Sustainable Computing: Informatics and Systems 33 (2022) 100618

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

journal homepage: www.elsevier.com/locate/suscom

Interact: IT infrastructure energy and cost analyzer tool for data centers

Nour Rteil [a,c,*], Rabih Bashroush [a,b], Rich Kenny [c], Astrid Wynne [c]

[a] University of East London, London, E16 2RD, United Kingdom
[b] Uptime Institute, London, E1 7LS, United Kingdom
[c] Techbuyer, Harrogate, HG2 8PB, United Kingdom

ARTICLE INFO

Keywords:
Energy efficiency
Environmental impact
Refurbished servers
Circular economy
Server refresh cycles

ABSTRACT

The environmental impact of data centers has become a major concern. In their Q1 2021 report, Intel estimated that there are around 100 million servers deployed globally. In this paper we provide an overview of Interact, a machine learning tool that helps data center operators easily analyze, manage, and optimize the energy, cost, and Carbon footprint of their current servers and infrastructure. We discuss in detail the machine learning algorithms and the energy models used. We also demonstrate the key features of the tool and how it can help data centers become more sustainable by reducing server inefficiencies.

1. Motivation

The environmental impact of data centers is becoming a rising concern in Europe and worldwide. The European Data Center sector has made advances towards increased sustainability in a number of ways. As an example, agreements on best practice include the Climate Neutral Data Pact, a consortium of data center operators and trade associations leveraging technology and digitalization to achieve the goal of making Europe climate neutral by 2050, released information on best practice to support the EU Green Deal. This covers a commitment to energy efficiency, 100 % carbon free energy, water conservation, the reuse and repair of servers and heat recycling as the five pillars of the pact [1].

The 2021 announcement builds upon a set of best practice guidelines, first published in the European Commission JRC Technical Reports in 2018. The document, entitled “2018 Best Practice Guidelines for the EU Code of Conduct on Data Center Energy Efficiency” has detailed guidelines on energy efficiency and other metrics.

The EU has also published several policies and directives on circular economy and practical steps taken to facilitate repair and reuse. The Ecodesign Directive [2], which came into effect in all European member states in March 2020 has a specific section (Lot 9) on servers. The legislation energy efficiency criteria for new servers as well as legislation such as the public provision of firmware be available for servers from 2 to 8 years after the release date, which supports repair, resale and remanufacture.

In addition, the EU Circular Economy Action Plan includes specific ambitions relating to electronic goods and ICT, including targets to extend product lifetimes [3]. While in the past, data center professionals would have been concerned about this due to the huge efficiency gains with every successive generation of servers, recent evidence suggests that the slowdown in Moore’s Law means that product lifetimes can be extended without a negative effect on energy efficiency. This will be discussed in detail below.

The UK published a Circular Economy Package policy statement in July 2020 [4], which identified steps for the reduction of waste and establishing an ambitious and credible long-term path for waste management and recycling.

This was followed in September by the more sector relevant Greening government: ICT and digital services strategy 2020–2025, which included commitments towards zero to landfill on electronic waste and a percentage increase in refurbished or remanufactured goods in the public sector IT estate [5]. The UK Government has also made a commitment towards Right to Repair legislation for consumer goods [6]. If adopted, this paves the way for business related legislation more in line with the EU Ecodesign Directive.

In reality, some Hyperscale data centers are making remarkable commitments to sustainable objectives. Google, for example, pledged to operate on Carbon-free energy [7] in all regions by 2030. Microsoft has also committed to achieving 100 % renewable energy by 2025 [8]. Hyperscalers are also noticeably minimizing the inefficiencies in their power and cooling facilities. Today, Google’s average PUE is 1.11 across all their large-scale data centers.

However, despite clean energy and KPIs targeting inefficiencies in the data center’s infrastructure being widely adopted by the industry

* Corresponding author at: University of East London, London, E16 2RD, United Kingdom.
E-mail address: n.rteil@uel.ac.uk (N. Rteil).

https://doi.org/10.1016/j.suscom.2021.100618
Received 16 August 2021; Received in revised form 7 October 2021; Accepted 30 October 2021
Available online 12 November 2021
2210-5379/

Table 1
Dataset snippet before cleaning.

Hardware Vendor	System	Nodes	Form Factor	Test Method	Avg. watts @ 100 %	Avg. watts @ idle	Performance /power @ 100 %	# Chips	# Cores Per Chip	# Threads	Processor	Processor MHz	Memory (GB)	DIMMS	Power Supplies	Power Supply Rating
Dell Inc.	Dell PowerEdge R710	1	2U	Single Node	227	62.2	4208	2	6	2	Intel Xeon X5675	3066	24	6 × 4096 MB	1	570
Dell Inc.	PowerEdge M520	16	Blade	Multi Node	4026	713	5174	32	8	2	Intel Xeon E5−2470	2300	384	6 × 4096 MB	0	0
Fujitsu	PRIMERGY TX30057	1	Tower	Single Node	240	52.6	5446	2	8	2	Intel Xeon E5−2660	2200	24	6 × 4096 MB	1	450
HPE	ProLiant DL360 Gen10	1	1U	Single Node	237	38.9	12,518	1	28	2	Intel Xeon Platinum 8180	2500	48	6 × 8182 MB	1	800

[9], measures targeting other areas, such as inefficiencies in servers, circular economy, and heat reuse, are still lagging.

Targeting inefficiencies particularly in servers is crucial to improve the overall data center's energy efficiency, especially since servers are the main energy consumers in data centers. This is currently being partially targeted by powering off idle servers and consolidation to increase server's utilization rate. However, there are additional ways to reduce server inefficiencies, such as using energy efficient servers that execute high operations per Watt, as well as optimizing the Software that these servers are running.

In our previous research [10], we presented the energy and cost-saving opportunities available through upgrading server components and refreshing servers to newer models (servers that are one generation younger than the latest model). In the next phase of work, we took these observations one step ahead with the creation of an IT Infrastructure Energy and Cost Analyser tool (Interact), which automatically generates the energy and cost analysis for any data center depending on their existing IT infrastructure.

In this paper, we explain how Interact works. We also demonstrate how it can help data centers target their servers' inefficiencies and optimize their server refresh cycles in terms of energy, cost, and carbon footprint.

The mathematical algorithms used to model the server's energy efficiency, followed by the equations used for the comparative analysis are discussed in Section 2. In Section 3, we go over the main features of the tool and in Section 4, we present three use cases to demonstrate how Interact can be utilized in real-life scenarios. Finally, the conclusion and future improvements are discussed in Section 5.

2. Development of data center server energy evaluation tool: interact

Interact offers a unique tool that is primarily focused on the energy efficiency of servers based on their hardware configuration. It is a continuation of our previous research [10] that evaluated the impact of hardware components and server refresh in terms of energy consumption, carbon footprint, and cost.

Interact is based on machine learning to accurately estimate the server's power and performance values based on a specific hardware configuration. In the following subsections, the training dataset is presented, model features are explained in detail, different regression models are evaluated, and the best model is presented and validated against real-life benchmarking experiments.

2.1. Modeling energy efficiency of servers

The main purpose of the machine learning model is to accurately estimate the power and performance values of a specific server configuration. These values are then used to calculate the server's annual energy consumption and cost estimations.

2.1.1. Dataset

The published server results of the SPECpower_ssj2008 benchmark [11] were used as a starting point for our dataset. The SPEC Power benchmark evaluates the power and performance characteristics of single and multi-node servers and is used as a toolset to improve server efficiency. Though released in 2007, SPECpower_ssj2008 is still one of the most widely used energy efficiency benchmark amongst hardware vendors and computer manufacturers. Table 1 shows a snippet of the dataset.

2.1.2. Data exploration and cleansing

The original data set extracted from SPEC as of 16/02/2021, contained 737 records. The first step of data cleansing consisted of removing non-compliant results (i.e. where Result = 0), which left us with 697 records.

Table 2
Statistical summary of the cleaned dataset.

Category	Statistical summary of the cleaned data set (total data set size m = 688)	Minimum	Maximum	Mean	Range	Median	Standard deviation
	Input features						
CPU	Number of chips	1	8	2.2	7	2	1.1
	Cores per chip	1	64	15.2	63	8	15
	Threads per core	1	4	1.8	3	2	0.4
	Processor speed (MHz)	1600	3800	2532.2	2200	2500	369.1
RAM	Memory capacity (GB)	4	1536	93.7	1532	24	142.7
	Number of DIMMs	2	48	8.9	46	6	6.9
Power supply	Power supplies installed	0	6	0.9	6	1	0.6
	Power supply rating (watts)	0	3000	582.4	3000	560	426.5
Server	Server release year	2004	2020	2013.2	16	2012	4.1
Storage	Number of disk drives	1	4	1	3	1	0.3
	Capacity of disk drives (GB)	14	1920	209.9	1906	160	197.6
	Output label						
	Power at active idle (watts)	9.3	993	91.5	983.7	68.3	83
	Power at 100 % utilisation load (watts)	44.7	2148	344	2103.3	264.5	249.8
	Performance/power at 100 % utilisation load (ssj_ops/watt)	68.1	32,361	7569.4	32292.9	5297	6454.1

Table 3
Summary of selected features for the model.

User Input	Extracted features
Server model	Hardware vendor Release year Form factor
CPU model	Cores per chip Threads per core Frequency
Number of chips populated	Number of chips
Memory capacity	Memory capacity

The second step was converting multi-node results to single-node results. For that purpose, the following fields were divided by the number of nodes:

- Power at active idle
- Power at 100 % utilization load
- Number of chips
- Memory capacity

Then, missing values for the form factor were manually filled in. If the form factor is unknown, the record was removed.

Moreover, there are a few outlier records spotted but only one obvious outlier, which had a performance/power at 100 % load value of 2, was removed.

Since hardware vendor and form factor columns contain text values, these were converted to the numerical format using the one-hot encoding technique.

The resulting cleaned dataset contained 688 records, summarized in the Table 2.

2.1.3. Feature selection

To balance out feature importance whilst maintaining practicality, we decided to exclude power supply details, disk drive details, and number of DIMMs since this information is not widely available among data centers and collecting it would add unnecessary complexity for the users. Table 3 summarizes the final selected features from the user input.

2.1.4. Models evaluation

In this section, we describe the sampling strategy used to split the data between training and testing sets, the different machine learning models used for assessment, and the error metric results for each model.

a Sampling strategy

Since we are dealing with a relatively small dataset (~700 records), we carried out a repeated k-fold cross-validation [12], with 10 folds and 3 repeats to split the data between training and testing sets. Doing so ensures the model is trained and assessed across all records, of different samples and variances. The final evaluation error metric is calculated by taking the average error values of each run.

b Machine learning models

The data is assessed using these 4 common models that are generally used to solve regression problems.

- Random Forest (RF): Random forest is an ensemble of decision tree algorithms. Several decision trees are created where each tree is created from a different bootstrap sample of the training dataset. A prediction on a regression problem is the average of the prediction across the trees in the ensemble [13]. For evaluation purposes, we used the random forest regressor provided by the Scikit-learn python library, with the default hyperparameters [14].
- Gradient Boosting (GB): Gradient boosting is also an ensemble of decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm, like a neural network [15]. For evaluation purposes, we used the gradient boosting regressor provided by the Scikit-learn python library, with the default hyperparameters [16].
- K-Nearest Neighbors (KNN): The principle behind this algorithm is to use the most similar historical examples on new data. When a prediction is required, the k-most similar records to a new record from the training dataset are then located. The similarity between records can be measured in several ways, including the Euclidean distance. Once the neighbors are discovered, the summary prediction can be made by returning the average [17]. For evaluation purposes, we used the k-nearest Neighbors regressor provided by the Scikit-learn python library, with the default hyperparameters [18].
- Artificial Neural Network (ANN): An ANN is based on a collection of connected nodes called artificial neurons, transmitting signals to each other. The signal at a connection is a number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. Neurons and connections typically have a weight that adjusts as learning proceeds. Neurons are aggregated into layers and signals travel from the first layer (the input layer) to the last layer (the output layer), while traversing through hidden layers. Different layers may perform different transformations on their inputs. For evaluation purposes, we used the Keras Deep Learning python library

Table 4
Error evaluation for each model.

Label	Model	Average Mean absolute error (MAE)	Standard deviation for MAE scores	Average Mean squared error (MSE)	Standard deviation for MSE scores
Power at idle (W)	RF	−14.13	4.35	−1328.88	1509.78
	GB	−15.73	4.02	−1086.29	1029.91
	KNN	−26.89	5.33	−3142.02	1539.24
	ANN	−18.43	4.28	−1387.26	882.45
Power at full load (W)	RF	−26.45	6.17	−3459.35	3626.79
	GB	−30.77	5.65	−3115.88	2652.81
	KNN	−78.65	11.19	−18350.92	7807.01
	ANN	−49.05	9.52	−6006.14	3443.8
Performance per watt at full load (ssj_ops/W)	RF	−450.186	100.169	−903660.45	881136.13
	GB	−529.711	92.664	−894948.49	728622.33
	KNN	−1517.084	236.676	−6526411.09	2295719.8
	ANN	−926.416	129.322	−3404429.8	1217161.5

Table 5
Hyperparameter values for the selected model.

Label	N-estimators	Learning rate	Subsample	Max depth
Power at idle	100	0.1	0.7	7
Power at full load	500	0.01	0.7	9
Performance per power at full load	500	0.01	0.5	9

Table 6
Summary of error metrics for the selected tuned model.

Label	Average MAE	MAE standard deviation
Power at idle (W)	−12.808	4.149
Power at full load (W)	−24.978	6.542
Performance per power at full load (ssj_ops/W)	−417.157	91.498

with the baseline model. The baseline model has a single fully connected hidden layer with the same number of neurons as input attributes. The network uses good practices such as the rectifier activation function for the hidden layer. No activation function is used for the output layer because it is a regression problem and we are interested in predicting numerical values directly without transform [19].

a Error evaluation

Features have been standardized before evaluation using Scikit Learn Standard Scalar library, which standardizes features by removing the mean and scaling to unit variance [20]. For the error metric, we are using the average mean absolute error and the average mean squared error for each run of the cross-validation. The results are summarized in the Table 4.

2.1.5. Selected model

The model with the least mean squared error was the Gradient Boosting model. The 4 key hyperparameters with the biggest impact on this model's performance [16], described below, were further tuned using Sklearn's GridSearch python library to reduce the mean absolute error.

- N-estimators: the number of decision trees used in the ensemble.
- Learning rate: the rate that controls the amount of contribution that each model has on the ensemble prediction.
- Subsample: the percentage of a subset of the training dataset that each tree is fit with.
- Max depth: the tree depth that controls how specialized each tree is to the training dataset: how general or overfit it might be.

Table 7
Summary of actual vs predicted values for idle power.

Model	Memory details	Actual idle power	Predicted idle power
DL380 G9	x2 16GB	76.2	66.9
DL380 G9	x4 16GB	73.7	58.5
DL380 G9	x6 16GB	74.4	59
DL380 G9	x8 16GB	75.1	82.3
DL380 G9	x12 16GB	76.1	83.4
Mean Absolute Error (MAE)		**10.89**	
Root Mean Squared Error (RMSE)		**3.25**	
Mean Absolute Percentage Error (MAPE)		15 %	

Table 8
Summary of actual vs predicted values for full power.

Model	Memory details	Actual full power	Predicted full power
DL380 G9	x2 16GB	333.8	261.6
DL380 G9	x4 16GB	369.2	300.1
DL380 G9	x6 16GB	377.2	389
DL380 G9	x8 16GB	391.2	362.5
DL380 G9	x12 16GB	401.1	418.7
Mean Absolute Error (MAE)		**39.8**	
Root Mean Squared Error (RMSE)		**5.96**	
Mean Absolute Percentage Error (MAPE)		11 %	

Table 9
Summary of actual vs calculated power at 25 % load.

Model	Memory details	Actual avg. power	Calculated avg. power
DL380 G9	x2 16GB	118.1	115.6
DL380 G9	x4 16GB	127.4	118.6
DL380 G9	x6 16GB	132.5	141.5
DL380 G9	x8 16GB	137.8	152.4
DL380 G9	x12 16GB	144.4	167.2
Mean Absolute Error (MAE)		**11.56**	
Root Mean Squared Error (RMSE)		**3.23**	
Mean Absolute Percentage Error (MAPE)		**8 %**	

Table 5 summarizes the optimal hyperparameters for each output value that was used to train the final models.

Table 6 shows the improved average Mean Absolute Error (MAE) scores across 10 folds for the tuned gradient boosting model.

2.1.6. Model validation

We have used the results from the benchmarking experiments carried out in our previous research [10] to validate the accuracy of the selected models. The predicted and actual results for idle power and power at full load are summarized in the Tables 7 and 8.

To calculate the %error of the average power consumption, the average power was calculated at 25 % load level using the predicted values for the idle power and full power as described in the subsequent Section (2.2.1) and compared to the reported average server power at 25 % load from the benchmarking reports. Results are summarized in Table 9.

Considering that the experiments for the DL380 G9 configurations above were carried out in a different environment and settings than the ones published in the dataset, an 8 % error for the average power is sensible and as such we can consider our models to be valid.

2.2. Energy, cost, and CO2 calculations

In this section, we explain how we calculated the data center energy consumption, server workload, cost, and Carbon footprint from the predicted server power and performance values.

2.2.1. Energy and workload

The data center's energy is divided into energy consumed by its IT equipment (servers, networking, and storage) and energy consumed by everything else, mainly its mechanical and electrical infrastructure (cooling, UPS heat loss, transformers heat loss, lighting, etc.). In the following subsections, we explain how the total data center energy is estimated, as well as how the total server workload is calculated.

a Server energy

The first step to calculate the total energy consumption of a data center site is to calculate the annual server energy consumption for each server, using the predicted idle power and power at full load values for that server. The total energy consumption is then calculated by adding together the energy of each server in the data center.

The server's annual energy consumption (Es) is calculated according to this equation [21]:

$$Es = [(Pi*\beta) + (Pf*\alpha)]*Rs*8.76$$

Where Pi is the idle power (W), Pf is the power at full load (W), a is the server's utilization rate and β is calculated as 1-α, and Rs is the proportion of time the servers are powered on. To convert watt to kWh per year, the value is multiplied by 8.76 (365*24/1000).

b IT energy

IT energy is calculated by adding the energy consumed by servers, networking devices, and storage devices. Interact currently does not require a list of the networking and storage devices present in a data center. Instead, the tool focuses on estimating the server's energy and calculating other IT equipment as a percentage of that energy whilst factoring in the proportion of these other devices in the data center. Therefore, the IT energy is calculated according to the equation below:

$$Eit = Es/\tau$$

Where Es is the total server energy and τ is the proportion server energy is of the IT energy, and is calculated as 1 – Rs – Rn, where Rs is the proportion of IT energy for storage devices and Rn is the proportion of IT energy for networking devices. The default value is set so that servers attribute to 70 % of the IT energy and networking and storage attribute to the remaining 30 %. The user has the option to set these 2 parameters when configuring their data center, depending on the degree to which their site relies on storage and/or networking.

c Mechanical and electrical energy

If the total energy consumption of the data center is known, the mechanical and electrical energy is simply calculated by subtracting the IT energy from the total energy. If the total energy is unknown, the mechanical and electrical energy is calculated by adding the electrical energy with the mechanical energy. Since UPS efficiency has the biggest impact on electrical energy, the electrical energy is estimated based on how much energy is lost due to UPS inefficiencies, and is calculated according to the equation below:

$$Ee = Eit*(1 - UPS\ efficiency)$$

Electrical energy lost due to transformers is currently not being accounted. Energy consumed by lighting is ignored as it is often very low.

The mechanical energy is estimated based on the energy consumed by air conditioning units, and is calculated according to the equation below:

$$Em = \eta*\mathrm{Pn}*\mu*\mathrm{Ra}*2.5$$

Where η is the number of air conditioning units, Pn is the average BTU rating of an air conditioning unit, μ is the air conditioning utilization rate, Ra is the proportion of the time air conditioners are powered on, and 2.5 is used to convert BTU to kWh.

d Total energy

If unknown, the total energy of the site is calculated by adding the mechanical and electrical (M&E) energy and the IT energy.

e Server workload

The predicted performance per power at full load is a very important energy efficiency metric for servers. It is also used to calculate the server workload (Ws) according to the equation below.

$$Ws = PPW*Pf*\alpha*Rs$$

Where PPW is the performance per watt at full load (ssj_ops/watt), Pf is the power at full load (watt), α is the servers utilization rate, and Rs is the proportion of time the servers are powered on. The overall site's workload is calculated by summing up the workload for all servers.

2.2.2. Cost calculations

Total cost is calculated according to the equation below:

$$Ct = Ce + Cp + Cm + Cs$$

Where Ce is the annual electricity cost and is calculated by multiplying the total annual energy by the electricity cost per kWh. The electricity cost per kWh is determined based on the location and the energy supplier of the site and can be edited by the user. Cp is the procurement cost for servers and is calculated by adding the prices of the servers present in a site. Cm is the average annual maintenance cost per server and is set by the user. Maintenance cost includes server parts repairs and upgrades, licensing, and IT labor cost. Finally, Cs is the annual space/rack cost and is calculated by multiplying the number of racks by the annual cost per rack. The annual cost per rack is specified by the user and excludes electricity cost.

2.2.3. CO2 calculations

The total estimated carbon footprint is calculated by adding scope 2 and scope 3 carbon emissions. Annual scope 2 CO2 emissions are calculated by multiplying the annual total energy by the CO2 per kWh value for the site. CO2 per kWh is determined based on the site's location [22] and can be adjusted by the user. Scope 3 CO2 emissions are calculated by multiplying the number of servers by 922 CO2e in Kg [23] Refurbished server models will not include the scope 3 emissions in the total carbon footprint calculation.

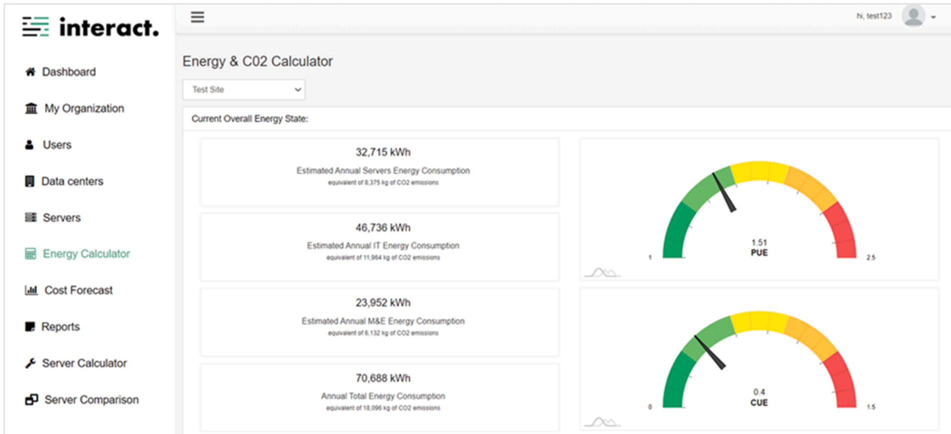

Fig. 1. Energy & CO2 calculator page.

Fig. 2. Interact's recommended refresh scenario to reduce energy.

3. Interact features

Interact provides unique capabilities that help data center operators analyze their current infrastructure's energy, carbon footprint, and cost, as well as provide the user with tailored server refresh recommendations to improve their energy and cost efficiency.

3.1. *Analysis of existing infrastructure's energy, CO2 footprint and cost*

Interact helps data centers estimate their energy consumption and associated carbon emissions. The tool also calculates the associated Power Usage Effectiveness (PUE) and Carbon Usage Effectiveness (CUE) of the site as shown in Fig. 1.

Another pertinent feature provided by Interact is allowing users to view the cost projections of their data center and enabling them to manipulate some parameters to see the direct effect of these parameters on cost. Fig. 2 shows some of the cost projections for a site according to its existing servers and specified parameters.

3.2. *Assessing servers*

Interact allow users to compare the energy consumption and performance between several servers from different vendors. The tool currently includes more than 400 stored server specs and over 1400 CPU specifications, to allow the user to easily find, select and evaluate a model of their choice.

The tool also identifies the data center's 10 % least efficient servers, responsible for consuming higher watts per operation and are worth upgrading to improve the site's overall energy efficiency.

3.3. *Vendor-neutral server recommendations to improve energy efficiency and reduce costs*

Of most widespread appeal to industry is Interact's ability to provide data centers with server refresh recommendations to reduce their total energy and/or cost. A list of over 240 preconfigured server models is evaluated (the list includes several vendors, with new and refurbished conditions), and the best model is identified for the user. The list is regularly updated to include more configurations and models.

Fig. 3 shows an example of a recommended server model scenario for improving the data center's overall energy efficiency and reducing the carbon footprint. In this scenario, the recommended server model for energy also reduces the overall cost over the assessment period, but it is not always the case.

We can observe for this particular example, that while the procurement cost to upgrade the servers is very high (£ 86,827), the cost saved from operating the new servers (electricity cost) as well as maintenance and space (by having the number of servers needed to execute the same workload cut by 80 %) is significant. Upgrading to new servers in this case justifies the procurement cost over time. This is an expected result,

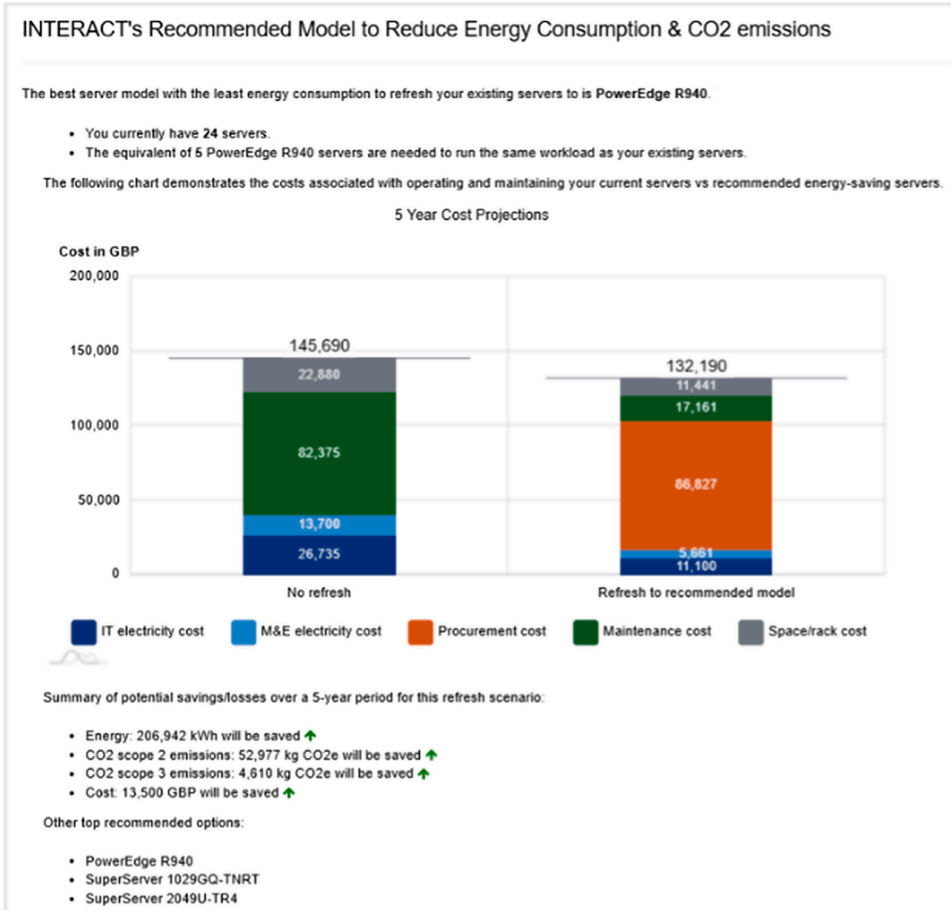

Fig. 3. Cost forecast page.

as explained in detail in our previous study [10]. Fig. 4 below is another example that shows the best server model for reducing cost.

The user also has the option to choose to assess the server refresh scenario in terms of cost, energy, and Carbon footprint using any model of their choice. It is worth mentioning here that some server models, despite their high compute capacity per watt and efficiency, often come at a too high purchase price to justify operational cost savings. It is often the case that refurbished servers guarantee the highest energy and cost savings in most of the scenarios.

4. Use cases

This section highlights three common data center use cases, and how Interact can be utilized to support data centers in each case.

4.1. Lift and shift

"Lift and shift," also known as "rehosting," is the process of migrating an exact copy of an application or workload (and its data store and OS) from one IT environment to another—usually from on-premises to public or private cloud [24]. For data center operators, this is likely to be one of two options:

- Lift and shift of existing infrastructure between sites.
- Lift and shift of client applications from their existing infrastructure to their managed data center.

In the case of shifting between two sites, the ability to optimize the IT required to move the applications and virtual machines is performance-based rather than necessarily physical requirement-led. This is the same as if the customer was moving to a public cloud. The physical hardware may become irrelevant but the workload matters.

The main reason for the move is often cost and performance/resiliency-related. In many cases moving to the public cloud will involve the same risks as moving to a new physical location if consolidation and performance benefits are being realized. In standard lift and shift moves the kit is moved "as-is" to limit negative impact on complex or legacy workloads.

The benefit of physical moves is that costs are more readily in control, we see a lot of this in hybrid environment moves. Some element of the infrastructure is moved to the cloud (items that are cloud-ready such as Virtualized workloads, containerized apps, and microservices) while other parts are kept on-premise or co-located.

Sizing for a co-location move where you want to gain energy efficiencies, performance benefits or size consolidation is difficult. However, the ability to measure current infrastructure against vendor-neutral suggestions is now possible using the Interact tool.

Lift and shift is often necessary when there are compliance reasons to keep physical infrastructure or where applications are being moved to end of life so re-architecting for the cloud makes little sense. There can also be API restrictions or latency issues for IoT or AI data. All these require local hardware provision but future planning around the infrastructure is highly beneficial.

4.2. Rack level consolidation

Needing extra capacity is a common problem for many operators. Finding the best answers to provide space and estate reductions with hardware provision is important. Increasing rack density is particularly important for edge-based deployments where space is at a premium, or

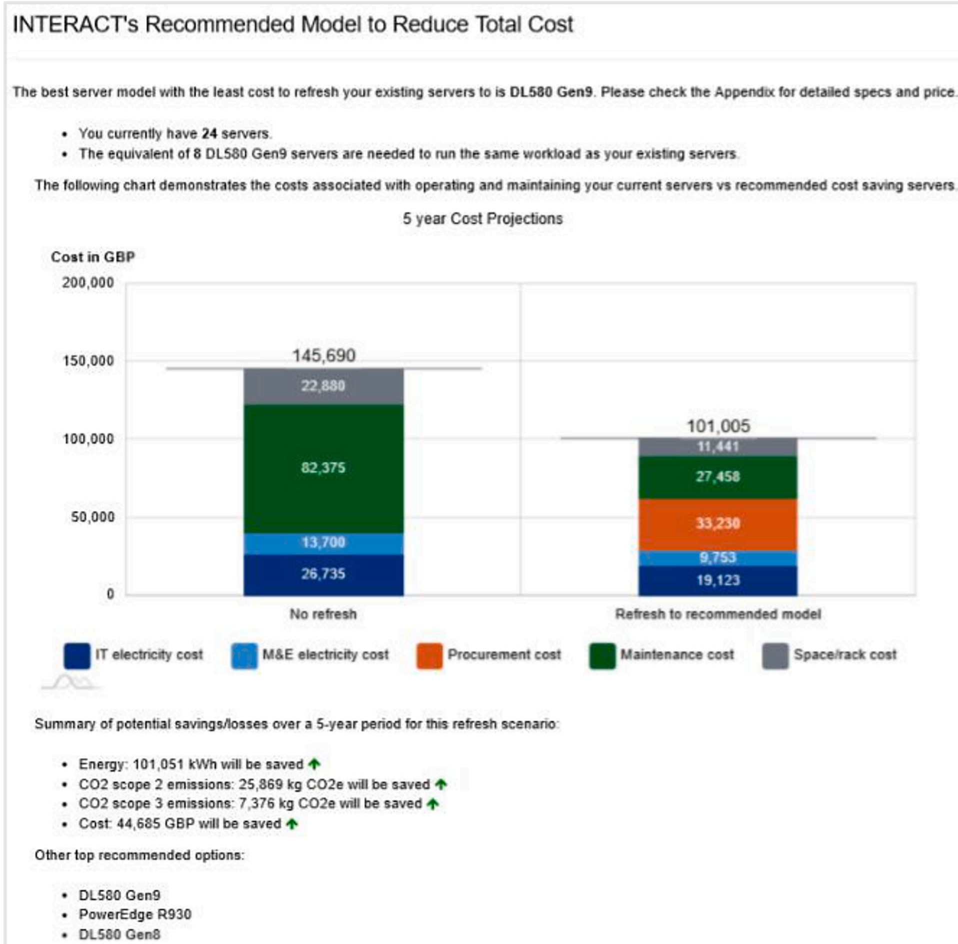

Fig. 4. Interact's recommended refresh scenario to reduce cost.

to make the best use of liquid cooling solutions.

In every case we have seen there has been an ability to increase rack density in current data estates. Either through consolidating less performing servers into higher capacity servers, or through correctly sizing the current need to account for future scalability. In many cases, we have been able to reduce the estate's physical footprint by up to 82 %. In a few rare cases even over 90 %. This allows providers to reduce the physical footprint of their infrastructure and use the same amount of space to facilitate serving more customers with the same fixed facility costs.

Sometimes additional servers or higher performance servers are required but at other times upgrades can be easier because you do not need to migrate workloads to a new machine. Correct sizing and configuration can increase the performance on a rack level and allow redundant servers to be turned off or decommissioned.

By running the data through the interact tool, we were able to run analysis alternatives, which gave the following information as a singular but not uncommon example.

- £534 for a component upgrade resulted in 18 % increase in total performance
- £4729 for a new server resulted in 25 % increase in total performance
- 3% increase in running costs for new RAM vs 25 % increase in running costs for a new server

This is one option for consolidation. Another is component level changes of existing hardware to expand capacity or available performance.

Rack level consolidation is also relatively straightforward to calculate by inputting the servers to analyze per rack and then getting smart recommendations to drive efficiency. For example, identifying the 10 % worst servers in terms of energy efficiency and performance and then working to reduce the number of servers required to generate the same workload and decrease the number of racks required.

We have seen this be particularly effective when a multisite business aims to reduce their physical footprint from 3 data centers to 2 and then have the cloud as their third option. Thereby reducing their physical costs but still working to the same levels of resilience and performance.

4.3. Five-year strategy

A large managed service provider in Northern England recently carried out a snapshot report on the energy efficiency of their server estate. The initial report gave some interesting data on energy and space savings opportunities [25] within the data center as an indication of the size and scope of possible cost savings, through operational and energy costs with associated carbon reductions. It suggested the overall server estate could be reduced by up to 86 %, with a near 4.7 M kWh energy reduction over five years and a £650,000 return.

We all know that replacing an entire server estate in this way is unrealistic. So, the next phase was to highlight immediate changes with high impact results, which would form part of a wider 5-year refresh strategy. We used the Interact tool to zero in on the 10 % worst performing servers within the data center to identify action points and calculate the immediate and long-term benefits to the organization. The tool identified 85 servers, manufactured between 2003–2014, which were performing poorly on energy use, with a combined energy

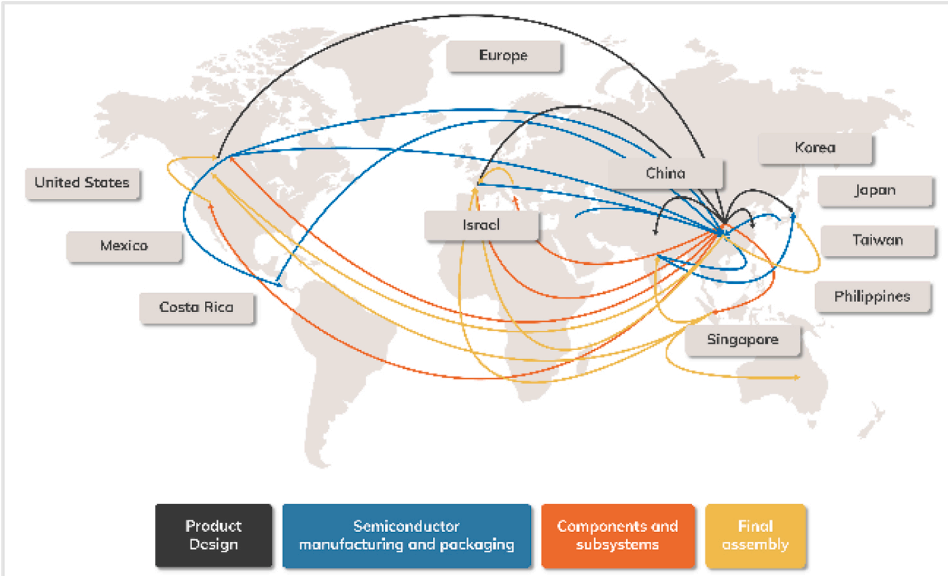

Fig. 5. The global value chain for the electronics industry [23].

consumption of around 250 kW h per year.

The results were startling. Interact was also able to demonstrate a 77 % reduction in energy consumption and associated carbon emissions by switching over from the current infrastructure. This translated to 809,823 kW h of energy and 240,315 kg CO2e. The 809,823 kW h is roughly enough energy to power 219 UK homes for a year [26].

More interesting was the cost savings the suggested changeover represented. Over five years, the suggested solution was 70 % cheaper than continuing with the current infrastructure. Reduced energy costs, M&E energy costs, maintenance, and rack savings represented over £180,000, and that was even factoring in the cost of the replacement machines. There was an additional financial benefit on the 83 % reduction in the number of servers. Current growth levels were suggesting the organization might have to acquire more building space. This report demonstrated they may not have to.

Interact demonstrated that every month the data center chose not to replace the inefficient servers, it was losing money. Comparing the monthly savings from the existing servers to the replacement models, transferring to the suggested option would pay back the cost of procurement within four months.

The data center was also wasting energy and carbon for every month that it continued using the existing, less energy-efficient servers. Energy loss was around 14,849 kW h (enough to power 4 average UK homes for a year). Excess carbon emissions were around 3800 kg. Better yet, the suggested solution was a refurbished model, meaning that there was minimum supply chain carbon in replacing the infrastructure. Estimated embodied carbon savings, as a result, were 11,986 kg CO2e.

Whilst the figures presented here are particular to the customer's facility, they do highlight an important issue. When executing a 5-year refresh program, it is important to carry out research into which machines are performing worst, what solutions are available on replacement, and what the exact cost and energy projections on this are. Too often, capacity planners operate a like-for-like transfer or – potentially worse – over-provision with IT hardware that is unsuitable for the workload. Understanding the server estate, measuring performance, and comparing against a full vendor-agnostic database of makes and models is essential for the bottom line and to reduce environmental impact.

5. Limitations and future work

The SPEC power benchmark measures server performance in terms of server-side java operations per second (ssj_ops), limiting the scope of results to servers performing transactional workload similar to SPEC's SSJ worklet.

Moreover, though published results have been reviewed by the SPEC organization prior to publication, SPEC makes no warranties about the accuracy or veracity of this data [27]. With the lack of third-party auditors to validate the published results, the SPEC power dataset, and consequently the estimated figures generated by Interact, should not be used for energy, carbon or financial reporting. Rather, these scores need to be used only as indicators for the purpose of improving energy efficiency.

Also, the available dataset contains more records for servers for certain hardware vendors (ex. HPE, Dell, Fujitsu) than others (ex. Cisco), and certain server types (ex. rack servers) than others (ex. blades), resulting in possible higher accuracy for certain models.

A possible extension of the current model is to include software/configuration features (such as Operating System and BIOS power profiles) and environmental factors (such as operating temperature and pressure) to study the effect these parameters have on power usage and performance. For now, the scope of the presented model is limited to the default BIOS profile which is set to energy efficiency (performance per watt) to most servers.

Another potential development is to report on the materials consumed when opting for a server refresh. This is a crucial point to consider in sustainability reports since servers and server components are costly in terms of both carbon and Critical Raw Materials (identified by the EU as in low or politically unstable supply).

Servers are made up of steel, aluminum and plastic, which all have a high environmental cost to produce with raw materials according to the JRC [23]. The same report also states that it is not possible to recover 100 % of these materials with current recycling technologies.

Steel and aluminum are two of the top five most energy intensive manufacturing processes according to the IPCC [28]. Plastic is a by-product of the petrochemical industry, responsible for 3.6 % of global greenhouse gas emissions according to Our World in Data [29].

In addition to this, many of the materials within servers have other environmental costs. For example, Aluminum production produces a mixture of metal- and silicon-rich oxides known as "red mud", which is highly caustic and is impossible to break down with current technologies. The build-up of red mud has been identified as a significant risk in aluminum producing areas [30].

In addition, the supply chain is cross continental during production and at end of life, meaning that there is a large amount of international shipment involved with obvious implications on transport related emissions [23] as seen in Fig. 5.

Quantifying the material cost and energy cost involved inthe manufacturing and transport of new server and component, and cross referencing this against the energy and materials saved by using refurbished components and remanufactured machines would be useful in measuring positive environmental impact outside of use phase energy savings.

Authorship statement

Conception and design of study: NR, RB, RK, AW
Acquisition of data: NR
Analysis and/or interpretation of data: NR, RB, RK
Drafting the manuscript: NR, RB, RK, AW
Revising the manuscript critically for important intellectual content: RB, RK
Approval of the version of the manuscript to be published: NR, RB, RK, AW.

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgments

This research is co-funded by InnovateUK as part of the Knowledge Transfer Partnership program under partnership number 11485.

References

[1] Climateneutraldatacentre.net. 2021. Climate Neutral Data Centre Pact – The Green Deal need Green Infrastructure. [online] Available at: <https://www.climateneutraldatacentre.net/> [Accessed 16 April 2021].
[2] European Commission - European Commission. 2021. About the energy label and ecodesign. [online] Available at: <https://ec.europa.eu/info/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/energy-label-and-ecodesign/about_en> [Accessed 16 April 2021].
[3] Environment. 2021. Circular economy action plan. [online] Available at: <https://ec.europa.eu/environment/strategy/circular-economy-action-plan_en> [Accessed 16 April 2021].
[4] GOV.UK. 2021. Circular Economy Package policy statement. [online] Available at: <https://www.gov.uk/government/publications/circular-economy-package-policy-statement/circular-economy-package-policy-statement> [Accessed 16 April 2021].
[5] GOV.UK. 2021. Greening government: ICT and digital services strategy 2020-2025. [online] Available at: <https://www.gov.uk/government/publications/greening-government-ict-and-digital-services-strategy-2020-2025> [Accessed 16 April 2021].
[6] R. Harrabin, 'Right to Repair' Law to Come in This Summer [online] BBC News. Available at: <https://www.bbc.com/news/business-56340077> [Accessed 16 April 2021], 2021.
[7] U. Hölzle, Announcing 'round-the-clock Clean Energy for Cloud, Blog] Google Cloud Blog, 2020. Available at: <https://cloud.google.com/blog/topics/inside-google-cloud/announcing-round-the-clock-clean-energy-for-cloud> [Accessed 16 April 2021].
[8] Azure.microsoft.com, Azure sustainability –Sustainable technologies | Microsoft Azure [online] Available at: <https://azure.microsoft.com/en-gb/global-infrastructure/sustainability/#environmental-impact> [Accessed 16 April 2021], 2021.
[9] Rabih Bashroush, Andy Lawrence, "Beyond PUE: Tackling IT's wasted terawatts.", Uptime Institute Intelligence, Report 34, January 2020. Available at: < http://go.the451group.com/K00PhDY00e9313obt0Fg0Y0>.
[10] R. Bashroush, N. Rteil, R. Kenny, A. Wynne, Optimizing server refresh cycles: the case for circular economy with an aging Moore's Law, IEEE Trans. Sustain. Comput. (2020) 1, https://doi.org/10.1109/TSUSC.2020.3035234.
[11] Spec.org, SPECpower_ssj® 2008 [online] Available at: <https://www.spec.org/power_ssj2008/> [Accessed 16 April 2021]., 2021.
[12] Scikit-learn.org. sklearn.model_selection.KFold — scikit-learn 0.24.1 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html> [Accessed 16 April 2021].
[13] J. Brownlee, How to Develop a Random Forest Ensemble in Python [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/random-forest-ensemble-in-python/> [Accessed 16 April 2021], 2020.
[14] Scikit-learn.org. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.24.1 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html> [Accessed 16 April 2021].
[15] J. Brownlee, How to Develop a Gradient Boosting Machine Ensemble in Python [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/> [Accessed 16 April 2021]., 2020.
[16] Scikit-learn.org. sklearn.ensemble.GradientBoostingRegressor — scikit-learn 0.24.1 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html> [Accessed 16 April 2021].
[17] J. Brownlee, K-Nearest Neighbors for Machine Learning [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/> [Accessed 16 April 2021]., 2020.
[18] Scikit-learn.org. sklearn.neighbors.KNeighborsRegressor — scikit-learn 0.24.1 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html> [Accessed 16 April 2021].
[19] J. Brownlee, Regression Tutorial with the Keras Deep Learning Library in Python [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/> [Accessed 16 April 2021]., 2016.
[20] Scikit-learn.org. sklearn.preprocessing.StandardScaler — scikit-learn 0.24.1 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html> [Accessed 16 April 2021].
[21] R. Bashroush, A comprehensive reasoning framework for hardware refresh in data centers, IEEE Trans. Sustain. Comput. 3 (4) (2018) 209–220, https://doi.org/10.1109/TSUSC.2018.2795465.
[22] Carbon Footprint, Country Specific Electricity Grid Greenhouse Gas Emission Factors [ebook] Available at: <https://www.interactdc.com/static/emissions_factors_2020.pdf> [Accessed 16 April 2021], 2020.
[23] European Comission, Environmental Footprint and Material Efficiency Support for Product Policy. Analysis of Material Efficiency Requirements of Enterprise Servers [online] Available at: <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/environmental-footprint-and-material-efficiency-support-product-policy-analysis-material> [Accessed 16 April 2021]., 2015.
[24] IBM Cloud Education, Lift and Shift: An Essential Guide [online] Available at: <https://www.ibm.com/cloud/learn/lift-and-shift> [Accessed 16 April 2021]., 2019.
[25] R. Kenny, Blog] Results of INTERACT Analysis of 35 Rack Data Centre, Available at: <https://www.interactdc.com/news/details/4/> [Accessed 16 April 2021]., 2021.
[26] C. Topping, Average Electricity Usage in the UK: How Many kWh Does Your Home Use? OVO Energy, 2021 [Blog] Ovoenergy.com. Available at: <https://www.ovoenergy.com/guides/energy-guides/how-much-electricity-does-a-home-use.html> [Accessed 16 April 2021].
[27] Spec.org. SPEC Results Disclaimer. [online] Available at: <https://www.spec.org/spec/disclaimer.html> [Accessed 16 April 2021].
[28] M. Fischedick, J. Roy, A. Abdel-Aziz, A. Acquaye, J.M. Allwood, J.-P. Ceron, Y. Geng, H. Kheshgi, A. Lanza, D. Perczyk, L. Price, E. Santalla, C. Sheinbaum, K. Tanaka, Industry, in: O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel, J.C. Minx (Eds.), Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014.
[29] H. Ritchie, Blog] Sector by Sector: Where Do Global Greenhouse Gas Emissions Come From?, Available at: <https://ourworldindata.org/ghg-emissions-by-sector> [Accessed 16 April 2021]., 2020.
[30] R. Service, Blog] Red Mud Is Piling up. Can Scientists Figure Out What to Do With It?, Available at: <https://www.sciencemag.org/news/2020/08/red-mud-piling-can-scientists-figure-out-what-do-it> [Accessed 16 April 2021]., 2020.