^{I} s_i \tag{10}$$

Using equations (8,9) we can reformat ζ_σ in more concrete terms. This is depicted in equation (10)

$$\zeta_\sigma = (\sigma - n) \sum_{i=0}^{I} a_i r_i p_i l_i - \sum_{i=0}^{I} (s_i + d_i) \tag{11}$$

There are also additional costs which are discussed in the next section which will affect the calculations of ζ_σ which can be incorporated into the equation. This is depicted in equation 11 where d_i is additional costs associated with the server's replacement such as deployment costs.

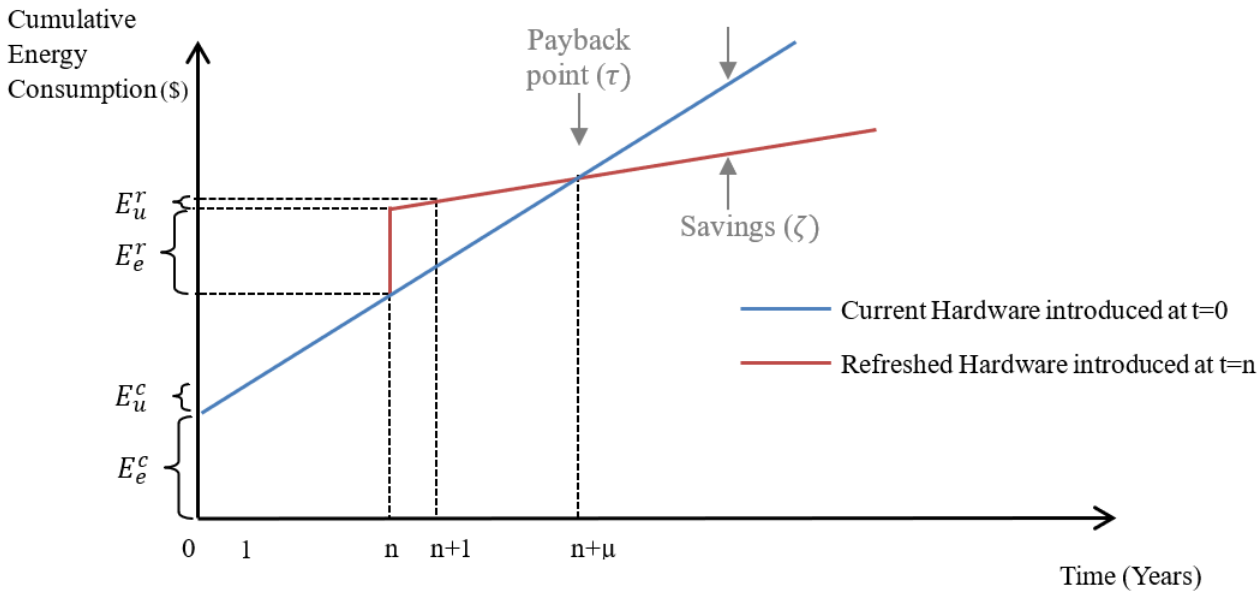

Fig. 1: Cumulative energy consumption of existing hardware vs refreshed hardware

4 Cost of Procurement

In order to determine the optimal time to replace servers, it is important to factor all aspects of server procurement into cost calculations. Beyond the cost of the hardware it is also important to consider the following:

- Deployment cost (E_d^r). This includes the physical deployment of the server as well as the costs associated with server, network and storage configuration. Data center power and cooling must also be considered as well as other system administration tasks.
- Downtime cost (E_o^r). This includes the costs associated with planned and unplanned downtime. This calculation includes the cost of restoring services, lost employee productivity and lost revenue.
- Business Administration (E_b^r). This includes the labor cost associated with creating orders, obtaining purchase approvals, negotiating vendor contracts and tracking the procurement process.

The deployment cost varies considerably as it depends on the hourly labor cost as well as the data center design. It is, however, closely related to the hardware cost (E_h^r) as the time required to install and configure hardware is dependent on how much hardware is being installed and configured. Previous case studies have found that the deployment cost ranges from 4% to 14% of the hardware cost [29]. Thus, in our analysis, we assume that the deployment cost is 10% of the hardware cost ($E_d^r = 0.1E_h^r$).

The Downtime cost also varies considerably as it depends on how much downtime is required for the installation and configuration of the servers. Previous case studies have found that the downtime for the installation of between fifty and a hundred servers can range from twelve hours to no downtime at all [29]. This will depend on the data center design. For example, if the hardware is being installed in a new data center with the current installation set to be decommissioned when the new data center comes online, there will be no downtime. The cost of downtime is dependent on the labor costs associated with fixing the problem and the revenue that is lost while the service is down [30]. As an accurate estimate is not possible without knowing the revenue of the service, we do not consider this cost in our analysis. It can, however, be a significant cost which will be considered in future work.

Business administration costs are also difficult to calculate. Labor is usually the largest cost in small data centers [31]. It is difficult, however, to determine how much of this cost is associated with business administration for procurement. Thus, we assume a conservative estimate that the business administration cost is the same as the deployment cost ($E_b^r = E_d^r = 0.1E_h^r$). It may be considerably more than this but a more accurate measure of this cost will be considered in future work.

5 Case Studies

In this section, various case studies representing real-life data center scenarios are evaluated. In this section, we present data on the computational capacity and energy usage of three organizations with small data centers. A small data center is defined as data centers where servers are housed in small areas of less than 1,000 square foot [7]. They can also be classified by the number of servers with a small data center classified as having less than twenty five servers [7]. As significantly more than 25 servers can be housed in 1,000 square foot if racks are used we have chosen to use this definition as it accurately reflects the data centers studied. While the individual energy consumption of these data centers is small, approximately 40% of servers in the US are housed in these data centers [7]. Thus, improvements in energy efficiency in these data centers will have a profound effect on the overall energy consumption of data centers. We also present information on the energy usage of supporting hardware such as CRAC units and UPS. We then examine

the cost of updating their hardware, the corresponding energy consumption reduction from utilizing more modern server designs and finally the time required for the savings from the energy consumption reduction to exceed the cost of updating the servers.

5.1 Hardware Profile

The hardware profile of three small organizations is depicted in Table 1. A number of interesting conclusions can be drawn from these profiles. Firstly, the hardware is quite heterogeneous. There are seventeen server types among the one hundred and thirty two servers deployed in the three data centers and only one server type is used in more than one data center. This trend has been found to occur in larger cloud organizations [32]. One of the proposed causes of this heterogeneity is asynchronous upgrades further exacerbating the inherent diversity of computational requirements found in organizations [33]. Asynchronous upgrades are a particular problem in small organizations as budgets are frequently limited and full hardware refreshes are not possible without a significant business case.

Secondly, the utilization rates of the servers which range from 20% to 40% are above the average CPU utilization levels in data centers reported as ranging from 6% to 12% [13]. They are, however, still sufficiently low that the organizations could benefit from consolidation [3]. Thirdly, the energy proportionality [34] of the servers varies hugely. This is to be expected due to the heterogeneity of the hardware. More modern servers use less power to complete the same computations. The performance per watt of similar Intel Core i7 processors is depicted in Table 2. From the table, we can see that the performance per watt has gradually increased since 2011. This is further depicted in Figure 2. It should be noted that there is not enough data to infer a specific rate for the growth of performance per watt. The data, however, is sufficient to illustrate the increased performance per watt which has resulted from increased transistor density. It should also be noted that the Intel Core i7 brand includes lower power and ultra low power processors which are deliberately underclocked to reduce their power consumption. These processors, however, are rarely used in data center servers as they are mostly used to increase battery life in mobile devices. Finally, recent work has shown that other processor designs such as ARM use less power than Intel architectures and this can also affect the energy proportionality of the servers [35]. These processors, however, also tend to be used solely in mobile devices.

Finally, we can see that the servers in these data centers are quite old. The age is an estimate based upon the date when this data was collated in 2017 and the release date of the server type. We can, however, conclude that the servers are considerably older than those found in larger cloud providers where the life cycle of a server is typically 3-5 years [36]. The average age of the servers in these organizations is 9 years. Even assuming an error in the age estimate of 2 years they are considerably older than the maximum age of servers found in large cloud providers. This illustrates that small organizations are particularly suitable for hardware refreshes to reduce energy consumption for

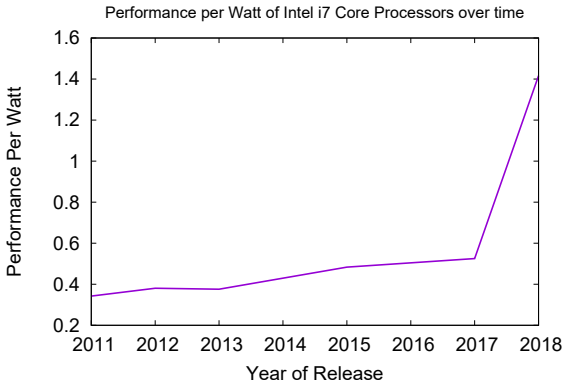

Fig. 2: Performance per watt for core i7 Intel Processor.

two reasons. Firstly, as the life cycle of a server is extended beyond what is typical in other organizations greater energy savings can be made as the data center will be operating with reduced energy consumption for a longer time before another hardware refresh is required. Secondly, greater consolidation of servers is possible as newer processors will be considerably more powerful allowing a single server to perform the work of several older servers.

5.2 Supporting Hardware Profile

In addition to energy savings which can be made by consolidating the number of servers operating, further savings can be made by considering the supporting hardware. This can also be consolidated to reflect the lower power consumption of the servers. The supporting hardware profile of the three organizations is depicted in Table 3. A number of interesting conclusions can be drawn from these profiles. Firstly, the supporting hardware is quite heterogeneous. While there are not as many different types due to the lower overall number of supporting devices there are four different computer room air conditioner (CRAC) types of the twenty four CRACs used and there are six different uninterrupted power supply (UPS) designs of the forty four UPSs used. This will occur partly for the same reason it occurs in servers, namely, asynchronous updates [33]. Another reason for this is the deployment of the UPS system. It is common to utilize two UPS devices to power servers to make servers more resilient to the failure of a UPS [38]. Based upon the supporting devices' power rating and the sum of the power ratings of the servers it appears that Organization 1 has taken this approach. It appears, however, that Organization 2 has taken the simpler approach of connecting all servers in the data center to a single UPS. Thus, this design strategy requires a UPS with a high power rating which further exacerbates the heterogeneity.

Secondly, if we compare the power rating of the CRAC devices, UPS devices and servers we can see that in the majority of cases that these organizations are over provisioned with cooling and UPS devices. This is depicted in Figure 3. For UPS devices this can partially be explained by typical UPS deployments to make servers resilient in the event of

Quantity	Server Type	100% utilized Power Rating	Idle Power Rating	Utilization (0-1)	Idle Power/100% utilized Power Rating as Percentage	Estimated Age of Server(years)
11	Server Type 1	301	125	0.2	41.5%	7
17	Server Type 2	258	172	0.4	68.3%	12
8	Server Type 3	258	172	0.4	68.3%	12

(a) Hardware Profile of Organization 1

Quantity	Server Type	100% utilized Power Rating	Idle Power Rating	Utilization (0-1)	Idle Power/100% utilized Power Rating as Percentage	Estimated Age of Server(years)
13	Server Type 4	237	75.6	0.4	31.9%	8
7	Server Type 5	258	172	0.4	68.3%	11
14	Server Type 6	244	143	0.4	58.6%	11
10	Server Type 7	117	75	0.4	64.1%	7
3	Server Type 8	258	172	0.4	68.3%	12
5	Server Type 9	258	172	0.4	68.3%	9
2	Server Type 10	258	172	0.4	68.3%	15
6	Server Type 11	258	172	0.4	68.3%	4
4	Server Type 12	258	172	0.4	68.3%	1

(b) Hardware Profile of Organization 2

Quantity	Server Type	100% utilized Power Rating	Idle Power Rating	Utilization (0-1)	Idle Power/100% utilized Power Rating as Percentage	Estimated Age of Server(years)
10	Server Type 4	237	75.6	0.3	31.9%	8
6	Server Type 13	263	53.3	0.3	20.3%	5
6	Server Type 14	117	75	0.4	64.1%	7
2	Server Type 15	117	75	0.4	64.1%	5
5	Server Type 16	276	157	0.4	56.8%	11
3	Server Type 17	266	56.9	0.4	21.4%	7

(c) Hardware Profile of Organization 3

TABLE 1: Hardware Profile of Organizations.

Year of Release	Processor Type	Thermal Design Power (W)	GFLOPs	GFLOPs per Watt
2011	i7-2600	95	32.52	0.3423
2012	i7-3770	77	29.29	0.3804
2013	i7-4770	84	31.57	0.3758
2015	i7-6700	65	31.42	0.4834
2017	i7-7700	65	34.14	0.5252
2018	i7-8700	65	92.07	1.4165

TABLE 2: Performance per watt for core i7 Intel Processor. The figures for GFLOPs are taken from [37]

Quantity	Device Type	Power Rating	Efficiency	Utilization (0-1)
7	CRAC Type 1	5500	1	1
5	CRAC Type 2	5500	1	1
8	UPS Type 1	2700	0.93	0.5
4	UPS Type 2	4200	0.93	0.5
7	UPS Type 3	1920	0.93	0.5
9	UPS Type 4	700	0.93	0.5

(a) Supporting Hardware Profile of Organization 1

Quantity	Device Type	Power Rating	Efficiency	Utilization (0-1)
8	CRAC Type 3	7100	1	1
4	UPS Type 5	12000	0.93	0.5

(b) Supporting Hardware Profile of Organization 2

Quantity	Device Type	Power Rating	Efficiency	Utilization (0-1)
4	CRAC Type 4	2800	1	1
3	UPS Type 1	2700	0.93	0.5
5	UPS Type 3	1920	0.93	0.5
4	UPS Type 6	8000	0.93	0.5

(c) Supporting Hardware Profile of Organization 3

TABLE 3: Supporting Hardware Profile of Organizations.

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
11	Server Type 1 Replacement	5021	55231
17	Server Type 2 Replacement	2308	39236
8	Server Type 3 Replacement	1429	11432

(a) Replacement Hardware Cost Organization 1

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
13	Server Type 4 Replacement	2241	29133
7	Server Type 5 Replacement	2325	16275
14	Server Type 6 Replacement	897	12558
10	Server Type 7 Replacement	1142	11420
3	Server Type 8 Replacement	1670	5010
5	Server Type 9 Replacement	2308	11540
2	Server Type 10 Replacement	2308	4616
6	Server Type 11 Replacement	825	4950
4	Server Type 12 Replacement	2325	9300

(b) Replacement Hardware Cost for Organization 2

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
10	Server Type 4 Replacement	2241	22410
6	Server Type 13 Replacement	1386	8316
6	Server Type 14 Replacement	897	5382
2	Server Type 15 Replacement	1142	2284
5	Server Type 16 Replacement	1429	7145
3	Server Type 17 Replacement	1429	4287

(c) Replacement Hardware Cost for Organization 3

TABLE 4: Replacement Hardware Cost for Organizations

Organization	Hardware Refresh Cost (€)	Deployment Cost (€)	Business Cost(€)	Administration	Total Cost (€)
1	105,899	10,590	10,590		127,079
2	104,802	10,480	10,480		125,762
3	49,824	4,982	4,982		59,788

TABLE 5: Refresh Costs for organizations

simultaneous UPS and power failure [38]. With cooling, it may also be partially explained by a resilient data center design. The design of the data center may specify that it should continue to function in the event of a single CRAC failure. The power required by a CRAC can be calculated using the following formula:

$$C = \frac{Q}{COP(T_{sup})} + P_{fan}$$

Where Q is the amount of power the servers consume, T_{sup} the temperature of the air that the CRAC units supply, P_{fan} the power required by the fans of the CRAC units and COP is the coefficient of performance (COP), that is the ratio of heat removed to work necessary to remove the heat, is a function of the temperature of the air being supplied by the CRAC unit. The COP of a typical chilled-water CRAC is given by the formula below [39]:

$$0.0068T_{sup}^2 + 0.0008T_{sup} + 0.458$$

Assuming a typical supply temperature of 15° ($T_{sup} = 15$) this yields a COP of 2 indicating that the power rating of the CRAC unit should be half of the server power rating with some additional power requirements for powering the fans and a margin of safety. It should not, however, significantly exceed the server power rating as is the case in Organizations 1 and 2. This can be seen in Figure 3. In the case of Organization 1, the cooling power rating is more than six times the power rating of the servers. This is considerably more than necessary even assuming cooling device redundancy. We can also see from Table 3 that the individual

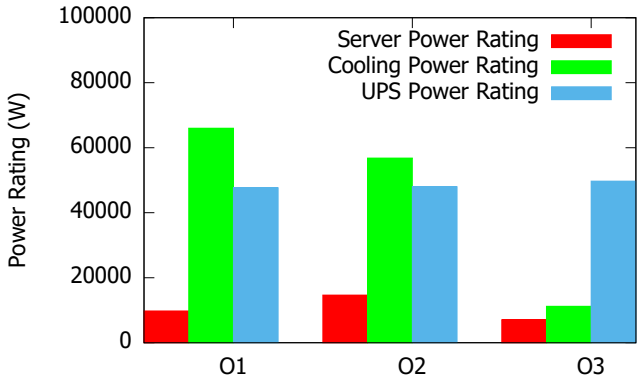

Fig. 3: Comparison of the server power rating, Cooling Power Rating and UPS Power Rating at different organizations.

power rating of CRAC Type 1 is significantly higher than is necessary. We can also see from Table 3 that utilization of the CRAC units is 100% indicating the CRAC units are over utilized. This indicates that greater consolidation of servers is possible by both reducing the overall power rating of CRAC units and setting the utilization level of CRAC units at an appropriate level. If we assume that the data center design implements cooling device resilience, CRAC units with a power rating equal to the sum of the power rating of servers in the data center should be sufficient to cool the servers assuming a design with adequate airflow.

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
7	Server Type 1 Replacement	5021	35147
13	Server Type 2 Replacement	2308	30004
4	Server Type 3 Replacement	1429	5716

(a) Equal Computation Replacement Hardware Cost Organization 1

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
10	Server Type 4 Replacement	2241	22410
4	Server Type 5 Replacement	2325	9300
12	Server Type 6 Replacement	897	10764
7	Server Type 7 Replacement	1142	7994
1	Server Type 8 Replacement	1670	1670
2	Server Type 9 Replacement	2308	4616
1	Server Type 10 Replacement	2308	2308
4	Server Type 11 Replacement	825	3300
1	Server Type 12 Replacement	2325	2325

(b) Equal Computation Replacement Hardware Cost for Organization 2

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
8	Server Type 4 Replacement	2241	17928
4	Server Type 13 Replacement	1386	5544
4	Server Type 14 Replacement	897	3588
1	Server Type 15 Replacement	1142	1142
3	Server Type 16 Replacement	1429	4287
1	Server Type 17 Replacement	1429	1429

(c) Equal Computation Replacement Hardware Cost for Organization 3

TABLE 6: Equal Computation Replacement Hardware Cost for Organizations

Organization	Hardware Refresh Cost (€)	Deployment Cost (€)	Business Cost(€)	Administration	Total Cost (€)
1	70,867	7,087	7,087		85,041
2	64,687	6,469	6,469		77,625
3	33,918	3,392	3,392		40,702

TABLE 7: Equal Computation Refresh Costs for organizations

If we examine Figure 3 we can see that the cooling power rating of Organization 3 is at approximately the correct level. In Organization 3 the UPS power rating is more than seven times the power rating of the servers indicating that the organization is over provisioned with UPS devices. It should be noted that this power rating refers to the maximum load which can be drawn from the UPS. Another important factor in UPS systems is their energy storage capacity which in conjunction with the load will determine how long the data center will operate in the event of a power failure. In this study, we exclude storage capacity as the optimal choice will depend on the transition time from main power to backup generators which can vary significantly depending on the data center. In large data centers, it is typically ten to fifteen seconds [40] but this requires systems and expertise not readily available in smaller organizations. This data was not available for the data centers studied and so is excluded. If we examine Table 3, however, we can see that the utilization level of the UPS devices is 50%. If we assume that a typical UPS deployment is used for redundancy and this is taken into account, then the power rating of the UPS devices is approximately at the correct level. It is, however, possible for savings to be made by purchasing smaller, less expensive UPS devices. This will reduce the time necessary for hardware refreshes to result in overall cost savings for the organization.

Finally, based upon the efficiency of the UPS devices they were purchases at a similar time. Advances in UPS technology [41] have improved the efficiency of UPS. The efficiency of a UPS is also dependent on the loading of the UPS and an efficiency of 93% given a 50% workload indicates that relatively modern UPS devices are being used in these data centers. It is possible, however, to improve the energy savings by increasing the utilization of a more modern UPS. By using a more modern UPS an efficiency of 98% can be achieved [41]. This will also reduce the time necessary for hardware refreshes to result in an overall cost savings for the organization.

5.3 Cost of Updating Hardware

In order to determine the cost of updating hardware, we examine the cost of replacing each server type with its most recent version. The cost for each organization is depicted in Table 4. When selecting the replacement model of each server type the most recent generation of the server type was selected where possible. In cases where the series of the server had been discontinued the closest equivalent from the current server ranges of the manufacturer was selected. When determining the cost of the server the price is taken directly from the manufacturer's website where possible. In cases where this is not possible the price is the average of the at least three re-sellers of the server. It should be noted, however, that considerable economies of scale can be achieved when purchasing servers [42] so these costs should be considered conservative estimates. Finally, each server type has several configurations and the closest configuration to the original is server is used to determine the replacement cost.

When examining Table 4 it is interesting to note that low-end servers are mostly used in these organizations. The

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
5	Server Type 1 Replacement	5021	25105
12	Server Type 2 Replacement	2308	27696
3	Server Type 3 Replacement	1429	4287

(a) High Utilization Replacement Hardware Cost Organization 1

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
8	Server Type 4 Replacement	2241	17928
2	Server Type 5 Replacement	2325	4650
11	Server Type 6 Replacement	897	9867
6	Server Type 7 Replacement	1142	6852
1	Server Type 8 Replacement	1670	1670
1	Server Type 9 Replacement	2308	2308
1	Server Type 10 Replacement	2308	2308
3	Server Type 11 Replacement	825	2475
1	Server Type 12 Replacement	2325	2325

(b) High Utilization Replacement Hardware Cost for Organization 2

Quantity	Server Type	Unit Cost (€)	Total Cost(€)
7	Server Type 4 Replacement	2241	15687
3	Server Type 13 Replacement	1386	4158
3	Server Type 14 Replacement	897	2691
1	Server Type 15 Replacement	1142	1142
2	Server Type 16 Replacement	1429	2858
1	Server Type 17 Replacement	1429	1429

(c) High Utilization Replacement Hardware Cost for Organization 3

TABLE 8: High Utilization Replacement Hardware Cost for Organizations

Organization	Hardware Refresh Cost (€)	Deployment Cost (€)	Business Cost(€)	Administration	Total Cost (€)
1	57,088	5,709	5,709		68,506
2	50,383	5,038	5,038		60,459
3	27,965	2,797	2,797		33,559

TABLE 9: High Utilization Refresh Costs for organizations

average replacement server cost for organization 1 is €2,942. The average replacement server cost for organization 2 is €1,638. The average replacement server cost for organization 3 is €1,557. The average replacement server cost across all three organizations is €1,974. This is considerably less than the figure used for calculating the cost of a cloud in the seminal work by Greenberg *et al.* [14]. This is in line with the preferred building blocks of low end servers used in warehouse scale computing [40]. It should be noted that the average replacement server cost for organization 1 is approximately double that of the other organizations. This difference is largely due to the use of servers from a relatively niche manufacturer for specialized tasks.

The total cost for replacing the servers in each organization is a considerable amount given their relative funding levels. The total cost for replacing the servers in organization 1 is €127,079. The total cost for replacing the servers in organization 2 is €125,762. The total cost for replacing the servers in organization 3 is €59,788. The breakdown of the hardware refresh cost, deployment cost and business administration cost is depicted in Table 5. This cost, however, can be significantly reduced. These costs assume that each server is replaced with a modern version of its type. Figure 2 shows that the performance per watt has increased significantly since 2011. Thus, even if we ignore the value for 2018 as an outlier 65% of the servers can provide the same computing resources as the servers currently operating in the organizations. This percentage is simply calculated by comparing the performance per watt of 2011 which is 0.3423 with the performance per watt of 2017 which is 0.5252.

If servers are replaced in a proportional fashion where the total number of servers being replaced is calculated as 65% of the number of servers operating in the organizations rounding up so that there will be 65% or more servers operating in the organization after the refresh. Servers which are not replaced are decided based upon two rules. Firstly, if there are more than one of the server types operating in the organization then the number of servers of this type is reduced by one. If this would remove too many servers, the most expensive server is removed. Secondly, if the number of servers removed is still not equal to the total number required to achieve 65% then the most expensive server type of the remaining types is removed. This methodology is used as it maintains the heterogeneous server mixture which is needed for diverse workloads while reducing the expense of refreshing the hardware as much as possible. The server types utilized in each organization are depicted in Table 6. If this methodology is used the total cost for replacing the servers in organization one is €85,041, the total cost for replacing the servers in organization two is €77,625 and the total cost for replacing the servers in organization three is €40,702. The breakdown of the hardware refresh cost, deployment cost and business administration cost is depicted in Table 7. This represents an average reduction in the cost of refreshing the hardware of 34.4%. In addition, it will further reduce the energy consumption of the organizations after the hardware refresh. This will reduce the time until the payback point is reached where the savings from reduced energy consumption is greater than the cost of the hardware refresh.

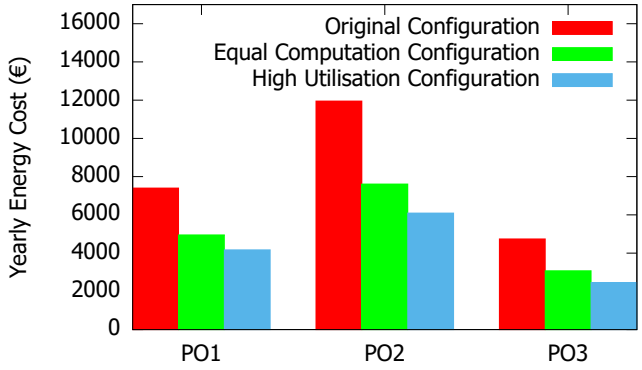

Fig. 4: Yearly energy consumption of servers of organizations with different configurations.

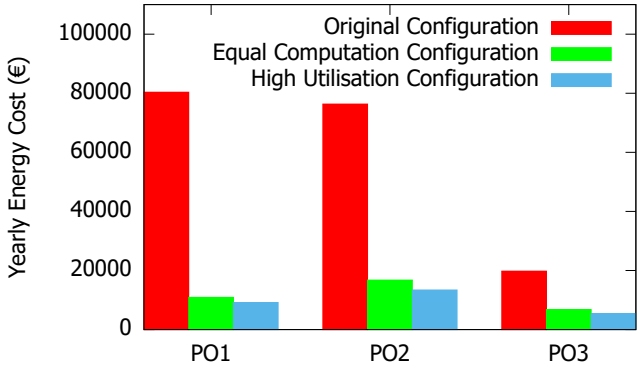

Fig. 5: Yearly energy consumption of organizations with different configurations.

Hardware refresh costs can be further reduced by considering best practice utilization levels. While the server utilization levels are considerably higher than the average CPU utilization levels in data centers reported as ranging from 6% to 12% [13] they are still lower than the best practice representation of the utilization level of 50%[1] [3]. By reducing the total number of servers by 20% of the equal computation refresh scenario the utilization levels can be increased from 40% to 50%. Using the same methodology, the expense of refreshing the hardware can be further reduced. The server types utilized in each organization are depicted in Table 8. If this methodology is used the total cost for replacing the servers in organization one is €68,506, the total cost for replacing the servers in organization two is €60,459 and the total cost for replacing the servers in organization three is €33,559. The breakdown of the hardware refresh cost, deployment cost and business administration cost is depicted in Table 9. This represents an average reduction in the cost of refreshing the hardware of 47.3%. This will also reduce the time until the payback point is reached in a similar fashion to the previous scenario.

5.4 Energy Savings

To determine the energy savings which can be made by updating the hardware we need to examine the current energy consumption of the organizations. Figure 4 depicts the electricity cost for the three server configurations discussed in the previous section at each organization. In the original configuration, the servers are simply upgraded to their latest version and the number of servers is maintained. In the equal computation configuration, the computational resources available are maintained and the number of servers shrinks accordingly due to the improved performance of the processors. In the high utilization configuration, the number of servers further shrinks from the equal computation configuration to increase the utilization of the servers from 40% to 50%. It should be noted that the price of electricity varies slightly between each organization. The price of electricity for organization one is 0.12209195€/kWh. The price of electricity for organization two is 0.12541€/kWh and the price of electricity for organization three is 0.1314€/kWh. The prices, however, are quite similar and do not substantially affect the operating cost of the servers and supporting equipment.

From Figure 4 we can see that although the overall cost of energy consumed by servers in a year varies somewhat between the three organizations, the relative reduction in the energy cost when the different configurations are used are broadly similar. The reduction of energy cost when the equal computation configuration is used instead of the original configuration ranges from 35.1% to 37.2%. The reduction of the energy cost when the high utilization configuration is used instead of the original configuration ranges from 43.7% to 49%. Both configurations represent significant savings, but further savings can be achieved by modifying the utilization levels of supporting equipment in particular cooling equipment.

Figure 5 depicts the electricity cost of the servers, cooling equipment and UPS equipment for the three server configurations discussed in the previous section at each organization. In this figure, the original configuration represents a configuration where the servers are simply upgraded and the utilization levels of the supporting cooling equipment are maintained at the reported level. The equal computation configuration represents a configuration where the computational resources available are maintained and the number of servers shrinks accordingly due to the improved performance of the processors. In addition, the utilization of the cooling equipment is set to be 20% higher than the energy consumption of the servers. This is a somewhat conservative level as the cooling costs can be lower than the energy consumption of the servers if a sufficiently high supply temperature is utilized. Indeed, the general trend in recent years has been a gradual increase in the recommended allowable temperature [43]. For example, if we take the maximum allowable temperature described in [43] of $45°C$ and use the formula for describing the power required for a CRAC we find that power required by a CRAC is 7% of the power that the servers consume if the power required by the fans are excluded. Using such a high supply temperature, however, requires a specific architecture and careful management and hence, our conservative estimate represent a reasonable compromise between energy consumption and

1. The actual utilization level recommendation is 55% but 5% is needed to account for consolidation overhead.

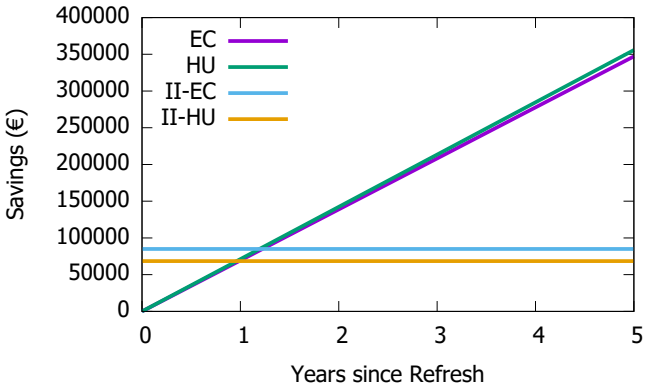

Fig. 6: Return on Investment for organization one under different scenarios.

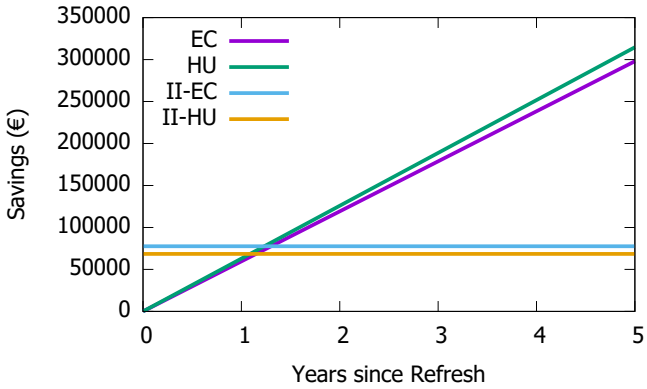

Fig. 7: Return on Investment for organization two under different scenarios.

preventing faults in hardware. For the high utilization configuration, the number of servers further shrinks from the equal computation configuration to increase the utilization of the servers from 40% to 50%.

From Figure 5 we can see that the savings which can be made when considering supporting equipment are very large, even with a relatively conservative estimate of the cooling energy required. In organization one, the reduction in energy cost when the equal computation configuration is used instead of the original configuration is 86.4%. For organization two the reduction is 78.1%. In organization one, the reduction in energy cost when the high utilization configuration is used is 88.6%. In organization two the reduction in energy cost when the high utilization configuration is used is 82.4%. It should be noted that a general reduction in the overall cooling costs is responsible for a large portion of this cost reduction. We can see from Figure 3 that organization three has cooling energy costs which are more closely related to the server energy costs and this is reflected in the energy cost reduction in organization three. In organization three the reduction in energy cost when the equal computation configuration is used instead of the original configuration is 65.2% and the reduction in energy cost when the high utilization configuration is used is 72.7%. These are, however, very significant savings and it should be noted that further savings can be made by increasing the supply temperature of cooling equipment to reduce the utilization levels of the cooling equipment.

5.5 Return on Investment

To determine the time required to achieve an ROI for the equal computation and high utilization configuration we plot energy savings over a five year period and include the initial investment (II) for the two configurations so that the time required to achieve an ROI is clearly visible. This is depicted in Figures 6, 7 and 8. From Figures 6 and 7 we can see that organizations one and two achieve an ROI in less than two years. In addition, with the high utilization configuration organization one has saved over €285,000 after five years while organization two has saved over €245,000. In both cases, this is over three times their current annual electricity cost and thus, represents a significant saving for both organizations.

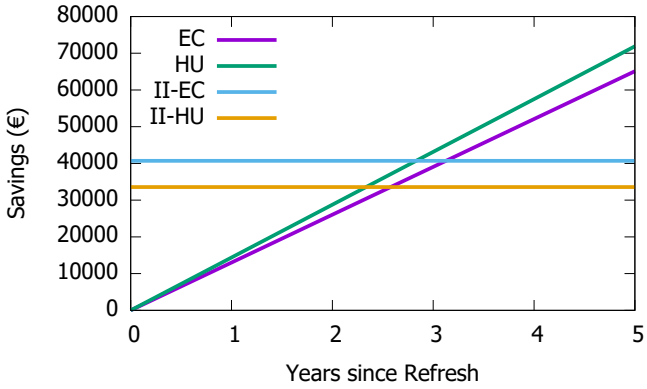

Fig. 8: Return on Investment for organization three under different scenarios.

In organization three the savings are more modest. In the high utilization configuration, it takes slightly less than three years to achieve an ROI and after five years organization three has saved over €38,000. While the savings are more modest in absolute terms this is still approximately two times their current annual electricity cost. In general, the lifetime of a server is approximately three years [14]. If the high utilization configuration is used, this represents an ROI for the organization. Considering the age of the servers depicted in Table 1, however, the lifetime of a server in small organizations is likely to be significantly longer than three years. Thus, the examination of the energy savings for five years could be considered a conservative estimate and show that significant savings can be made in this time period.

6 Study Limitations

This work exhibits a number of limitations as with any research which are discussed in this section along with measures designed for their mitigation.

Firstly, the energy savings depicted in the previous section depend on the specific power consumption of the replacement servers suggested for the organizations. One of the main limitations is that without instrumentation of the servers it is difficult to accurately determine the exact power consumption of the server. While the figures obtained

in Table 1 were obtained via instrumentation and are thus, accurate it is likely that newer servers will consume less power due to improved processor design and the general trend depicted in Figure 2. It is difficult, however, to specify this saving accurately without instrumentation. In order to address this limitation, we use the power ratings of the servers currently used by the organizations. We can safely, assume, that the power consumption will not increase and our results can be considered somewhat conservative. It is also interesting to note that all the results show a linear relationship between savings in operational expenditure and time. Previous work has shown that the relationship between the efficiency of servers and the power of the servers is non-linear [44]. If we used power measurements from instrumentation of the new server models, we would likely see this non-linear relationship in our figures.

Secondly, the energy savings which can be attained by reducing the cooling energy supplied to the machines relative to the energy consumption of the server is dependent on the architecture of the room where the servers are housed. For example, if aisle containment [45] is utilized this allows a much higher supply temperature to servers as the danger of thermal "hot-spots" is considerably less. In order to address this limitation, we assume a utilization level for cooling equipment which is significantly less than the theoretical maximum. Further savings may be possible if the servers are housed in a facility with modern thermal management systems but this cannot be stated definitively. It should be noted that "hot-spots" are more likely due to the higher power density of the new servers so the current cooling structure may not be sufficient. This will be explored in future work.

Thirdly, it is also possible to save energy by replacing network and storage equipment. Similar trends to those found in servers where the performance per watt has been steadily increasing in recent years can be found for both of these devices [3]. While data on these devices was available it was found that the savings offered by replacing these devices were significantly less than the savings achieved by replacing servers. As the savings offered by servers are not substantially affected by the inclusion of a storage and network refresh, we have focused our attention on a server refresh only. Further savings, however, are possible.

Finally, new EcoDesign legislation for servers and online storage devices requires servers to reduce their idle power. In many cases, next generation servers do not have as much computational power as the previous generation as a result of different chip designs which prioritize a lower idle power [15]. Performance per watt, however, is still increasing as indicated by the results of this work. We have assumed that the price of the processor will remain proportional to its computational power and endeavored to select replacement servers which have sufficient computational power to support workload consolidation without performance impairment.

7 Conclusion

This work provides a new perspective on the energy consumption of small data centers and savings that can be achieved with relatively little investment. The work demonstrates that reducing the server population and thereby consolidating workloads can decrease energy costs sufficiently that an ROI after a hardware refresh can be achieved in as little as two years and that organizations can save over three times their annual electricity cost after five years. Additionally, the work demonstrates that consolidation of workloads onto a smaller number of more powerful servers can reduce the energy consumption of data centers. In future work we aim to examine how some of the limitations of our study affect the results. In particular, we would like to instrument the new proposed servers as significant savings may be achieved due to the improved power design. We also aim to examine how the replacement of network and storage devices can achieve an ROI. While the savings offered by replacing these devices are not as substantial as the savings achieved by replacing servers they are nevertheless significant and merit further research.

Acknowledgments

This work was funded as part of the EURECA project which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 649972

References

[1] R. Bashroush, E. Wood, and A. noureddine, "Data center energy demand: What got us here won't get us there," *IEEE Software*, vol. 33, no. 2, 2016.

[2] P. Bertoldi, "The european programme for energy efficiency in data center: The code of conduct," *EU DG JRC Institute for Enery and Transport*, 2016.

[3] A. Shehabi, S. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin, and W. Lintner., "United state data center energy usage report," Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775, Tech. Rep., 2016.

[4] J. Doyle, R. Shorten, and D. O'Mahony, "Stratus: Load balancing the cloud for carbon emissions control," *IEEE Transactions on Cloud Computing*, vol. 1, no. 1, pp. 1–1, Jan 2013.

[5] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. ACM, 2012, pp. 211–222.

[6] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening geographical load balancing," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*. ACM, 2011, pp. 233–244.

[7] M. Ganeshalingam, A. Shehabi, and L.-B. Desroches, "Shining a light on small data centers in the us," 2017.

[8] C. Li, Y. Hu, L. Liu, J. Gu, M. Song, X. Liang, J. Yuan, and T. Li, "Towards sustainable in-situ server systems in the big data era," *ACM SIGARCH Computer Architecture News*, vol. 43, no. 3S, pp. 14–26, 2015.

[9] "Iso/iec 30134-2:2016; information technology – data centres – key performance indicators – part 2 : Power usage effectiveness (pue)."

[10] "Datatank immersion cooling data centre winning the best ict award 2014 for most efficient data center - pue 1.01."

[11] N. Jones, "How to stop data centres from gobbling up the world's electricity." *Nature*, vol. 561, no. 7722, p. 163, 2018.

[12] R. Bashroush and E. Woods, "Architectural principles of energy-aware internet-scale applications," *IEEEE Software*, vol. 34, no. 3, 2017.

[13] R. Huang and E. Masanet, "Data center it efficiency measures," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2015.

[14] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review*, vol. 39, no. 1, pp. 68–73, 2008.

[15] R. Bashroush, "A comprehensive reasoning framework for hardware refresh in data centers," *IEEE Transactions on Sustainable Computing*, vol. 3, no. 4, pp. 209–220, Oct 2018.

[16] G. Wang, L. Zhang, and W. Xu, "What can we learn from four years of data center hardware failures?" in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2017, pp. 25–36.

[17] J. Alter, J. Xue, A. Dimnaku, and E. Smirni, "Ssd failures in the field: symptoms, causes, and prediction models," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2019, pp. 1–14.

[18] K. V. Vishwanath, A. Greenberg, and D. A. Reed, "Modular data centers: how to design them?" in *Proceedings of the 1st ACM workshop on Large-Scale system and application performance*, 2009, pp. 3–10.

[19] P. Nikolaou, Y. Sazeides, A. Lampropoulos, D. Guilhot, A. Bartoli, G. Papadimitriou, A. Chatzidimitriou, D. Gizopoulos, K. Tovletoglou, L. Mukhanov *et al.*, "On the evaluation of the total-cost-of-ownership trade-offs in edge vs cloud deployments: A wireless-denial-of-service case study," *IEEE Transactions on Sustainable Computing*, 2019.

[20] A. Qouneh, C. Li, and T. Li, "A quantitative analysis of cooling power in container-based data centers," in *2011 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2011, pp. 61–71.

[21] R. Khalid, A. P. Wemhoff, and Y. Joshi, "Energy and exergy analysis of modular data centers," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 7, no. 9, pp. 1440–1452, 2017.

[22] D. Laganà, C. Mastroianni, M. Meo, and D. Renga, "Reducing the operational cost of cloud data centers through renewable energy," *Algorithms*, vol. 11, no. 10, p. 145, 2018.

[23] Y. Guo, M. Pan, Y. Gong, and Y. Fang, "Dynamic multi-tenant coordination for sustainable colocation data centers," *IEEE Transactions on Cloud Computing*, 2017.

[24] C. Qiu and H. Shen, "Dynamic demand prediction and allocation in cloud service brokerage," *IEEE Transactions on Cloud Computing*, 2019.

[25] M. Xu, A. N. Toosi, and R. Buyya, "ibrownout: An integrated approach for managing energy and brownout in container-based clouds," *IEEE Transactions on Sustainable Computing*, vol. 4, no. 1, pp. 53–66, 2018.

[26] Y. Xiao and M. Krunz, "Distributed optimization for energy-efficient fog computing in the tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2390–2400, 2018.

[27] T. Wang, L. Qiu, G. Xu, A. K. Sangaiah, and A. Liu, "Energy-efficient and trustworthy data collection protocol based on mobile fog computing in internet of things," *IEEE Transactions on Industrial Informatics*, 2019.

[28] H. R. Arkes and C. Blumer, "The psychology of sunk cost," *Organizational Behavior and Human Decision Processes*, vol. 35, no. 1, pp. 124 – 140, 1985. [Online]. Available: http://www.sciencedirect.com/science /article/pii/0749597885900494

[29] vmware, "Reducing server total cost of ownership with vmware virtualization software," Tech. Rep., 2007.

[30] D. A. Patterson *et al.*, "A simple way to estimate the cost of downtime." in *LISA*, vol. 2, 2002, pp. 185–188.

[31] J. Hamilton, "Cooperative expendable micro-slice servers (cems): low cost, low power servers for internet-scale services," in *Conference on Innovative Data Systems Research (CIDR'09)(January 2009)*. Citeseer, 2009.

[32] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *Proceedings of the Third ACM Symposium on Cloud Computing*. ACM, 2012, p. 7.

[33] W. Wang, B. Li, and B. Liang, "Dominant resource fairness in cloud computing systems with heterogeneous servers," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 583–591.

[34] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, no. 12, pp. 33–37, 2007.

[35] Z. Ou, B. Pang, Y. Deng, J. K. Nurminen, A. Yla-Jaaski, and P. Hui, "Energy-and cost-efficiency analysis of arm-based clusters," in *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*. IEEE Computer Society, 2012, pp. 115–123.

[36] K. V. Vishwanath and N. Nagappan, "Characterizing cloud computing hardware reliability," in *Proceedings of the 1st ACM symposium on Cloud computing*. ACM, 2010, pp. 193–204.

[37] "Astroids@home," https://asteroidsathome.net/boinc/cpu_list.php. Retreived October 2018.

[38] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH computer architecture news*, vol. 35, no. 2. ACM, 2007, pp. 13–23.

[39] J. D. Moore, J. S. Chase, P. Ranganathan, and R. K. Sharma, "Making scheduling" cool": Temperature-aware workload placement in data centers." in *USENIX annual technical conference, General Track*, 2005, pp. 61–75.

[40] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 8, no. 3, pp. 1–154, 2013.

[41] L. Giuntini, "Efficiency in data centers: Highefficiency operating mode for double-conversion ups," in *PCIM Europe Conference*, 2011.

[42] J. Hamilton, "Internet-scale service efficiency," in *Large-Scale Distributed Systems and Middleware (LADIS) Workshop (September 2008)*, 2008.

[43] ASHRAE Technical Committee (TC) 9.9Mission Critical Facilities, Data Centers, Technology Spaces, and Electronic Equipment, "Data center power equipment thermal guidelines and best practices," ASHRAE, Tech. Rep., 2016.

[44] D. Meisner and T. F. Wenisch, "Does low-power design imply energy efficiency for data centers?" in *IEEE/ACM International Symposium on Low Power Electronics and Design*. IEEE, 2011, pp. 109–114.

[45] J. Niemann, "Hot aisle vs. cold aisle containment," *American Power Conversion, West Kingston, RI, APC White Paper*, vol. 35, 2008.

Joseph Doyle graduated from Trinity College Dublin in 2009 with a B.A.I., B.A. degree in Computer and Electronic Engineering as a gold medalist. He was awarded a Ph.D in 2013 from Trinity College Dublin. He was a post-doctoral researcher in Trinity College Dublin and University College London from 2013 to 2014 and 2014 to 2016, respectively. He was Senior Lecturer in the University of East London, London, U.K. from 2016 to 2018. He is a co-founder of Dithen Ltd. (London, U.K.) and is also Lecturer at Queen Mary University of London, London, U.K. He is an RSE cloud fellow. His research interests include cloud computing, virtual machine classification, green computing, and network optimization.

Rabih Bashroush received BEng degree in computer and communications engineering from AUB, in 2001 and PhD degree in systems engineering from Queen?s University Belfast, in 2005. He is the Director of Research at Uptime Institute, and holds a faculty position with UEL. Before joining Uptime, he spent two decades holding various positions in academia (e.g. at QUB and CMU) as well as industry (e.g. Philips Research and Danfoss). Dr Bashroush was named in the UK Universities top 100 Best Breakthrough list for his work on energy conservation in ICT and was the recipient for the Industry Initiative of the Year DCD Global Awards 2018. He advises various governments (e.g. EU Commission, UK, etc.) and influenced a number of international policies and legislation (e.g. EcoDesign, EMAS, and GPP) on ICT energy conservation. He serves on the main standardization committees for CEN/CENELEC/ETSI as well as ISO/IEC. He serves as an editorial board member on various journals, and on the steering committee of the IEEE Transactions on Sustainable Computing journal. He has been a member of the IEEE and IEEE Computer Society since 1998.