Combined IT and Power Supply Infrastructure Sizing for Standalone Green Data Centers

Marwa Haddad[a], Georges Da Costa[b], Jean-Marc Nicod[a], Marie-Cécile Péra[c], Jean-Marc Pierson[b], Veronika Rehn-Sonigo[a], Patricia Stolf[b], Christophe Varnier[a]

[a] *FEMTO-ST institute, Université de Bourgogne Franche-Comté*
CNRS / UFC / ENSMM, Besançon, France
[Marwa.Haddad, Jean-Marc.Nicod, Veronika.Sonigo, Christophe.Varnier]@femto-st.fr
[b] *Université de Toulouse, Institut de Recherche en Informatique de Toulouse (IRIT), Toulouse, France*
[georges.da-costa, jean-marc.pierson, patricia.stolf]@irit.fr
[c] *FEMTO-ST institute, Université de Bourgogne Franche-Comté*
FCLAB / CNRS / UTBM, Belfort, France
Marie-Cecile.Pera@univ-fcomte.fr

Abstract

In this work, we propose a two-step methodology for designing and sizing a data center solely powered by local renewable energy. The first step consists in determining the necessary IT equipment for processing a given IT workload composed of batch and service tasks. We propose an adapted binary search algorithm and prove its optimality to find the minimum number of servers to handle the IT workload. When the IT sizing is computed, the second step consists in defining the supplying electrical infrastructure using wind turbines and photovoltaic panels as primary sources. Batteries and a hydrogen system are added as secondary sources for short- and long-term energy storage, respectively. In this electrical sizing step, first a set of primary source configurations is determined using a binary search algorithm, then the secondary sources are calibrated so that levels of charge are constant during one day and one year, respectively. Experiments using real IT workload traces and actual meteorological data are conducted to illustrate the provided methodology to decision makers for choosing the best configuration for their data center.

Keywords: Renewable energy, Infrastructure sizing, Green data center

1. Introduction

The growing demand for online services leads to a significant increase in the power consumption of data centers. In 2018, the global data center energy use has been estimated to 205 TWh, which represents about 1% of the world electricity consumption [27].

Economical, political and customer pressure pushes data center operators to improve their carbon footprint. One way of coping with the related increase of the carbon footprint is to add renewable energy sources in the power supply chain.

Many companies, being either big players like Google, Amazon, Facebook, etc, or smaller players, have moved to either partially operating with renewable energies for a share of their energy consumption, or rely on remote renewable power production sites. Ultimately, the renewable energy sources should be co-located with the data center as it avoids losses in the transport and distribution of electricity. Ideally, renewable energy sources are to be directly installed on site.

The question of location and size of the data center is nowadays mainly commercially directed: land costs, financial advantages given by the State, market of electricity, environmental conditions (mainly for cooling reasons), and the power that can be drawn from the power line influence these choices. It is not only a research area anymore, many consulting companies offer their service to build a data center, though the integration of renewable energies is still in its infancy.

In general, the target IT workload is roughly estimated, usually using a basic peak demand, in or-

der to cope with uncertainty and future probable usage. The number of servers in server rooms is adapted to this overestimate, leading to resource waste when the actual load is run on the servers. In the best case, and in the context of renewable energy powered data centers without connection to the grid, some authors ([21], [8], and [36]) proposed to dynamically manage IT workload according to power availability, so that idle servers can be shut down and the load consolidated on fewer servers. In this context, an overview has been published by Ishfaq et al. [1] about power/energy/thermal-aware policies. [21] minimizes the makespan of High Performance Computing (HPC) tasks, while [8] minimizes the number of due date violations for batch tasks, in both cases constrained by a power envelope. Sharma et al. [36] propose a more optimistic approach where web applications do not suffer from the regular on/off power cycles of the machines, while their execution is constrained by renewable energy. Other existing approaches consider a "follow the renewable" concept ([26], [25]) by balancing the load among several data centers and using the right mix of renewable energies. In addition, a theoretical and experimental study proposed by Khargharia et al. [23] optimizes for power and performance of a distributed platform. However, to the best of our knowledge, the initial sizing of a data center powered by renewable energy has never been studied, and the existing works aim at coping with the rough estimate: We support that the presence of renewable sources campaigns for less useless servers and that the IT sizing must be revisited accordingly.

Conversely, more works have been conducted for the sizing of a power plant integrating renewable power sources. These include optimal solutions, heuristics, or metaheuristics to find the proper size and number of electrical components (see Section 2). They are based on prediction models of the weather conditions (solar and/or wind), some of them in the context of data centers (solar panels and batteries in [14]). However these studies do not integrate the IT workload, nor include several combinations for the on-site power supply.

In this work, we design an on-site data center that is solely powered by local renewable energy and we investigate its sizing. The sizing consists of two steps. First, the necessary IT equipment for processing a given IT workload is determined, giving the estimated power over time needed for the IT infrastructure. In a second step the electrical infrastructure is defined to produce enough energy to power the IT infrastructure taking into account the matching over time of the IT power consumption and renewable power production. We investigate the sizing of power-plants consisting of wind turbines (WT) and photovoltaic panels (PV) as primary sources. To cope with the fluctuations in the energy production, we add batteries for short term storage and hydrogen tanks for long term storage and seasonal variations. Those secondary sources are also taken into account in the sizing process.

The main contribution of our research is to provide a methodology to propose a set of infrastructure sizing combinations given an IT workload and a data center location (with its weather conditions). These different combinations can then be tested against a variety of IT workloads to finally choose the best one for the case at hand, depending on the metrics selected by the decision maker.

The rest of this paper is organized as follows: In Section 2 we detail the related work in electrical sizing with renewable energies. Section 3 provides the decision problem description while Section 4 details the underlying IT and power supply models. The sizing methodology is described in Section 5. Section 6 is providing results of the methodology under different IT workload and weather conditions. Finally Section 7 concludes and gives perspectives on the work.

2. Related work

The problem addressed in this paper is twofold: designing and sizing an IT infrastructure and a power plant including only renewable sources as primary sources. As mentioned above, the problem of the initial IT sizing on this basis has never been addressed before. However, a great deal of work has been carried out by researchers for more than ten years on the design of power supply infrastructures only or partially based on renewable energy [39]. In practice since sun and wind are free and accessible everywhere on earth, even in the most remote areas, these energies are the two main renewable energies that are commonly chosen for the construction of such power plants [3] even if other renewable energies exist [32, 10]. Anoune et al. highlight in [3] that separating sun and wind leads to an over-sizing of the system.

Due to the intermittency, sizing a Hybrid Renewable Energy System (HRES), whether in grid connected or in standalone systems, is a very important issue. Many research papers have been addressed and published on this hot topic in order to find the most suited power infrastructure to the context of use and its appropriate size. They take the power production intermittency and its forecasts into account and allow to maximize (or minimize) predefined performance criteria, not only the traditional economical cost, etc. For instance, Sawle et al. in [34] published a literature analysis where the design of HRES connected to the grid is given. This review illustrates that hybrid systems based on hybrid renewable sources give good indicators in terms of energy cost and reliability. Similar results were also obtained by Erdinc and Uzunoglu in [12] where authors point out the advantage of an optimal design for renewable energy in terms of cost, after an analysis of "optimum sizing approaches in the literature".

In addition, since sun and wind are free and accessible everywhere, it is advantageous and realistic to create HRES based on stationary power generation for isolated areas. This point has been discussed by Anoune et al. in [3] where the authors highlight that separating sun and wind leads to an oversizing of the system. In their work they review the most common typologies and present mathematical models and comparisons between existing implementations based on different sub-optimal optimization techniques (nature inspired optimization techniques). Another analysis about methods to optimize the sizing of standalone HRES is given by Bernal-Agustín and Dufo-López in [6]. This study shows that these systems have a high availability when associated with back-up sources such as batteries and then become a viable and credible alternative to the classical energy sources.

To summarize, a state of the art on hybrid renewable energy source sizing is defined in the previous reviews for all types of applications, and they do not take particular features and usage of data centers into account. Many researchers aim at proposing various sizing methods in order to reach optimal solutions of their own systems. These methods could be categorized as follows.

2.1. Probabilistic method

Yang et al. [48] propose a probabilistic method in which they prove the importance of choosing a suitable typical meteorological year (TMY) in order to get an accurate assessment of performance in a hybrid PV-wind energy system. Another probabilistic approach is suggested by Tina et al. [42]. Their method is based on convolution techniques using probability density functions to assess the long term performance of hybrid solar and wind power systems.

2.2. Analytical method

Several computer tools have been developed in order to help decision makers to analyze the integration of sources for optimizing, designing and evaluating the performance of PV-wind hybrid systems as discussed in comprehensive reviews by Bernal-Agustín and Dufo-López [6], Erdinc and Uzunoglu [12], Sinha and Chandel [40], Al-Falahi et al. [2], and Anoune et al. [3]. The most popular and most widely used tools are the commercial software named HOMER for Hybrid Optimization Model for Electric Renewable, developed by the National Renewable Energy Laboratory (NREL) [28] and the Hybrid Power System Simulation Model (HYBRID2) [34]. Indeed, HOMER is defined as a set of "the most powerful tools for this purpose" by the authors [5]. The paper is a state-of-the-art review of existing work based on the use of HOMER for HRES planning. However, all these softwares have strong limitations such as black box coding, different working platforms, and they are not as flexible as optimization techniques which can be used as per research criteria.

2.3. Iterative methods

Many hybrid renewable systems are designed using genetic algorithms to achieve a sizing as close as possible to the optimal solution, depending on the target objective. For instance, Kaldellis et al. [20] propose to minimize the system cost by means of electrical load under some design constraints. Similar works can be found by Dufo-López and Bernal-Agustín [11] and Yang et al. [47, 46]. Ashok [4] obtained a hybrid system among different combinations for a rural community, minimizing the total life cycle cost and ensuring system reliability: a numerical algorithm based on the Quasi-Newton method is used to solve the optimization problem [31]. Numerous methods are based on Particle Swarm Optimization (PSO). Sawle et al. [34] mentioned studies using this popular optimization technique with results obtained on the HOMER software.

2.4. Hybrid method

Finally, many researchers [7, 22, 38] modify genetic algorithms in order to give the designer the choice of the configuration. This can be done by considering non-dominating Pareto sets in which a criteria has to be selected in order to find the appropriate solution. In [22] by Katsigiannis et al., the optimization objective is twofold and consists in minimizing the system cost of energy and greenhouse gas (GHG) emissions by six different constraints. The main originality comes from the assessment of GHG emissions based on life cycle analysis. A similar work has been proposed by Wang and Singh in [45] where the set of non-dominated Pareto solutions is obtained using a PSO algorithm. The optimization objective is also twofold (technological and economical), or even threefold (technical, economical and environmental).

Khalaj et al. [16] present a whole data center design by minimizing the total amount of power consumption of the data center, including the cooling system, and based on an Integer Linear Programming approach. Nevertheless, none of the aforementioned methods treats a crossed data center IT and HRES sizing. This new proposed paradigm is the originality of this paper. It harnesses the existing relationship between power demand and power production for the sizing process.

3. Problem Statement

3.1. Framework

This work has been developed in the context of the DATAZERO project [30]. It aims at investigating possible solutions to design and operate a data center which is only supplied with renewable energy. The goal is to study how to build and manage such a data center without any connection to the grid while taking into account intermittent power production over time. The complete removal of the connection to the power grid imposes new challenges, such as the sizing of storage devices and renewable energy sources in order to provide enough energy to the data center while ensuring the client Quality of Service (QoS). The proposed sizing approach aims at preventing over-sizing or under-utilization of the data center by avoiding redundant equipment.

The energy supply of the data center is divided into two main types, primary and secondary sources (see Figure 1). We focus on solar and wind for primary renewable sources as justified before [3], so we use photovoltaic panels and wind turbines. To complement them, it is mandatory to associate short- and long-term storage devices such as batteries and hydrogen systems as is conventionally accepted [24]. This choice is all the more justified as our project involves the construction of a completely energy standalone data center, therefore without external energy input. Thus, the long-term storage device cannot be a conventional internal combustion engine, but must be a reversible storage device as the hydrogen system (i.e., the overproduction of electricity is stored using electrolysers in the form of hydrogen, or the production of electricity using fuel cells when it runs out). In the latter, electrolyzers (EZ) are used to store in the hydrogen tank excess energy in form of hydrogen obtained through electrolysis. Fuel cells (FC) allow to transform hydrogen into electricity by the reverse chemical reaction. This is imposed by the fact that renewable energy is intermittent by nature. As a matter of fact, the data center must operate despite the alternation of day and night and the differences in the energy production during the seasons.

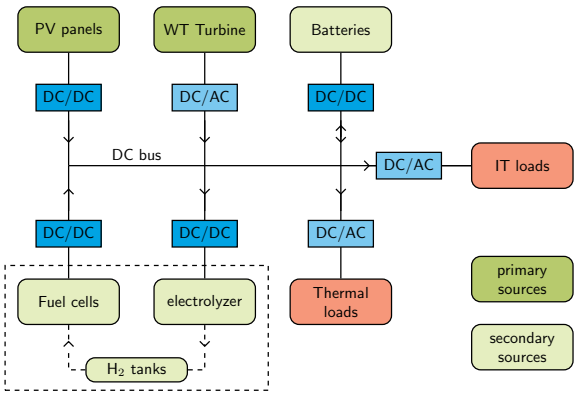

Figure 1: Electrical architecture of DATAZERO. Dark green boxes indicate primary energy sources, secondary sources are shaded in light green (source: Robin Roche [30]).

3.2. Problem description and hypothesis

In this work we focus on the sizing of such a data center. The sizing can be defined as a decision problem that aims at identifying the needed IT elements and the associated electrical devices to satisfy a given computation service that has to be provided over time. We consider as inputs: (i) an estimate of the user demands over time (one year) called "the workload", (ii) the location of the data center and

its historical weather conditions (solar and wind) at least over one year, (iii) a desired Quality of Service (QoS) for the running applications. We answer to the following question: based on these inputs, which IT and electrical infrastructures are needed to process a given workload under a constrained QoS? The answer to this question is called the sizing of a standalone data center. It aims at computing: (i) the number of servers, (ii) the area of photovoltaic panels, (iii) the number of wind turbines, (iv) the capacities and the power of the batteries, (v) the size of the power hydrogen system (power of electrolyzer and fuel cells) associated with the hydrogen tank capacity.

3.3. Decision problems

This general sizing problem can be addressed by the following steps: First, considering a workload and a scheduling strategy, the decision problem is to find the smallest set of servers that is able to process the workload within a given QoS. This architecture and the resulting schedule provide a power envelope or profile. This profile is one input for a second decision problem which consists in defining the power supply architecture. Two levels are to be distinguished: (i) short- and (ii) long-term variation of the power production: (i) Because of the variation of the daily power production, the battery capacity is first sized considering the worst day conditions (due to unbalanced workload and/or season and/or bad weather). This problem can be viewed as a min-max optimization problem (minimize the sizing for the worst day). (ii) To take into account the fluctuation of the renewable power production within one year due to seasonal variations, we have to consider long term energy storage. Consequently since there is no connection to the power grid, the data center should store the excess production when possible for later use. The decision problem we face is to define several combinations of primary sources (photovoltaic panels and wind turbines) and short- and long-term storage elements (batteries and hydrogen system) capable of ensuring the autonomy of the data center. One combination can be selected later by the decision maker using criteria such as the economical cost, the footprint, the loss of power supply probability of the system, etc.

These criteria are either chosen individually or in combination to help to select the right configuration for the renewable energy system.

4. Models

In the description of the sizing decision problem, we have introduced several inputs. This section is dedicated to both IT and power supply models. First we detail the models used in the IT decision part and then the models used in the electrical decision part. The decision horizon $\mathcal{H}$ within which decisions are made is discretized into K indivisible time slots whose durations are Δt with $\mathcal{H} = K\Delta t$. For the sake of simplicity, we assume that one time slot takes one unit of time ($\Delta t = 1 u.t.$). In practice, we assume in the following that $\mathcal{H}$ is one year (365 days, 8760 hours), $\Delta t = 1h$ and hence $K = 8760$.

4.1. IT models

The role of a data center is to deliver digital services, to produce results after processing tasks, etc. This set of work that the data center has to process is called its workload. In the following we first give the properties of the workload and then the architecture on which the workload is supposed to be executed. Notations used in this part are given in Table 1.

The workload is the result of users' submissions and consists of two distinguished types, (i) services $S_i \in \mathcal{S} = \{S_1, \ldots, S_r\}$ and (ii) tasks $T_i \in \mathcal{T} = \{T_1, \ldots, T_n\}$. The QoS requested by a user is different for services and tasks. (i) Each service S_i (such as web services, databases) is defined as a load $ws_{i,k}$ over time with no flexibility, i.e., an amount of work or number of million instructions (MI) to be executed during the time slot k for all times t such that $(k-1)\Delta t \leq t < k\Delta t$. We consider that services can not be delayed as they are in a direct interaction with the users. (ii) Tasks are considered as applications that can be delayed, providing flexibility and malleability. We define the flexibility as a time window during which the execution of a task can be deferred. We consider this flexibility as a constant δ for all tasks. Each task T_i has a requested number of instructions $wt_{i,k}^{req}$ in MI at the time slot k for all times t such that $(k-1)\Delta t \leq t < k\Delta t$. Once scheduled, tasks can be delayed up to a maximum of δ *u.t.*, i.e., $c_i^{sch} - c_i^{req} \leq \delta$ where c_i^{req} and c_i^{sch} are respectively the completion time of T_i, should T_i be executed at its requested time (i.e., as soon as possible), and the completion time when T_i is actually scheduled. The amount of instructions to perform task T_i when scheduled is now $wt_{i,k}^{sch}$ at each time

Table 1: Main notations for the IT model

$\mathcal{H}$	decision horizon $\mathcal{H} = K\Delta t$
K	number of time slots $\Delta t = 1\,\text{h} = 1 u.t.$
$\mathcal{W}$	the workload ($\mathcal{W} = \mathcal{S} \cup \mathcal{T}$)
$\mathcal{S}$	set of services, $\lvert\mathcal{S}\rvert = r$
S_i	one service of $\mathcal{S}$ with $1 \leq i \leq r$
$ws_{i,k}$	amount of work of S_i at time k [MI]
ws_k	total amount of work of all services at time k [MI]
$\mathcal{T}$	set of tasks, $\lvert\mathcal{T}\rvert = n$
T_i	one task of $\mathcal{T}$ with $1 \leq i \leq n$
$wt_{i,k}^{req}$	amount of work of T_i requested at time k [MI]
wt_k^{req}	amount of work of tasks requested at time k [MI]
δ	scheduling flexibility [$u.t.$]
$wt_{i,k}^{sch}$	amount of work of T_i when scheduled during time slot k [MI]
wt_k^{sch}	total amount of work of tasks scheduled at time slot k [MI]
c_i^{sch}	completion time of T_i when scheduled with $c_i^{sch} - c_i^{req} \leq \delta$ [$u.t.$]
m	number of homogeneous machines
$\mathcal{M}$	set of homogeneous machines, $\lvert\mathcal{M}\rvert = m$
M_j	one machine of $\mathcal{M}$
$p_j = p$	max. power consumption of M_j [W]
nbI	max. number of inst. of M_j during any time slot ($nbI_j = nbI\ \forall j$) [MI]
$maxW$	max. number of inst. for $\mathcal{M}$ ($maxW = m \times nbI$) [MI]
PUE	Power Usage Effectiveness constant
D_k	power demand for the time slot k [W]
$\mathcal{D}$	$=\{D_k, 1 \leq k \leq K\}$

slot k with $1 \leq k \leq K$. The malleability of a task T_i means that at each time slot k the number of processors in charge of computing instructions of T_i can vary. The task model is a simplified version of the models found in [43, 15].

Figure 2 presents an example of an incoming workload composed of batch tasks $\mathcal{T}$ (in blue) and services $\mathcal{S}$ (in orange) within the period of one whole year, hour by hour (8760 h). The actual load (in green) represents the resulting workload (services plus batch tasks) after using the flexibility to delay batch tasks. The red line represents the maximum number of instructions that the entire set of servers is able to execute during one time slot. The purpose of this figure is to show that the use of a scheduler allows us to modify the execution of the requested workload and to limit the number of servers without reducing the required quality of service.

We assume that the hardware IT architecture consists of a set $\mathcal{M}$ of m homogeneous servers or machines $M_j \in \mathcal{M} = \{M_1, \ldots, M_m\}$. Each machine M_j consumes a maximum of power $p_j = p$ Watts for a corresponding maximum number of instructions $nbI_j = nbI$ MI. The addressed sizing problem for the IT part is to compute the smallest value for m such that a schedule exists and is able to process the workload $\mathcal{W}$ with the expected level of QoS.

We recall that the QoS is different for tasks and services. Services can not be delayed whereas tasks can be executed while their delay respects the flexibility value δ. Given a number of servers m and a possible schedule that meets the required QoS, it is possible to know which amount of power is needed at each time slot k for this schedule. The power demand for each time slot k ($1 \leq k \leq K$), denoted $\mathcal{D} = \{D_1, \ldots, D_K\}$, is another output additional to the IT architecture sizing. D_k is proportional to the number of instructions that are executed during time slot k. Knowing that nbI instructions are needed and knowing the maximum power consumption p for one machine, the average power cost for one instruction can be approximated by p/nbI [29]. Albeit this model is not precise [43] for a small number of hosts as a single computer is not power-proportional, with a larger number such as the one aimed in this research, the error becomes negligible. Please note that our approach does not depend on the actual power-model used and would stay valid using more precise power models [9]. Finally, the power demand has to take the cooling and utilities into account. The constant PUE [33] (Power Usage Effectiveness) measures this proportional extra power cost. PUE is the ratio between the total energy consumed by a data center and the energy consumed only by the IT part. It is assumed that idle nodes, that are switched off, consume 0 Watt.

Every power demand D_k needed for every time slot k ($1 \leq k \leq K$) can be expressed by summing the amount of instructions from services and tasks scheduled at each time slot k multiplied by the power consumption of one instruction. We recall that $ws_{i,k}$ (resp. $wt_{i,k}^{sch}$) is the number of instructions of the service S_i (resp. the task T_i) which is scheduled onto machine M_i in $\mathcal{M}$ at time slot k or

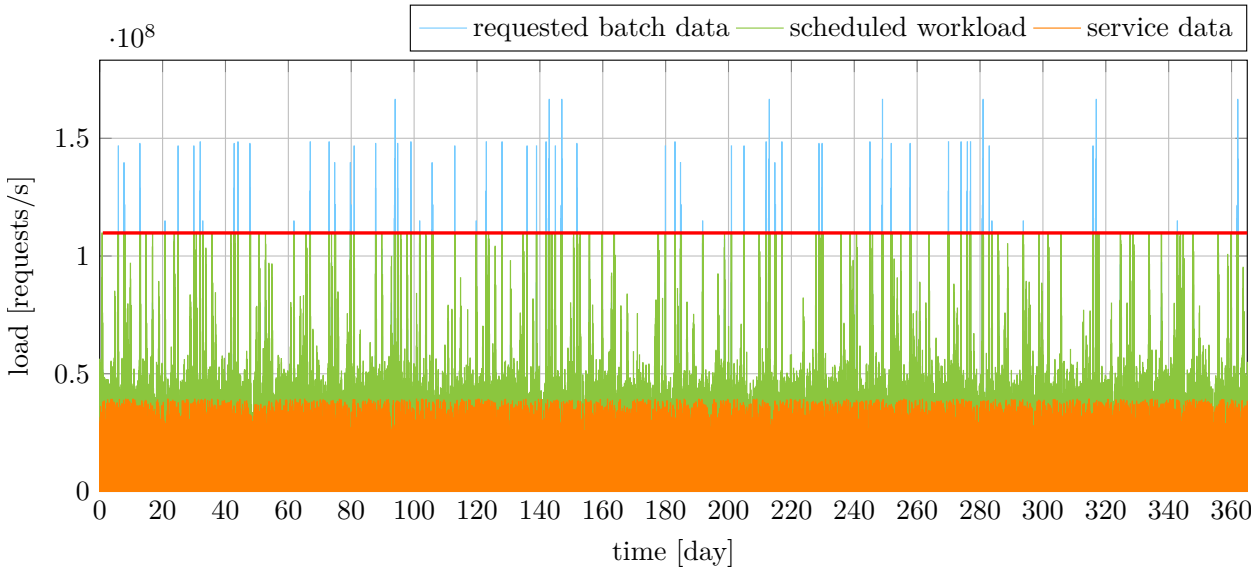

Figure 2: Example of the load (expressed as a histogram in number of requests per seconds) over time (in hours) which is executed after the IT sizing. The requests for services are in orange, the ones for batches in blue. It results in the green processed requests using 1098 servers considering one year discretized into $K = 8760$ time slots of one hour and a flexibility of 3 hours.

equal to 0 otherwise. This power demand D_k can be expressed as follow:

$$D_k = \frac{p}{nbI}\left(\sum_{i=1}^{r} ws_{i,k} + \sum_{i=1}^{n} wt_{i,k}^{sch}\right) \times PUE \quad (1)$$

The power profile $\mathcal{D}$ is then given by the set of all the power demands made at time slot k ($1 \leq k \leq K$) such that $\mathcal{D} = \{D_1, \ldots, D_K\}$.

The rest of this section aims at defining the power supply models that take $\mathcal{D}$ and weather conditions as inputs and define electrical devices needed to meet the data center power demand despite the renewable energy source intermittency.

4.2. Power supply models

The role of the power supply part of the data center infrastructure is to provide the computing part with electricity. The power supply models aim at describing the electrical architecture that has to be defined and sized to build a standalone data center without connection to the classical power grid. Table 2 summarizes the main notations used in the electrical part and in the rest of the paper. The infrastructure consists of (i) primary sources that supply the IT part of the data center with renewable energy sources such as wind and sun, and (ii) secondary sources that are back up power devices whose purpose is to provide power to servers when the renewable energy is not sufficient or to store energy otherwise.

As the data center is autonomous in terms of power supply, the connection to the classical power grid does not exist. So, in order to achieve its IT server power demand, the on-site power supply architecture of the data center only consists of wind turbines and photovoltaic panels to produce electricity from wind and sun, and batteries and a hydrogen system (electrolyzers, fuel cells and hydrogen tank) to assure the balance of the intermittency of the primary sources.

Due to the seasonal (long-term) and the daily (short-term) variations of the weather conditions (wind and sun), we decide (i) to dedicate the battery usage to the day and night alternation – the hours of overproduction will balance the hours of underproduction during the same day (e.g. the production will be smoothed over the day); (ii) to use the hydrogen system to balance underproduction days with overproduction days (e.g. the production will be smoothed over the season).

The role of the power supply is to satisfy the power

Table 2: Main notations for power supply models

V_k	wind speed at time slot k $[m/s]$
$\mathcal{V}$	$=\{V_k, 1 \leq k \leq K\}$
I_k	solar irradiation at time slot k $[W/m^2]$
$\mathcal{I}$	$= \{I_k, 1 \leq k \leq K\}$
q	number of wind turbines
Pr	the WT rated power production of one WT [W]
Pw_k	WT power prod. at time slot k [W]
Pw	$= \{Pw_k, 1 \leq k \leq K\}$
Apv	surface of the whole PV $[m^2]$
η_{pv}	PV efficiency
Ppv_k	PV power production at time slot k [W]
Ppv	$= \{Ppv_k, 1 \leq k \leq K\}$
Pre_k	renewable power production $Pw_k + Ppv_k$ at time slot k [W]
Pre	$= \{Pre_k, 1 \leq k \leq K\}$
BC_k	battery capacity at the end of the time slot k with $BC_0 = BCinit$ [Wh]
BC	capacity of the batteries
η_{ch}	battery charging efficiency
η_{dch}	battery discharging efficiency
Pch_k	charging power of the batteries during time slot k [W]
$Pdch_k$	discharging power of the batteries during the time slot k [W]
LOH_k	level of H_2 in the tank at the end of the time slot k with $LOH_0 = LOHinit$ [kg]
LOH	hydrogen tank capacity
η_{ez}	electrolyzer charging efficiency
η_{fc}	fuel cell discharging efficiency
Pez_k	charging power of electrolyzers during time slot k [W]
Pef_k	discharging power of fuel cells during time slot k [W]

demand $\mathcal{D}$ of the IT along the time horizon $\mathcal{H}$ discretized into K time slots. It is necessary to take the weather conditions of the data center location into account. The weather conditions are characterized by the solar irradiation $I_k \in \mathcal{I} = \{I_1, \ldots, I_K\}$ and the wind speed $V_k \in \mathcal{V} = \{V_1, \ldots, V_K\}$ for every time slot k $(1 \leq k \leq K)$ of $\mathcal{H}$. The goal of the design of the power architecture is to define the primary and secondary sources: number of wind turbines, surface area of photovoltaic panels, maximal power of both electrolyzers and fuel cells as well as batteries and hydrogen tank capacities.

Let q be the number of wind turbines of the same type (homogeneous wind turbine architecture) and Pr their rated power. Their averaged output power production Pw_k of time slot k depends on the wind speed V_k for all k $(1 \leq k \leq K)$. Let $Pw = \{Pw_1, \ldots, Pw_K\}$ be the power production of one wind turbine within the horizon $\mathcal{H}$. A turbine starts at the "cut-in" wind speed Vci, generating a power linearly increasing with wind speed from Vci to the rated wind speed Vr. When the wind speed varies between Vr and the "cut-out" wind speed Vco, the turbine produces a constant rated power Pr as an output electrical power. Once the wind speed exceeds Vco, the turbine stops generating for safety reasons. Thus, the power production Pw_k of a wind turbine at each time slot k is obtained using formula (2) $(1 \leq k \leq K)$:

$$Pw_k = \begin{cases} 0 & \text{if} \quad V_k \leq Vci \\ Pr\dfrac{V_k - Vci}{Vr - Vci} & \text{if} \quad Vci < V_k \leq Vr \\ Pr & \text{if} \quad Vr < V_k \leq Vco \\ 0 & \text{if} \quad Vco < V_k \end{cases} \tag{2}$$

Let Apv be the surface area of homogeneous photovoltaic panels and η_{pv} their associated efficiency. The averaged power produced Ppv_k by a surface of photo-voltaic panels A_{pv} at time step k is computed using formula (3) for all k $(1 \leq k \leq K)$:

$$Ppv_k = I_k \times Apv \times \eta_{pv} \tag{3}$$

Let BC_k be the capacity of the batteries at the end of the time slot k $(1 \leq k \leq K)$. It represents a given energy level in Wh. Let $BC_0 = BCinit$ be the initial battery capacity at the beginning of the time horizon $\mathcal{H}$. BC_k depends on the previous capacity of the battery BC_{k-1}, for all $1 \leq k \leq K$, and the level of charge $Pch_k \times \Delta t$ or discharge $Pdch_k \times \Delta t$ during time slot k (with respective efficiencies η_{ch} and η_{dch}). Considering the fact that we assume that batteries are dedicated to daily balance the renewable energy over- and under-production, the state of charge is cyclic and returns to the same level every 24 hours (every midnight for instance). The consequence of this assumption is that the self discharge of batteries within the duration of one day is so small that it can be neglected, even if this discharge remains within the model. Moreover, considering one time slot k, if $Pch_k \neq 0$, $Pdch_k = 0$ and *vice versa* (i.e., no charge and discharge at the same

time slot). Formula (4) allows to compute the battery capacity BC_k for each time slot k ($1 \leq k \leq K$) with $BC_0 = BCinit$. With all these values we can totally define the battery operations by computing the greatest amplitude BC of BC_k as well as the greatest needed charge PCH and discharge $PDCH$ powers over one day within $\mathcal{H}$. The details of the computation of the battery sizing is given in Section 5.2.3.

$$BC_k = BC_{k-1}(1-\alpha) + \Big(\eta_{ch} Pch_k - \frac{Pdch_k}{\eta_{dch}}\Big)\Delta t \quad (4)$$

where α is the self discharge rate.

Let LOH_k be the level of hydrogen in the tank at the end of time slot k, for all k ($1 \leq k \leq K$). It represents a given hydrogen mass in kilogram [kg]. Let $LOH_0 = LOHinit$ be the initial level of hydrogen at the beginning of the time horizon $\mathcal{H}$. LOH_k depends on the previous level of hydrogen in the tank LOH_{k-1}, for all k ($1 \leq k \leq K$). Given the time slot k ($1 \leq k \leq K$), given the electrolyzers' charging power Pez_k and the fuel cells' discharging power Pfc_k, the H_2 mass density ρ ($33\,\mathrm{kWh.kg^{-1}}$), the levels of charge and discharge are respectively $\eta_{ez} Pez_k \times \Delta t/\rho$ and $Pfc_k \times \Delta t/\eta_{fc}/\rho$. Considering the fact that hydrogen is dedicated to balance the seasonal renewable energy production, we assume that the level of hydrogen is cyclic and returns to the same level at the end of the considered time horizon $\mathcal{H}$. Formula (5) allows to give the level of hydrogen LOH_k for each time slot k ($1 \leq k \leq K$) with $LOH_0 = LOHinit$:

$$LOH_k = LOH_{k-1} + \Big(\frac{\eta_{ez} Pez_k}{\rho} - \frac{Pfc_k}{\eta_{fc}\rho}\Big)\Delta t \quad (5)$$

Both, IT and power supply models are used in the following in order to determine different data center sizing configurations. A data center sizing configuration corresponds to the number of servers, the number of wind turbines, the photovoltaic panel surface, and the sizing of the short-term and long-term storage devices (size and power) that allow to meet the IT power demand $\mathcal{D}$ of the data center given the level of QoS requirements of the data center. The next section describes the sizing strategy for the IT and the power supply parts.

5. Sizing Methodology

5.1. IT infrastructure sizing

The IT sizing approach proposed here is based on a scheduling algorithm of a workload consisting of malleable tasks and rigid services. This case provides a good illustration of the workload adaptation. The methodology remains the same if the scheduling algorithm is replaced by another version.

The aim of the proposed IT sizing process is to find the minimum number of servers or machines m necessary to fully respect the quality of service (QoS) for both the set of services and the set of tasks, i.e., to execute the set of tasks according to the flexibility δ. In the following equations, wt_k^{req}, wt_k^{sch} and ws_k represent respectively the amount of work that has to be performed by the IT platform concerning the tasks before the scheduling process, the amount of work that will be actually performed after the scheduling process decisions, and the services for every time slot k ($1 \leq k \leq K$) within the time horizon $\mathcal{H}$. This amount of work is a number of instructions in MI (millions of instructions).

$$wt_k^{req} = \sum_{i=0}^{n} wt_{i,k}^{req} \qquad \forall k \text{ s.t. } 1 \leq k \leq K \quad (6)$$

$$wt_k^{sch} = \sum_{i=0}^{n} wt_{i,k}^{sch} \qquad \forall k \text{ s.t. } 1 \leq k \leq K \quad (7)$$

$$ws_k = \sum_{i=0}^{r} ws_{i,k} \qquad \forall k \text{ s.t. } 1 \leq k \leq K \quad (8)$$

The QoS implies that the completion time of any scheduled task can not be delayed more than δ time slots.

5.1.1. Motivating example

Figure 3 shows on the left the incoming workload $\mathcal{W} = \mathcal{T} \cup \mathcal{S}$. The load is expressed as a number of instructions for each time slot k, $k = 1, \ldots, 9$. The red bar represents the available computing capacity ($maxW$) of the IT platform such that $maxW = m \times nbI$.

The right side of Figure 3 illustrates how the amount of work can be moved from the requested

time slot to the scheduled time slot when it is possible to respect the flexibility δ, i.e., for all tasks T_i, $c_i^{sch} - c_i^{req} \leq \delta$. The chosen workload example can be executed on the considered IT platform with $\delta = 2u.t.$ because after the scheduling process, the whole amount of work never exceeds the constraint $maxW$ value and respects by construction the QoS.

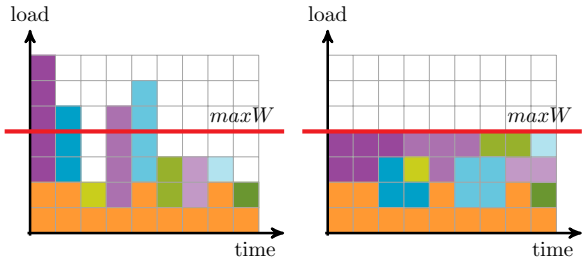

Figure 3: Workload before (wt_k^{req}), left) and after (wt_k^{sch}, right) the scheduling process considering a flexibility $\delta = 2u.t.$. Colored boxes are the tasks and service demands for each $u.t.$. The orange ones represent the (inflexible) services.

The next section presents the IT sizing methodology based on the principles introduced here.

5.1.2. Optimal IT sizing algorithm

Considering a given workload $\mathcal{W}$ and a given flexibility δ, we propose an optimal approach to minimize the number of servers needed to proceed this workload. Algorithm 1 is a binary search approach that is able to converge on the smallest IT platform size. The optimal number of machines m can be achieved using this algorithm such that $minM \leq m \leq maxM$ and its complexity is $O(\log_2(maxM - minM + 1))$. These lower and upper bounds, whose values are respectively given by Equations (9) and (10), are as tight as possible as proven respectively by Lemmas 5.1 and 5.2. The corollary of these two lemmas is that the number of iterations to reach the optimal value for m is minimal (see Corollary 5.2.1).

$$minM = \max_{1 \leq k \leq K} \lceil ws_k / nbI \rceil \tag{9}$$

$$maxM = \max_{1 \leq k \leq K} \lceil (wt_k^{req} + ws_k) / nbI \rceil \tag{10}$$

Lemma 5.1. *minM is the largest lower bound of the number of machines needed to complete a given workload $\mathcal{W}$ considering a malleable execution model.*

Proof. Services consist of instructions that cannot be deferred in time. This means that it is not possible to delay their execution to another time slot. Conversely, the execution of tasks can be deferred in time and then in the most extreme case, it can be considered that there are no more tasks to be executed within a given time slot k. In this context the minimum amount of work that has to be done by the IT infrastructure is ws_k at each time slot k ($1 \leq k \leq K$). The minimum infrastructure size (number of servers $minM$) is then given by the time slot k in which the amount of work devoted to services is the highest. As one machine can only proceed nbI instructions within one time slot, $minM$ is given by Equation (9).

Moreover m can be equal to $minM$ in the case where $h = \text{argmax}_{1 \leq k \leq K} \lceil ws_k / nbI \rceil$ and $wt_k + ws_k \leq ws_h$ with $h \neq k$ and $1 \leq k \leq K$. As cases could exist such that $m = minM$, then $minM$ is the largest lower bound for m. This concludes the proof. □

Lemma 5.2. *maxM is the smallest upper bound on the number of machines needed to complete a given workload $\mathcal{W}$ considering a malleable execution model.*

Proof. $wt_k^{req} + ws_k$ is the amount of work at the workload submission time of any time slot k ($1 \leq k \leq K$). In a case where it is not possible to move instructions from one time slot to another, the largest amount of work that determines m is given by $m = maxM = \max_{1 \leq k \leq K} \lceil (ws_k + wt_k^{req}) / nbI \rceil$.

The scheduling step aims at ensuring that the workload always remains below $maxM$. To guarantee that, the amount of work executed in time slot k has to decrease according to the flexibility and the scheduling strategy.

So m should be less than $maxM$ as $wt_k^{rep} \geq wt_k^{sch}$ ($1 \leq k \leq K$). This concludes the proof. □

Corollary 5.2.1. *The number of iterations of Algorithm 1 to reach its results is minimal.*

Proof. m is obtained by a binary search algorithm such that $minM \leq m \leq maxM$ in $\lceil \log_2(maxM - minM + 1) \rceil$ iterations in the worst case. As the bounds $minM$ and $maxM$ on m are respectively as large and as small as possible (see Lemmas 5.1 and 5.2), the number of iterations is minimal. This concludes the proof. □

Algorithm 1: Binary search based IT sizing algorithm that gives the minimum number of servers to complete a given workload $\mathcal{W}$ considering a given constant flexibility δ

Input: $\mathcal{W} = \mathcal{T} \cup \mathcal{S}$: workload to complete
Input: p: the power consumption of one server
Input: δ: the QoS such that $c_i^{sch} - c_i^{req} \leq \delta$ with $1 \leq i \leq n$

1 **Function** `IT_sizing`$(\mathcal{W}, nbI, \delta)$**:**
2 $u \leftarrow maxM$
3 $l \leftarrow minM - 1$
4 **while** $u - l > 1$ **do**
5 $m \leftarrow \lfloor (u + l)/2 \rfloor$
6 $valid \leftarrow$ `test_IT`$(\mathcal{W}, m \times nbI, \delta)$
7 **if** *valid* **then** $u \leftarrow m$
8 **else** $l \leftarrow m$
9 **return** u

This binary search algorithm is based on Algorithm 2 that returns *true* if it is possible to find a schedule of $\mathcal{W}$ using only m machines and *false* if not. As services cannot be delayed, we only move tasks. The principle of Algorithm 2 is to consider each time slot k of $\mathcal{H}$ which is defined by $\mathcal{W}$ and whose amount of work wt_k as it has been requested. As it is not possible to delay tasks from the last time slot, we start from the second last one ($k = K - 1$) (line 6) and we finish backward with time slot $k = 1$ if the scheduling process guarantees the flexibility and if the workload execution using only m machines is possible. Given the time slot k, we seek to move as much task instructions from time slot k to time slot $h = k+\delta$ (limited by K (line 8)). If there is enough room (line 10) on time slot h to execute an additional amount of charge wt_k, wt_k is moved to that time slot and we can consider the next time slot $k - 1$ with now $wt_k = 0$. Otherwise, the amount of charge of the targeted time slot h is fully filled using all the available room (lines 11 and 12) and the rest of charge of wt_k that has not been transferred yet (line 13) is considered now to fill the previous time slot $h-1$ (line 14), and so on until h remains strictly greater than k. Then, the new computing quantity wt_k from time slot k is now less than or equal to its value before this scheduling step. Therefore, if $wt_k + ws_k$ remains larger than the computing capacity $maxW$ of the m machines within one time slot (line 15), the scheduling process can stop and returns false. On the other hand, if the process goes to its conclusion, i.e., the amount of work from time slot $k = 1$ is less than the available computing capacity $maxW$, m machines is enough to perform $\mathcal{W}$ respecting the given flexibility δ. Algorithm 2 is then returning true. Depending on the response, the binary search algorithm can test another configuration with more or less machines as before until converging to the smallest value for m. wt_k can be considered as wt_k^{sch} at the end of the process. Finally note that, by construction, instructions that belong to wt_{k1} and wt_{k2} with $k1 < k2$ are proceeded respectively on the two time slots $k1'$ and $k2'$ such that $k1' \leq k2'$.

Algorithm 2: Algorithm to check whether $\mathcal{W}$ can be scheduled without ever exceeding a certain level of work $maxW$ regardless of the time slot of $\mathcal{H}$, taking into account a given constant flexibility δ and the IT models

Input: $\mathcal{W} = \mathcal{T} \cup \mathcal{S}$: the workload to complete; $maxW$: maximal possible amount of work [MI]; δ: the flexibility in *u.t.*

1 **Function** `test_IT`$(\mathcal{W}, maxW, \delta)$**:**
2 $wt_k \leftarrow \sum_{i=0}^{n} wt_{i,k}^{req}\ \forall k\ 1 \leq k \leq K$
3 $ws_k \leftarrow \sum_{i=0}^{r} ws_{i,k}\ \forall k\ 1 \leq k \leq K$
4 possible $\leftarrow$ true
5 $K \leftarrow$ number of time slots of $\mathcal{H}$
6 $k \leftarrow K - 1$
7 **while** $k \geq 1$ & *possible* **do**
8 $h \leftarrow \min(k + \delta, K)$
9 **while** $h > k$ & $wt_k > 0$ **do**
10 room $\leftarrow \max(maxW - ws_h - wt_h, 0)$
11 work2Move $\leftarrow \min(\text{room}, wt_k)$
12 $wt_h \leftarrow wt_h + \text{work2Move}$
13 $wt_k \leftarrow wt_k - \text{work2Move}$
14 $h \leftarrow h - 1$
15 possible $\leftarrow wt_k + ws_k \leq maxW$
16 $k \leftarrow k - 1$
17 **return** possible

Theorem 5.3. *Algorithm 1 returns the smallest value for the number m of machines that is able to complete the workload $\mathcal{W}$ with the flexibility δ.*

Proof. Algorithm 1 is a binary search that is able to access each integer value $m \in [\![minM, maxM]\!]$ even if the solution is either $minM$ or $maxM$. The key point of this algorithm is the test function `test_IT`$(\mathcal{W}, maxW, \delta)$ that is able to know if a schedule is possible to execute the workload $\mathcal{W}$ using only m machines or not ($maxW = m \times nbI$).

Algorithm 2 aims at delaying tasks as late as possible, i.e., at most δ *u.t.* This is a scheduling at the latest, called $\mathcal{L}$-scheduling in the following.

If such a delay is possible, the maximum amount of work $maxW$ of the target time slot is not exceeded and instructions are moved from their requested time slot to the scheduled time slot (line 12 and line 13). If possible = *true* at line 17, it means that the IT platform has sufficient computing capacity $maxW$ to execute $\mathcal{W}$ with respect to δ.

Let $\mathcal{W} = \mathcal{T} \cup \mathcal{S}$ be a workload successfully scheduled with the computing capacity $maxW$ of the considered IT platform, but without using the $\mathcal{L}$-scheduling approach.

Then $\sum_{i=1}^{n} wt_{i,k}^{req} + \sum_{i=1}^{r} ws_{i,k} \leq maxW$ for every time slot k $(1 \leq k \leq K)$. Carry over a set of instructions from one time slot k to another if there is room, and if the duration between the end of their requested time slot and the end of the target time slot does not exceed the flexibility δ of $\mathcal{W}$ and respects the computing capacity. So, using a $\mathcal{L}$-scheduling does not change the value returned by the `test_IT` function.

On the other hand, if the computing needs of one time slot k exceed $maxW$ of the platform, the computing capacity is not sufficient for executing that workload. Indeed, if at least one instruction I can not be deferred to another time slot $k+\delta, \ldots, k+1$, then $wt_h^{sch} + ws_h$ is equal to $maxW$ for any h such that $k+1 \leq h \leq k+\delta$. The only possibility should be to delay instructions in a given time slot h to make room for the instruction I. But all instructions of time slot h $(k+1 \leq h \leq k+\delta)$ have already been delayed as much as possible because the $\mathcal{L}$-scheduling begins by the end $(k = K-1, k = K-2, \ldots, k = 1)$.

So if there is no room to move extra instructions from a time slot k to any time slot h $(k+1 \leq h \leq k+\delta)$, it means that it does not exist any schedule using only a computing capacity $maxW$ that respects the flexibility δ to execute in time such a workload. Algorithm 2 returns *false* if and only if there does not exist a schedule to execute $\mathcal{W}$ in such a computing capacity $maxW$ and a given flexibility δ.

Finally, the binary search is based on a test function that returns *true* when the platform is large enough and *false* only if the platform is not large enough. m is then the smallest possible value for the platform. This concludes the proof. □

The next section proposes a methodology to design the power supply architecture that is able to deliver electrical power when the IT part of the data center needs it. Using $wt_k = wt_k^{sch}$ also as a result of the IT sizing process given by Algorithm 1, it is possible to easily compute D_k for all k $(1 \leq k \leq K)$. Indeed $wt_k^{sch} = \sum_{i=1}^{n} wt_{i,k}^{sch}$. We recall that $\mathcal{D}$ is one of the inputs of the power supply sizing process.

5.2. Power supply sizing

In this section, the sizing methodology dedicated to the electrical part of the data center is presented. The sizing of the power supply architecture depends on different inputs: the data center power demand $\mathcal{D}$, the weather conditions that give the solar irradiation $\mathcal{I}$ and the wind speed $\mathcal{V}$ over the same time horizon $\mathcal{H}$ including K time slots with duration Δt $(\mathcal{H} = K\Delta t)$. Our methodology aims at finding the appropriated sizing for each element that composes the power supply system of the standalone data center: wind turbines, photovoltaic panels, batteries, and hydrogen system. The methodology consists in determining first the primary sources – number of wind turbines and surface area of the photovoltaic panels – and then designing the short- and long-term storage devices to reach the power demand $\mathcal{D}$ considering power supply models. The power supply of the data center is then only based on the renewable power production during each time slot $Pre_k = Pw_k + Ppv_k$ $(1 \leq k \leq K)$, the storage facilities being introduced to compensate for the inherent intermittency of sun and wind.

The next section presents the rule of the game dedicated to manage the power supply system at the different timescales: short- and long-term.

5.2.1. Daily power supply management

During one day d $(1 \leq d \leq 365)$ with Λ time slots (as $\Delta t = 1\,h$ in practice, $\Lambda = 24$), considering a given power demand $\mathcal{D}$ and a given primary source architecture ($nbWT$ wind turbines and a surface area Apv of photovoltaic panels), it is possible to know if there is over or under renewable power production for any time slot k of this day $(1 + (d-1)\Lambda \leq k \leq d \times \Lambda)$.

The rule of the game is that batteries aim at balancing day/night power production. To ensure it,

we assume that their level of charge has to be the same at the beginning of each day. This initial level of charge is defined during the sizing process to prevent the shortage of the batteries.

Let Op_d and Up_d be respectively the amount of over produced energy and the under produced energy during the day d. Equations (11) and (12) allow to compute Op_d and Up_d:

$$Op_d = \sum_{k=1+(d-1)\Lambda}^{d\times\Lambda} \mathbb{1}_{[D_k,+\infty[}(Pre_k)(Pre_k - D_k)\Delta t \tag{11}$$

$$Up_d = \sum_{k=1+(d-1)\Lambda}^{d\times\Lambda} \mathbb{1}_{[0,D_k[}(Pre_k)(D_k - Pre_k)\Delta t \tag{12}$$

where $Pre_k = Pw_k + Ppv_k$ is the renewable power production during the time slot k $(1 + (d-1)\Lambda \leq k \leq d \times \Lambda)$, and where $\mathbb{1}_{[D_k,+\infty[}(Pre_k)$ and $\mathbb{1}_{[0,D_k[}(Pre_k)$ are the two indicator functions of Pre_k respectively for the Op_d and Up_d expressions ($\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 if $x \notin A$).

A day d is qualified as an overproduction day if the following condition (13) is true:

$$Op_d \times \eta_{ch} \times \eta_{dch} \geq Up_d \tag{13}$$

Otherwise d is an underproduction day when the following condition (14) is true in turn:

$$Op_d \times \eta_{ch} \times \eta_{dch} < Up_d \tag{14}$$

Depending on the over- or under-production of a day, the rule of the battery usage is not the same:

- Overproduction day: batteries aim to supply the servers of the IT part in addition to WT and PV if needed to meet the power demand (light green part below the power demand, the red line in Figure 4 being constant for illustration purpose). Conversely, batteries are charged as quickly as possible as soon as there is sufficient renewable power production to bring the state of charge to the same level as at the beginning of that day as shown in Figure 4 by the deep green part above the power demand. The hydrogen system, i.e., electrolyzer (EZ), has to store the rest of the power overproduction in form of H_2 (blue part above the power demand, the red line in Figure 4);
- Underproduction day: unlike the day of overproduction, batteries are charged by using power overproduction of every overproduction time slot (deep green part in Figure 5) and supply the IT part in addition to primary sources and fuel cells (FC) as quickly as possible and as soon as possible as shown in Figure 5 by the light green part. Then FC take over when the batteries have supplied the amount of energy equivalent of the overproduction after considering charge and discharge efficiencies.

Note that in both cases, battery efficiencies, to charge (η_{ch}) or discharge (η_{dch}), have to be taken into account. Figures 4 and 5 summarize both cases described before giving the rules of the game observed to use short term and long term storage devices within a given day. Deep green and light green parts take efficiencies into account. These rules are the basis for the algorithm used to determine the amplitude of the battery state of charge each day and the level of hydrogen at the end of the time horizon $\mathcal{H}$. The largest battery amplitude gives the battery capacity BC for the short term storage device and the amplitude of the level of H_2 gives the size of the tank of hydrogen of the system. These values are obtained considering a given power demand $\mathcal{D}$ and the given renewable power production allowed by weather conditions and the primary architecture.

Let `storageSizing(`$\mathcal{D}$`,`Pre`)` be such an algorithm that returns the level of hydrogen at the end of $\mathcal{H}$. This algorithm is used day after day to size the primary sources such that the level of hydrogen LOH_K at the end of the time horizon has to be greater than or equal to its level at the beginning LOH_0 but as close as possible. The sizing of the storage devices is given in the subsequent section.

5.2.2. Sizing of the primary sources using a binary search approach

Primary sources consist of photovoltaic panels (PV) and wind turbines (WT). As this architecture is homogeneous (only one type for PV and WT), the number of configurations is not combinatorial. Primary sources aim at collaborating to supply the

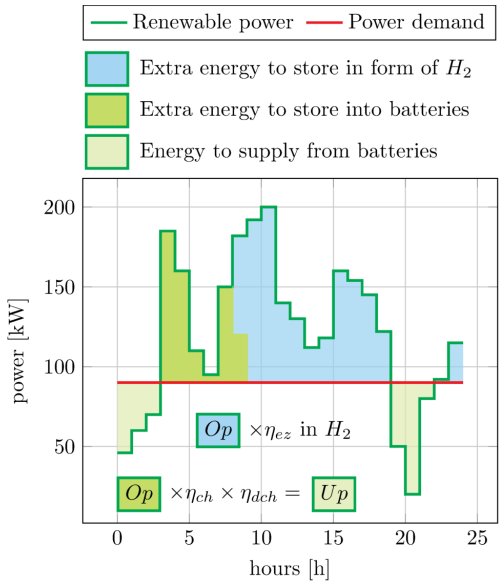

Figure 4: Rule of the game of an overproduction day for the usage of batteries and the hydrogen system. Power demand is constant for illustration purpose. The whole renewable production that does not meet the power demand is supplied thanks to batteries that are recharged using a fraction of the overproduction.

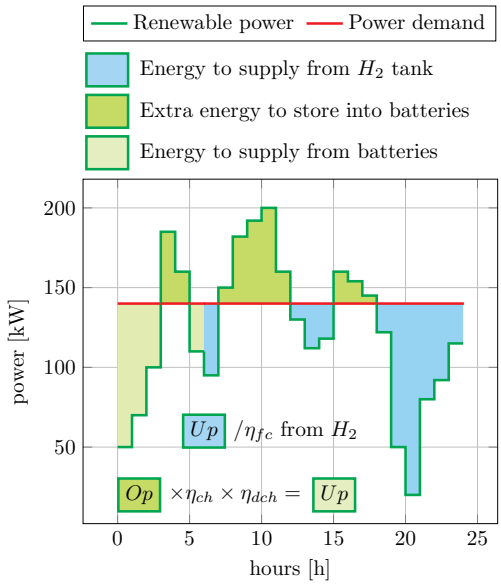

Figure 5: Rule of the game of an underproduction day for the usage of batteries and the hydrogen system. Power demand is constant for illustration purpose. The whole power production that exceeds the power demand is stored into batteries so as to partially compensate the underproduction.

power needed $\mathcal{D}$ by the data center to complete its computation demand, i.e., the workload $\mathcal{W}$. As mentioned before, both power production of photovoltaic panels and wind turbines are added to supply the data center and the back up power devices. The size of the primary sources has to be large enough to reach overproduction in order to allow sufficient energy storage for time slots where primary power production is not sufficient to power the entire data center. The primary sources must compensate for the day/night alternation as well as seasonal variations by a necessary overproduction.

Considering a given power demand $\mathcal{D}$ and weather conditions $\mathcal{I}$ and $\mathcal{V}$ for one given year, the principle is to find the appropriate primary source set that allows to end the year with a hydrogen level (LOK_K) that has to be greater than or equal to the hydrogen level at the beginning of the same year (LOH_0), i.e., LOH_K and LOH_0 are as close as possible to make the supply for another year possible. Hence, the following condition has to be respected:

$$LOH_K - LOH_0 \geq 0 \tag{15}$$

Let $maxWT$ be the minimal number of WT that does not need any PV to respect Equation (15). Then, there are only $maxWT+1$ different configurations including respectively $maxWT$, $maxWT-1$, $maxWT-2$, ..., 1, 0 WT. The $maxWT$ last configurations are complemented by the minimum surface area of photovoltaic panels that allow to respect Constraint (15).

Computation of $maxWT$. The data center energy need (E_{DC}) can be evaluated by computing the integral of the power demand $\mathcal{D}$ on the considered time horizon $\mathcal{H}$. By choosing historical weather conditions $\mathcal{V}$ on the same period of time, it is possible to know which quantity of energy E_{1WT} one WT is able to produce. Indeed, $maxWT$ is greater than or equal to the upper integer part of the ratio between E_{DC} and E_{1WT}:

$$maxWT \geq \left\lceil \frac{E_{DC}}{E_{1WT}} \right\rceil \tag{16}$$

$$E_{DC} = \sum_{k=1}^{K} D_k \times \Delta t \tag{17}$$

$$E_{1WT} = \sum_{k=1}^{K} Pw_k \times \Delta t \tag{18}$$

Since there is no reason for the data center power demand to coincide with the wind power, it is necessary to store a part of the produced energy. But a significant part of wind energy is lost due to the efficiency of the storage process. This loss can be covered rounding up the ratio given above up to the next integer value. If this is not enough, and depending on the type of wind turbine, another turbine may be necessary until Condition (15) becomes true. The configuration with only $maxWT$ WT, without PV, is the first configuration. The remaining configurations are given in the next paragraph.

The maxWT other configurations. Let $Q = \{Q_0, Q_1, \ldots, Q_{maxWT-1}, Q_{maxWT}\}$ the set of possible configurations that correspond to the $maxWT+1$ different options for the primary source configurations for the power supply architecture of the data center. The q^{th} configuration $Q_q = (q, Apv_q)$ consists of a surface area Apv_q in $[m^2]$ for the PV and q WT ($0 \leq q \leq maxWT$). Q_0 and Q_{maxWT} are two special cases with respectively no WT or no PV. To meet the data center power demand when the number of WT is less than $maxWT$, the point is to find the appropriate surface area of PV. By considering the same historical weather conditions as before (i.e., solar irradiation $\mathcal{I}$ and wind speed $\mathcal{V}$), a given surface area Apv_q of PV and the number q of WT of the q^{th} configuration of Q, it is possible to compute the overall renewable power production for any time slot k ($1 \leq k \leq K$):

$$Pre_k = I_k \times Apv_q \times \eta_{pv} + q \times Pw_k \tag{19}$$

Finally, by considering the daily power supply management rules (Section 5.2.1), it is possible to find the level of hydrogen LOH_K at the end of the horizon $\mathcal{H}$. If this level is higher than LOH_0, Apv_q is too large and vice-versa.

Algorithm 3 is a binary search algorithm that is able to find, for each possible number of WT, the smallest value for the surface area of PV that respects Constraint (15) and makes it possible to meet the data center power demand. This algorithm complexity is obviously logarithmic: $O(\log_2(maxApv))$ with $maxApv$ the largest possible surface area for PV. This algorithm returns the set Q of $maxWT+1$ configurations (q, Apv_q) including a number of q WT between 0 to $maxWT$ and the corresponding surface area values Apv_q of PV. Note that the configuration with $maxWT$ wind turbines does not contain any photovoltaic panels.

A value of $maxApv$ could be obtained by considering the surface area of PV required to produce the total amount of energy E_{DC} needed in the data center only using PV so that the renewable power production is not similar with the computer consumption.

A value of $maxApv$ could be obtained by considering the surface area of PV required to produce the total amount of energy E_{DC} needed in the data center only using PV taking into account the fact that the power production and the power consumption occurs at different times, hence the need to use the storage infrastructure.

In this worst case, energy production could be stored first in hydrogen using electrolyzers before being consumed by the data center using fuel cells. In this case $maxApv$ is given by Equation (20):

$$maxApv = \left\lceil \frac{E_{DC}}{E_{1PV} \times \eta_{ez} \times \eta_{fc}} \right\rceil \tag{20}$$

where E_{1PV} is the energy obtained by using $s_{pv} = 1\,m^2$ of PV during the time horizon $\mathcal{H}$.

$$E_{1PV} = \sum_{k=1}^{K} I_k \times s_{pv} \times \eta_{pv} \times \Delta t \tag{21}$$

Now, considering a given primary source configuration (x WT and a surface area APV_x of PV), it is possible to size the storage devices, batteries and hydrogen system.

5.2.3. Sizing of the storage system:

The strategy to design the storage system relies on computing time slots of overproduction and underproduction for each day during the time horizon. As a reminder of the rules of the storage usage given in Section 5.2.1, the batteries are used during the day to balance the hours of overproduction and underproduction (fluctuations between day and night) and the hydrogen system composed of electrolyzers

Algorithm 3: Computation of the set Q of possible configurations that respect Condition (15)

Input: $maxWT$, Pw, $\mathcal{D}$
Output: $Q = \{(0, Apv_0), \ldots, (q, Apv_q), \ldots\}$

Function `Electrical_sizing()`:
- $Q = \emptyset$
- **for** $q = 0$ *to* $maxWT - 1$ **do**
  - $u \leftarrow maxApv$
  - $l \leftarrow -1$
  - $LOH_0 \leftarrow LOHinit$
  - $LOH_K \leftarrow 0$
  - **while** $u - l > 1$ & $LOH_K \neq LOH_0$ **do**
    - $apv \leftarrow \lfloor (u + l)/2 \rfloor$
    - $Ppv \leftarrow \mathcal{I} \times apv \times \eta_{pv}$ (vector operation)
    - $Pre \leftarrow Ppv + q \times Pw$ (vector operation)
    - $LOH_K \leftarrow$ `storageSizing`$(\mathcal{D}, Pre)$
    - **if** $LOH_K < LOH_0$ **then**
      - $l \leftarrow apv$
    - **else**
      - $u \leftarrow apv$
  - $Q \leftarrow Q \cup \{(q, u)\}$
- **return** $Q \cup \{(maxWT, 0)\}$

and fuel cells is used to balance days of overproduction and days of underproduction (seasonal fluctuations).

Storage capacity. As the batteries operate on a day scale, the difference between the maximum and the minimum values of their capacity within the same given day determines the sizing of the battery for that day. The capacity BC of the batteries is equal to the maximum computed daily capacities:

$$BC = \max_{1 \leq d \leq K/\Lambda} (maxBC_d) \tag{22}$$

with $\forall h$ such that $1 + (d-1)\Lambda \leq h \leq d \times \Lambda$:

$$maxBC_d = \max_h(BC_h) - \min_h(BC_h) \tag{23}$$

As the hydrogen system operates on the seasonal scale, the sizing of the tank is equal to the difference between the maximum and the minimum level of hydrogen. This is expressed as follows for all k ($1 \leq k \leq K$):

$$LOH = \max_{0 \leq k \leq K} (LOH_k) - \min_{0 \leq k \leq K} (LOH_k) \tag{24}$$

As it is not possible to imagine before the sizing process the initial level for batteries and hydrogen tank to supply the data center using the mentioned input data, they are arbitrarily set to 0. But after the process it is then possible to set these initial levels, $BCinit$ at the beginning of each day and $LOHinit$ at the beginning of the year. They are computed as absolute values of the minimum values for BC_k (respectively LOH_k), $1 \leq k \leq K$, since at least one is necessary negative or null at that step of the process:

$$BCinit = \left| \min_{1 \leq k \leq K} BC_k \right| \tag{25}$$

$$LOHinit = \left| \min_{1 \leq k \leq K} LOH_k \right| \tag{26}$$

Power of storage devices. To complete the sizing of the storage system, batteries, and hydrogen system, the power required for each device has to be defined to be sure that the appropriate power is delivered when the renewable sources are able to meet the data center power demand. Considering the daily storage usage as defined by the rules of the game (Section 5.2.1), day after day for the entire duration of the time horizon $\mathcal{H}$, the nominal required power globally for each device (PCH and PDCH for the power the batteries, PEZ for electrolyzers, PFC for fuel cells) are:

$$\begin{aligned} PCH &= \max_{1 \leq k \leq K} (Pch_k) \\ PDCH &= \max_{1 \leq k \leq K} (Pdch_k) \\ PEZ &= \max_{1 \leq k \leq K} (Pez_k) \\ PFC &= \max_{1 \leq k \leq K} (Pfc_k) \end{aligned} \tag{27}$$

5.3. IT and Power supply sizing summary

As a result, the optimal number of servers is obtained from the IT workload. Then, the power supply sizing process is able to propose $maxWT + 1$ configurations: one only with wind turbines (full WT configuration); one only with photovoltaic panels (full PV configuration); the others with both primary sources (1 WT, 2WT, etc.). Each configuration is known by its number of wind turbines,

its surface area of photovoltaic panels, and for each of them the associated storage devices, power and capacity for the batteries as a short term storage device, power and hydrogen tank size for the hydrogen system as a long term storage device (electrolyzers and fuel cells).

In the remainder of the paper, experiments show the behavior and characteristics of these configurations to help decision makers to choose the appropriate configuration for a standalone data center only supplied with renewable energy.

6. Experiments

6.1. Input data

In this section, we explain the data used in the experiments:

Workload The workload $\mathcal{W}$ is generated following the data from user requests recorded during the Soccer World Cup in 1998 and available on the web site[1]. We have used the same methodology as in [44].

The days with a high load in the trace (days 45 to 79) are first selected, then the load for each of the 365 days of our workload is randomly chosen among those selected days. The flexibility of the tasks δ is set to 3 hours. Service load is generated with a uniform distribution requesting an equivalent of work in the 100-400 servers range;

Servers The servers are quad-core processors running at $2.5GHZ$ and consuming $350W$.

Weather conditions To simulate the power production of the primary sources (PV and WT), one needs to download meteorological data such as the irradiance $\mathcal{I}$ and wind speed $\mathcal{V}$ for one year. These data can be obtained online from various databases. In our case, the irradiance data is downloaded from the National Solar Radiation Database (NSRDB) [35], and the wind speed data is downloaded from the wind prospector from the National Renewable Energy Laboratory (NREL) [13]. These data are collected hourly, every day from 2004 to 2012. The chosen localization is Los Angeles with the coordinates: Latitude: 34.57; Longitude -118.02; Elevation 807. The selected year is 2004.

Power sources The input values of the primary sources used in the power supply sizing process are summarized in Table 3.

Table 3: Input values of the power supply sizing process

Notation	Value	Units
P_r	400	$[kW]$
V_r	14	$[m/s]$
V_{ci}	4	$[m/s]$
V_{co}	25	$[m/s]$
η_{fc}	0.6	–
η_{ez}	0.6	–
η_{ch}	0.8	–
η_{dch}	0.8	–
η_{pv}	0.15	–

6.2. Sizing of the IT infrastructure

The resulting sizing reaches 1098 servers with a maximum power consumption of 384.3 kW. The result is shown in Figure 2 with the initial workload (services in orange and batch tasks in blue) along with the shifted load (due to the limited number of servers) using Algorithms 1 and 2 which results in the actual load in green. The peak power consumption is 499.59 kW including the environmental consumption (PUE of 1.3).

6.3. Sizing of the power supply infrastructure

In the following, as an illustrative case, the chosen year of reference to size the power supply of the data center is 2004 using the IT workload $\mathcal{W}$ presented before. As shown in [18], mathematical models based on time series are suitable to model solar irradiation but are more difficult to apply to the wind. We have also shown that the obtained power supply sizing is different depending on the reference year. The choice of the year 2004 is led by the fact that during this reference year, the production of electrical power by the photovoltaic panels and the one by the wind turbines exhibit a complementarity. It emphasizes the effect of the hybridization of both primary sources. Indeed the power supply sizing depends on the weather conditions and then on the data center location. This work intends to show how the sizing process works and does not intend

[1] WorldCup'98 logs. http://ita.ee.lbl.gov/html/contrib/WorldCup.html

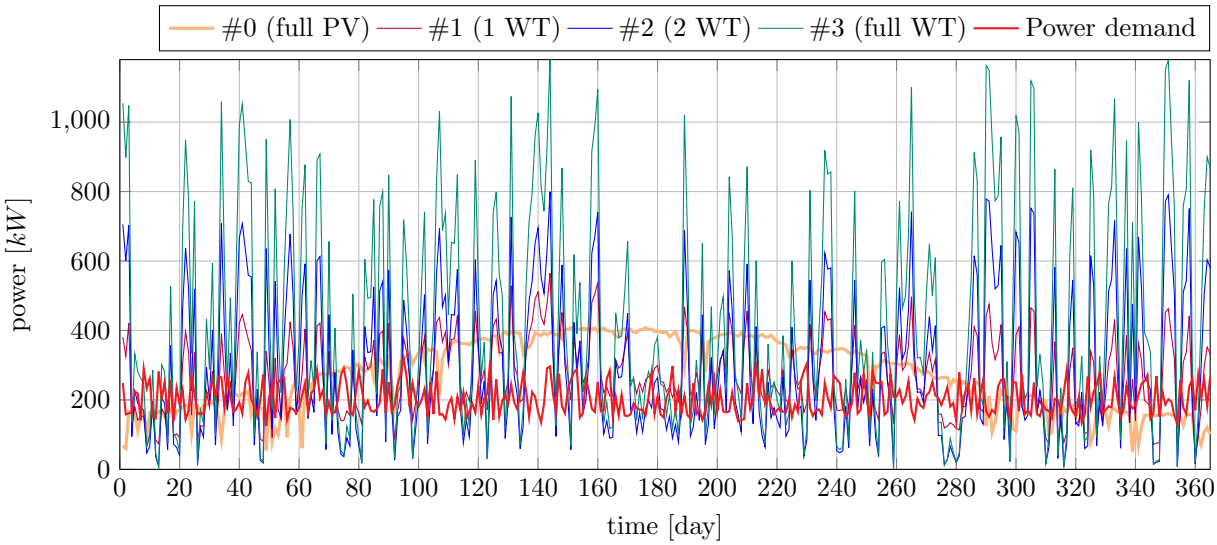

Figure 6: Renewable power production in 2004 for each possible configuration found by the sizing algorithm to meet the data center demand.

to infer general rules for the sizing itself. This latter choice stays with the decision maker who could iterate on this process to take his decision choosing between different locations and therefore weather conditions.

Considering one reference year for the weather conditions and a target workload demand, different configurations are obtained for primary and secondary sources. For the year 2004, Table 4 illustrates the possible configurations starting from a configuration with only photovoltaic panels (full PV), reaching a configuration with only wind turbines (full WT), and exploring configurations with both PV and WT. Each configuration is described by the number of WT, the surface area of PV, the power of electrolyzers, fuel cells and storage capacity (batteries and hydrogen tank).

Table 4: Four different sizings of the power supply based on the year 2004 and the workload of the 1998 soccer world cup servers. Configurations are indexed by the number of WT.

config	q	Apv_q [m^2]	BC [kWh]
#0	0	7258	5387
#1	1	3142	3050
#2	2	203	3889
#3	3	0	5613

config	PEZ [kW]	PFC [kW]	H_2 tank [kg]
#0	632	836	6296
#1	482	836	2884
#2	442	836	2216
#3	676	836	14643

As can be seen on Table 4, the choice of one primary source configuration (i.e., the surface of PV and the number of WT) has a huge impact on the secondary sources (i.e., the batteries and the hydrogen components).

6.3.1. Influence of the primary source configuration on the annual power production

Algorithm 3 has computed each possible primary source configuration to face the data center demand during the year 2004, considering the sun and the wind profiles. Figure 6 shows the power supplied by these primary sources, day after day for the four different configurations obtained, the power demand (in red) being the same. As expected, the power profile in the full PV configuration (#0) follows a bell shaped curve, due to the seasonal alternation of the day duration: the maximum is reached in summer and the lowest level in winter but from day to day, the evolution of the power production is smooth. On the contrary, in the full WT configuration (#3), the mean value of the power production is much more regular all along the year, even if a seasonality is sensible with a slight deficit of wind in summer (see Figures 7 and 8). But its variability is much higher from one day to another. In Configuration #1 and Configuration #2, the variability of the WT production dominates, but the higher the PV surface area, the lower the production peaks.

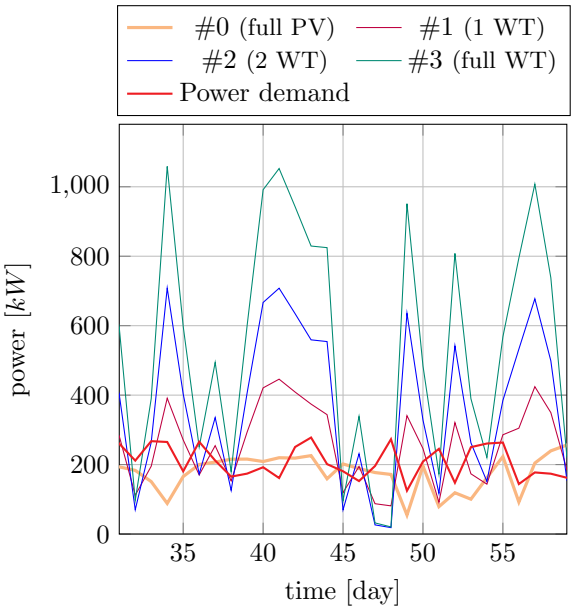

Figure 7: Renewable power production in February 2004 for each possible configuration found by the sizing algorithm to meet the data center demand.

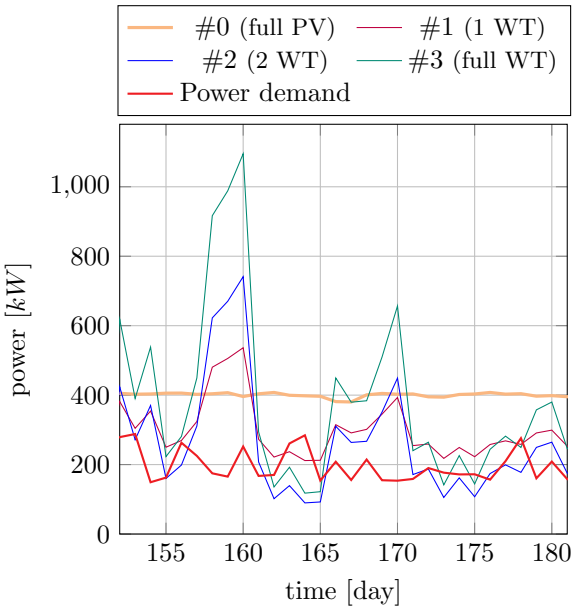

Figure 8: Renewable power production in June 2004 for each possible configuration found by the sizing algorithm to meet the data center demand.

6.3.2. *Influence of the primary source configuration on the battery sizing*

The rule governing the battery sizing is that its charge should return to its initial state at the end of each day. The initial state has been set to an intermediate value between fully charged and fully discharged, the reason being that it should be able to face any situation at dawn, a sunny day or a cloudy one. Then, the sizing of the battery depends on a daily variation. In the full PV configuration, the longer nights in winter are dominant in the sizing of the battery, i.e., the 172^{nd} day of the year at the summer solstice, leading to a 5387 kWh capacity. In the full WT configuration, the highest variation of the power production from one day to another governs it, leading to the largest capacity of 5613 kWh, being of the same order than Configuration #0. Configurations #1 and #2 lead to lower capacities as they take advantage of the complementarity of the primary production, the photovoltaic panels smoothing the daily variability of the WT and the wind power smoothing the seasonal variation of the PV production.

6.3.3. *Influence of the primary source configuration on the hydrogen component sizing*

It can be seen that the configuration has no influence on the sizing of the fuel cell. Indeed, for each configuration, there is one solely time slot in the IT workload which sets this value at the same peak power demand, according to Equation (27).

Concerning the hydrogen tank and the electrolyzer, their sizing is governed by two rules. First, all the renewable energy should be stored and cannot be lost. It means that the disconnection of a wind turbine or a PV module is not considered. Once the battery is fully charged, the electrolyzer has to convert all the excess primary power produced into hydrogen which should be stored in the hydrogen tank. Second, the level of the hydrogen tank should tend to go back to its initial level as imposed by Condition (15).

Figure 9 shows the evolution of the hydrogen storage during the year. The hydrogen storage of the full WT configuration diverges. As a matter of fact, the number of WT is set to comply with the demand of the data center but the power produced by one turbine cannot be modulated. This leads to an oversizing of the installation, as the third WT is used in place of only 203 m^2 of PV. As a consequence the constraint on the level of the hydrogen tank at the end of the year cannot be respected. To convert all the power overproduced by the oversized WT set, the electrolyzer sizing reaches the highest value of 676 kW.

In the case of the full PV configuration (#0), the constraint on the level of hydrogen back to its initial value is respected. The evolution of the storage follows the seasonal bell shaped curve of the solar production. The lowest storage level is reached at the winter solstice (i.e., 172^{nd} day of the year) and the highest at the summer solstice (i.e., 355^{th} day of the year). As the gap between the primary power available in summer and the power available in winter is rather high, the power need of the electrolyzer is of the same order as the full WT case (#3).

In Configuration #1, the amplitude of the storage is reduced compared to the full PV but the evolution is still dominated by the seasonality. In Configuration #2, the seasonality between summer and winter is almost damped and the constraint of the level of hydrogen back to the initial value is respected. In both cases, the power of the electrolyzer is about the same, about 30% reduced compared to full PV or full WT cases.

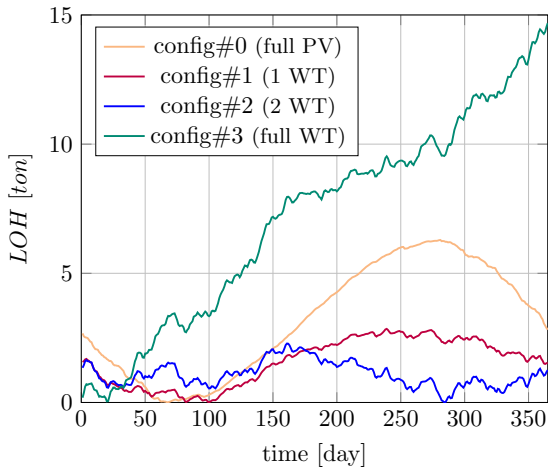

Figure 9: Evolution of the level of hydrogen in tons for each configuration day after day during year 2004.

6.4. Sizing assessment

In this section, in order to evaluate the sizing assessment, the same IT load as used in the previous section is shifted repeatably by one hour during a day until reaching a shift of 23 hours. As a result, we obtain 23 new IT power demands. These loads are used as inputs of the proposed sizing methodology.

6.4.1. 24 IT loads

First of all, in order to identify the nature of the considered workload (i.e., if the demand is greater during the night or during the day), the average of the load has been calculated per hour over the year. This computation provides the hourly average demand of power considering the whole year (Figure 10).

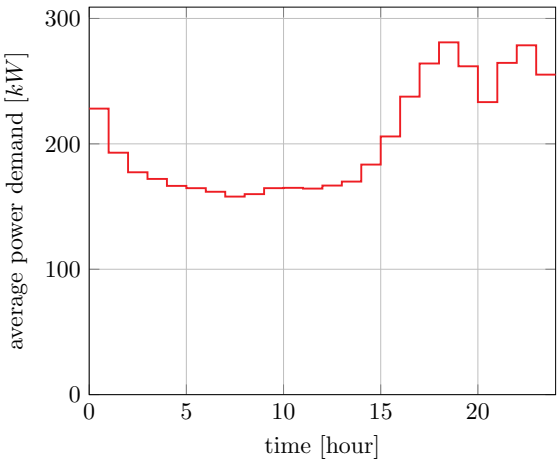

Figure 10: The average hours accumulated over the year of the IT load. Hour 0 corresponds to 0:00 am.

One can see that the majority of the power demand is executed after the second part of the afternoon. By shifting this workload hour by hour, the impact on the sizing of the power supply infrastructure is identified.

6.4.2. Power supply infrastructure

Based on the results obtained by Algorithm 3, one can see in Figure 11 that the configuration of the primary sources is hardly changed. As a matter of fact, the energy supplied to the load over the whole year is produced by the primary sources, i.e., the wind turbines and the photovoltaic panels. The shifting of the load has not changed significantly the global need of energy over the year, the impact is indirect due to the efficiency of the power conversion involving the storage. In Figure 11, it can be seen that the surface of PV is slightly decreased around the 9 hours shift in the full PV and the 1 WT configuration because it corresponds to a load profile following more or less the bell shaped profile of the solar illumination. The maximal number of wind turbines is submitted to Constraint (16): it remains equal to 3 because the variation of the annual energy load induced by the efficiency of the

storage conversions is much smaller than the yearly production of one additional wind turbine.

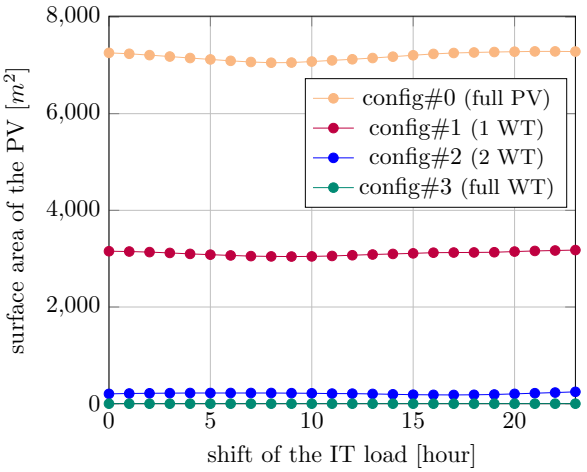

Figure 11: Evolution of the surface area of the photovoltaic panels as a function of the shift of the IT load.

In the same way, the capacity of the hydrogen tank is almost stable. Indeed, shifting the workload during the day only affects the daily but not the annual imbalance between demand and power supply. In Figure 12, one can see that the load shift has a strong impact on the battery sizing. In fact, from a shift of one hour to another, the battery capacity significantly changes. As expected, the highest impact concerns the full PV configuration when shifting the initial workload by 8 hours makes the workload following the bell shaped solar production, reducing the need for the daily storage. This synchronization is still reasonable for Configuration #1. The effect of the shifting on the battery sizing in Configurations #2 and #3 is lower but exists and is more complex to analyze as the wind profile is not as regular as the solar one. Nevertheless, there is a better synchronization between the workload and the wind production with a shifting beyond 10 hours.

To conclude, the variation of the IT load during the hours of the day has only an impact on the sizing of the battery capacity which balances the day and night alternation. The proposed sizing methodology provides different possible configurations for the renewable primary sources that are not affected by workload variations. The overall approach leads to a robust sizing. In case the workload has more variability, a negotiation can be initiated at runtime between the IT scheduling and the power supply storage management to adapt the power demand (i.e., by changing the task scheduling of the current workload) and to make the supply of the data center demand possible [41].

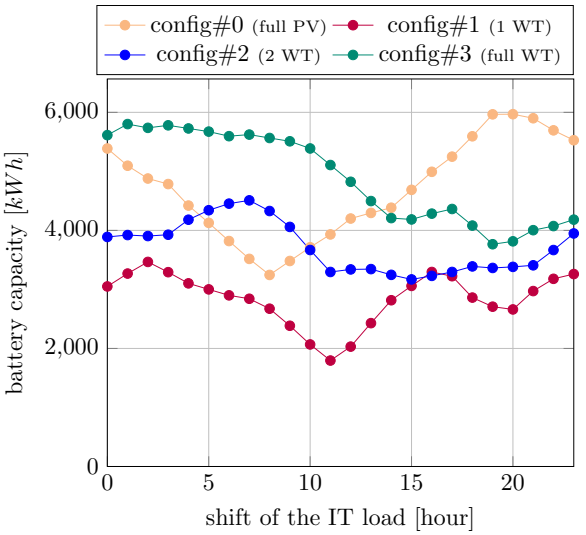

Figure 12: Evolution of the battery capacity as a function of the shift of the IT load.

6.5. Discussion

Now that we have analyzed the influence on different parameters on the sizing, it is legitimate to question the robustness of the approach defended in the paper and to extend this discussion to external parameters such as weather conditions (sun and wind) and submitted workloads day after day.

The first question concerns the influence of meteorological conditions. There is no reason that the meteorological conditions of future years should be the same as the one used for sizing the power supply part of the data center. What is the risk in such an unavoidable case? If the wind blows often enough, especially in winter and if the sun shines without clouds, the primary renewable energy production will be higher than expected by the sizing process. In this case, either the hydrogen level at the end of the year of the considered horizon $\mathcal{H}$ will be larger than expected or, in a worst case, the hydrogen tank capacity will not be large enough to store the whole energy overproduction. Indeed, the overproduction is stored in form of hydrogen as shown in Figure 4. Conversely, if the wind does not blow and the sky is cloudy especially in summer, the part of the renewable energy supplying directly

the data center demand will be reduced compared to the expected one. Then the hydrogen storage will be more frequently solicited to compensate the shortage of power. The stored hydrogen level will decrease and will not be regenerated enough by the overproduction days. So, it appears, that the risk of having meteorological conditions that differ from the one on which the sizing process is based on, is that the tank could be full or empty sooner than expected. A way to cope with this problem is to buy or sell hydrogen. The capacity of the tanks is high (we consider tons of hydrogen), so this decision can be easily scheduled without risking to jeopardize the safety of the data center supply. This decision has not to be seen as a failure in the data center management: the hydrogen market exists and new hydrogen usage is expanding, especially in the vehicle domain, electrical bus or cars using fuel cells and batteries [19]. Moreover, buying certified green hydrogen (i.e., hydrogen certified to be produced from renewable energy and not from fossil fuels) is convenient and keeps the computation within the data center totally green.

Another possible way to deal with meteological conditions is to apply the generic methodology proposed on several years to cover different meteorological scenarios and to prevent such a situation. But as shown in Sections 4 and 5, the primary source combination depends on the whole energy that has to be produced so as to meet the power demand using more or less storage facilities. So the sizing process will result in a worst case solution only based on the year with adverse weather conditions which leads to an oversizing of the infrastructure in most cases as shown by M. Haddad in her PhD thesis [17]. This oversizing versus most of the annual weather conditions leads to an oversized total cost of ownership and is hardly acceptable from the economical point of view. In any case, situations can occur in which the hydrogen tank will never be large enough and the data center becomes a hydrogen producer and uses a small part of the hydrogen produced to supply the computation facilities. But this is another economical model beyond this study.

An alternative is to consider each sizing combination set, one for each reference year, and to propose the average sizing to the decision makers. In this case, M. Haddad also shows in [17] that the hydrogen shortage or hydrogen overproduction is reduced when considering a year with either adverse or favorable weather conditions. In this case, the volume of hydrogen to buy or to sell each year for keeping Constraint 15 true is not huge. In some cases, a small oversizing of the tank capacity could be tolerated in order to increase slightly the system resilience. This is a promising option because using the most representative year in terms of weather conditions is very difficult. Indeed, weather models are very difficult to built, especially for the wind [18], and remain extremely location-dependent.

Finally, the last situation in which the sizing can be less efficient than expected is when the data center computation demand is not what was foreseen, leading to a miss-predicted power demand. If the demand is low, the power supply part will produce more than the data center consumes and hydrogen will be sold or stored, waiting for days where the demand will be higher. But this is not the most probable scenario. Conversely, if the demand is increasing day after day knowing that digital services and applications are more present everywhere each day, and knowing that the power supply infrastructure is built for years, buying hydrogen is not the right answer. In the particular case where computation demands are exceptionally high, they can be regulated using a negotiation process between parts at runtime as mentioned before. In the other cases, one option is to overestimate the demand but the risk is now to oversize the power supply part. On the other hand, since the life cycle of the servers is about five years, much less than the wind turbines (between thirty and fifty years), and since the computation efficiency of servers increases while their power consumption decreases, the increased demand can be contained by the power supply part even if designed years before. In addition, PUE today is smaller than it was ten years ago thanks to other cooling technologies, it should decrease more in the future. Sheikh et al. have proposed a comprehensive survey in [37] of thermal aware scheduling research that aim at improving the cooling efficiencies of a multi-core processing systems in data centers or computing centers. Changing partially a given surface area of photovoltaic panels will also be mandatory because of aging and acts to increase the global efficiency of the platform.

To conclude, it does not seem impossible to use the initial sizing for years using these recommendations, if the initial specifications of the data center are the right ones.

7. Conclusion and Future Work

In this research work, we designed an on-site data center solely powered by local renewable energy (sun and wind) and using short term and long term energy backups. The hybrid renewable energy system consists of photovoltaic panels and wind turbines as primary sources. Batteries and hydrogen system aim at storing energy to overcome the shortcomings of primary sources energy during days and to compensate seasonal variations in the renewable power production. Moreover, the IT power demand is also not constant. The proposed sizing methodology allows following the data center power demand day after day and provides the necessary production by the primary sources to start each year with the same level of hydrogen.

Specifically, we investigated both the IT and power supply infrastructures. This study was divided in two steps: (1) determining optimally the necessary number of servers of the data center for processing a given IT workload and (2) giving a set a power supply infrastructure needed to meet the IT power demand. To the best of our knowledge, this study is the first of its kind, proposing a proven optimal IT sizing together with associated power supply combinations. One originality of the paper is to propose a generic methodology for both IT and electrical sizing. First, the workload and the scheduling that obtain the minimal needed number of servers to process the workload by respecting a given quality of service could be replaced by others depending on the target IT applications or IT models without impacting the methodology. Second, the meteorological conditions of a reference year and the yearly power demand (hour by hour) could be changed, resulting in a different power sizing.

Another originality of this work is that the output of the sizing process is a set of infrastructure sizing combinations, given an IT workload and a data center location with its weather conditions. Indeed the specificity of the WT in comparison with PV is that the WT can be considered as a discrete entity (0, 1, 2 or 3 WT) while PV as continuous value (surface area): For each possible number of WT corresponds a surface area of PV and associated energy storage. Comprehensive experiments are conducted to show the pros and cons of each possible PV-WT combination. A discussion allows the decision maker to select the best data center infrastructure depending on the context.

Finally, we showed that the workload peak within a day has only an influence on the battery capacity, the sizing of other elements being robust. This corroborates the fact that a better synchronization between the power demand and the renewable power production leads to a smaller battery oversizing. This is a very promising perspective to increase cross-dependencies between IT and HRES sizing. Moreover, discarding extreme values (power demand and weather conditions) while using our proposed sizing methodology could probably reduce the sizing with only a small percentage of QoS violations.

Acknowledgments

This work was partly supported by the French Research Agency under the project Datazero (ANR-15-CE25-0012) and the EIPHI Graduate School (ANR-17-EURE-0002).

References

[1] Ishfaq Ahmad and Sanjay Ranka. *Handbook of Energy-Aware and Green Computing – Two Volume set*. CRC Press, 2012.

[2] Monaaf DA Al-Falahi, SDG Jayasinghe, and H Enshaei. A review on recent size optimization methodologies for standalone solar and wind hybrid renewable energy system. *Energy Conversion and Management*, 143:252–274, 2017.

[3] Kamal Anoune, Mohsine Bouya, Abdelali Astito, and Abdellatif Ben Abdellah. Sizing methods and optimization techniques for pv-wind based hybrid renewable energy system: A review. *Renewable and Sustainable Energy Reviews*, 93:652–673, 2018.

[4] S Ashok. Optimised model for community-based hybrid energy system. *Renewable energy*, 32(7):1155–1164, 2007.

[5] S Bahramara, M Parsa Moghaddam, and MR Haghifam. Optimal planning of hybrid renewable energy systems using homer: A review. *Renewable and Sustainable Energy Reviews*, 62:609–620, 2016.

[6] José L Bernal-Agustín and Rodolfo Dufo-López. Simulation and optimization of stand-alone hybrid renewable energy systems. *Renewable and Sustainable Energy Reviews*, 13(8):2111–2118, 2009.

[7] José L Bernal-Agustín, Rodolfo Dufo-López, and David M Rivas-Ascaso. Design of isolated hybrid systems minimizing costs and pollutant emissions. *Renewable Energy*, 31(14):2227–2244, 2006.

[8] S. Caux, P. Renaud-Goud, G. Rostirolla, and P. Stolf. IT optimization for datacenters under renewable power constraint. In *Euro-Par 2018: Parallel Processing - 24th International Conference on Parallel and Distributed Computing, Turin, Italy, August 27-31, 2018, Proceedings*, pages 339–351, 2018.

[9] Georges Da Costa, Jean-Marc Pierson, and Leandro Fontoura Cupertino. Effectiveness of neural networks

for power modeling for Cloud and HPC: It's worth it! *ACM Transactions on Modeling and Performance Evaluation of Computer Systems*, page (on line), 2020.

[10] Said Diaf, Gilles Notton, M Belhamel, M Haddadi, and Alain Louche. Design and techno-economical optimization for hybrid pv/wind system under various meteorological conditions. *Applied Energy*, 85(10):968–987, 2008.

[11] Rodolfo Dufo-Lopez and José L Bernal-Agustín. Design and control strategies of PV-diesel systems using genetic algorithms. *Solar energy*, 79(1):33–46, 2005.

[12] Ozan Erdinc and Mehmet Uzunoglu. Optimum design of hybrid renewable energy systems: Overview of different approaches. *Renewable and Sustainable Energy Reviews*, 16(3):1412–1425, 2012.

[13] Lee Fingersh, Dave Simms, Maureen Hand, Dave Jager, Jason Cotrell, Mike Robinson, Scott Schreck, and Scott M Larwood. Wind tunnel testing of NREL's unsteady aerodynamics experiment. In *20th ASME Wind Energy Symposium*, 2001.

[14] Greendatanet research project. http://www.greendatanet-project.eu/.

[15] Abdou Guermouche, Loris Marchal, Bertrand Simon, and Frédéric Vivien. Scheduling trees of malleable tasks for sparse linear algebra. In Jesper Larsson Träff, Sascha Hunold, and Francesco Versaci, editors, *Euro-Par 2015: Parallel Processing*, pages 479–490, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[16] Ali Habibi Khalaj, Khalid Abdulla, and Saman K. Halgamuge. Towards the stand-alone operation of data centers with free cooling and optimally sized hybrid renewable power generation and energy storage. *Renewable and Sustainable Energy Reviews*, 93:451–472, October 2018.

[17] Maroua Haddad. *Sizing and Management of Hybrid Renewable Energy System for Data Center Supply*. PhD thesis, Unioversié Bourgogne Franche-Comté, 2019.

[18] Maroua Haddad, Jean-Marc Nicod, Yacouba Boubacar Mainassara, Landy Rabehasaina, Zeina Al Masry, and Marie-Cécile Péra. Wind and solar forecasting for renewable energy system using sarima-based model. In *6th International conference on Time Series and Forecasting (2019)*, Gran Canaria, Spain, sep 2019.

[19] Marwa Haddad, Jean-Marc Nicod, and Marie-Cécile Péra. Hydrogen infrastructure: data-center supply-refueling station synergy. In *IEEE Vehicle Power and Propulsion Conference (VPPC'2017)*, pages 1–6, 2017.

[20] JK Kaldellis, D Zafirakis, and E Kondili. Optimum autonomous stand-alone photovoltaic system design on the basis of energy pay-back analysis. *Energy*, 34(9):1187–1198, 2009.

[21] A. Kassab, J. M. Nicod, L. Philippe, and V. Rehn-Sonigo. Scheduling independent tasks in parallel under power constraints. In *2017 46th International Conference on Parallel Processing (ICPP)*, pages 543–552, Aug 2017.

[22] YA Katsigiannis, PS Georgilakis, and ES Karapidakis. Multiobjective genetic algorithm solution to the optimum economic and environmental performance problem of small autonomous hybrid power systems with renewables. *IET Renewable Power Generation*, 4(5):404–419, 2010.

[23] Bithika Khargharia, Salim Hariri, Ferenc Szidarovszky, Manal Houri, Hesham El-Rewini, Samee Ullah Khan, Ishfaq Ahmad, and Mazin S Yousif. Autonomic power & performance management for large-scale data centers. In *2007 IEEE International Parallel and Distributed Processing Symposium*, pages 1–8. IEEE, 2007.

[24] Sang C Lee and Woo Young Jung. Analogical understanding of the ragone plot and a new categorization of energy devices. *Energy procedia*, 88(526-530), 2016.

[25] M. Lin, Z. Liu, A. Wierman, and L. L. H. Andrew. Online algorithms for geographical load balancing. In *2012 International Green Computing Conference (IGCC)*, pages 1–10, June 2012.

[26] Z. Liu, M. Lin, A. Wierman, Steven H. Low, and L. L. H. Andrew. Geographical load balancing with renewables. *SIGMETRICS Perform. Eval. Rev.*, 39-3:62–66, December 2011.

[27] Eric Masanet, Arman Shehabi, Nuoa Lei1, Sarah Smith, and Jonathan Koomey. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, February 2020.

[28] NREL. Hybrid optimization model for electric renewable energy, 2009.

[29] Luca Parolini, Bruno Sinopoli, and Bruce H Krogh. Reducing data center energy consumption via coordinated cooling and load management. In *Proceedings of the 2008 conference on Power aware computing and systems, HotPower*, volume 8, pages 14–14, 2008.

[30] Jean-Marc Pierson, Gwilherm Baudic, Stéphane Caux, Berk Celik, Georges Da Costa, Léo Grange, Marwa Haddad, Jerome Lecuivre, Jean-Marc Nicod, Laurent Philippe, Veronika Rehn-Sonigo, Robin Roche, Gustavo Rostirolla, Amal Sayah, Patricia Stolf, Minh-Thuyen Thi, and Christophe Varnier. DATAZERO: DATAcenter with Zero Emission and RObust management using renewable energy. *IEEE Access*, 7:(on line), juillet 2019.

[31] Singiresu S Rao. *Engineering optimization: theory and practice*. John Wiley & Sons, 2009.

[32] Shafiqur Rehman and Ibrahim M El-Amin. Study of a standalone wind and solar pv power systems. In *2010 IEEE International Energy Conference*, pages 228–232. IEEE, 2010.

[33] André Rouyer. Energy policy research and implications for data centres in emea. Technical Report WP#44, The Green Grid, jan 2012.

[34] Yashwant Sawle, SC Gupta, and Aashish Kumar Bohre. Review of hybrid renewable energy systems with comparative analysis of off-grid hybrid system. *Renewable and Sustainable Energy Reviews*, 81:2217–2235, 2018.

[35] Manajit Sengupta, Yu Xie, Anthony Lopez, Aron Habte, Galen Maclaurin, and James Shelby. The national solar radiation data base (NSRDB). *Renewable and Sustainable Energy Reviews*, 89:51–60, 2018.

[36] N. Sharma, S. Barker, D. Irwin, and P. Shenoy. Blink: Managing server clusters on intermittent power. *SIGARCH Comput. Archit. News*, 39(1):185–198, March 2011.

[37] Hafiz Fahad Sheikh, Ishfaq Ahmad, Zhe Wang, and Sanjay Ranka. An overview and classification of thermal-aware scheduling techniques for multi-core processing systems. *Sustainable Computing: Informatics and Systems*, 2(3):151–169, 2012.

[38] Jun-Hai Shi, Xin-Jian Zhu, and Guang-Yi Cao. Design and techno-economical optimization for stand-alone hybrid power systems with multi-objective evolutionary algorithms. *International Journal of Energy Research*,

31(3):315–328, 2007.

[39] Rajanna Siddaiah and RP Saini. A review on planning, configurations, modeling and optimization techniques of hybrid renewable energy systems for off grid applications. *Renewable and Sustainable Energy Reviews*, 58:376–396, 2016.

[40] Sunanda Sinha and SS Chandel. Review of software tools for hybrid renewable energy systems. *Renewable and Sustainable Energy Reviews*, 32:192–205, 2014.

[41] Minh-Thuyen Thi, Jean-Marc Pierson, Georges Da Costa, Patricia Stolf, Jean-Marc Nicod, Gustavo Rostirolla, and Marwa Haddad. Negotiation game for joint it and energy management in green datacenters. *Future Generation Computer Systems*, 2019.

[42] G Tina and S Gagliano. Probabilistic analysis of weather data for a hybrid solar/wind energy system. *International Journal of Energy Research*, 35(3):221–232, 2011.

[43] Violaine Villebonnet, Georges Da Costa, Laurent Lefevre, Jean-Marc Pierson, and Patricia Stolf. Big, medium, little: Reaching energy proportionality with heterogeneous computing scheduler. *Parallel Processing Letters*, 25(03):1541006, 2015.

[44] Violaine Villebonnet, Georges Da Costa, Laurent Lefèvre, Jean-Marc Pierson, and Patricia Stolf. Energy aware dynamic provisioning for heterogeneous data centers. In *2016 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 206–213. IEEE, 2016.

[45] Lingfeng Wang and Chanan Singh. Multicriteria design of hybrid power generation systems based on a modified particle swarm optimization algorithm. *IEEE Transactions on Energy Conversion*, 24(1):163–172, 2009.

[46] Hongxing Yang, Zhou Wei, and Lou Chengzhi. Optimal design and techno-economic analysis of a hybrid solar–wind power generation system. *Applied Energy*, 86(2):163–169, 2009.

[47] Hongxing Yang, Wei Zhou, Lin Lu, and Zhaohong Fang. Optimal sizing method for stand-alone hybrid solar–wind system with lpsp technology by using genetic algorithm. *Solar energy*, 82(4):354–367, 2008.

[48] HX Yang, L Lu, and J Burnett. Weather data and probability analysis of hybrid photovoltaic–wind power generation systems in hong kong. *Renewable Energy*, 28(11):1813–1824, 2003.