Demand esponse program for supercomputing and datacenters providing ancillary services in the electricity market

S.Montes de Oca, J. Muraña, P. Monzón, S. Nesmachnow, S. Iturriaga, G. Belcredi
IIE-INCO-Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
Email: smontes,jmurana,monzon,sergion,siturria,gbelcredi@fing.edu.uy

***Abstract*—In this article, we studied a negotiation approach for the participation of datacenters and super-computing facilities in smart electricity markets providing ancillary services, an important problem in modern smart grid systems. Different demand response strategy were studied for colocation datacenters to commit power reductions during a sustained period, according to offers proposed to tenants. The negotiation algorithm and a heuristic planning method for energy reduction optimization were experimentally validated over realistic problem instances that model different problem dimensions and flexibility of the datacenter clients. The obtained results indicate that the proposed approach is effective to provide appropriate frequency reserves control according to monetary incentives.**

***Index Terms*—Demand response, Actives agents on the electricity market, Ancillary services**

I. Introduction

The penetration of distributed generation based on renewable sources is increasing in most electricity markets. This replacement of traditional sources based on nuclear and fossil fuels for renewable, mainly wind and solar, is still expected to increase in the short time. Nevertheless, to effectively integrate these new technologies in the supplied side, new issues as operation, control and stability of the power network should be addressed. Besides this technological changes, electricity markets are undergoing an institutional transition. To enhance economic efficiency and improve services to the consumer, the electricity markets have being liberalized gradually, leading to the introduction of competition and opening in the wholesale markets first, and more slowly in the retail market. However, networks are regulated to protect consumers and to guarantee access to the grid for different actors.

Wholesale electricity markets were designed to meet short-term and future requirements of operating the electric power system reliably and at the lowest cost. Policy makers saw competition among suppliers as a mean to control pricing by attracting new sources and technologies from the private sector in an open, competitive, and transparent market. The wholesale market is structured in several sub-markets, with different horizon times in advance of the electricity purchase. Markets range from a few seconds or minutes for ancillary services, correcting on-line mismatches between generation and demand (electricity frequency stability), to a few year in advance for the capacity market (assuring the capacity of the future demand), including day-ahead and intra-day planning markets for the energy purchase [1].

From the demand-side point of view, the evolution of the energy sector under the paradigm of the *smart grid* is enabling the interaction between end-users and grid operators. Smart electricity network refers to a electrical grid that includes operation and management features to improve the controlling of production and distribution of energy [2]. Smart grids are the current state-of-the-art technology for electricity networks, the last step in their evolution from unidirectional systems of electric power transmission and distribution to holistic approaches that provide different services for demand-driven control. The main goal of the smart grid is to maintain a reliable, resilient and secure infrastructure to properly satisfy the demand growth and the integration of distributed energy resources such us: small generators based on renewable sources (typically wind and solar); big consumer with flexible load; electric vehicle fleet; or small and distributed loads through the coordination of an aggregator, relating smart devices and real-time information provided to clients [3]. Information and Communication Technologies have provided a key foundation for communicating and processing information that is very useful at different levels to implement the aforementioned services [4].

Within the smart grid paradigm, a large consumer with flexible power profile can participate in the electricity market. This is one of the main ideas behind the implementation of strategies oriented to modern smart electric networks, where consumers are associated to the roles of both active clients and market agents [2]. There are several examples of active consumers with the capacity to plan in advance its power profile for the next or following days: smart building centrally controlled, EV fleets, factories, datacenters, super-computing facilities, etc. Paradigms and strategies applying multi-hour

tariffs can also be implemented, handling time periods where it is preferable that consumers use energy. On the other hand, many consumers may not be capable to adapt its power profile in advance for the next day but can adapt quickly to sudden changes in the electricity network through control signals sent by the grid operator. These kind of customers are capable to adapt in the short time, helping the grid operator to rapidly recover from a potentially disruptive event, such us a frequency deviation or black start services. Is in this sense that a market agent or an active consumer can participate in the electricity market in several ways, receiving an income for providing different services.

This article shows the effects of applying demand response strategies on data centers and super-computing infrastructures as an example of flexible large consumers, allowing them to participate in the electricity market providing ancillary services for frequency restoration through a demand response program. As a relevant case study, we address the possibilities of planning strategies for colocation data centers and super-computing infrastructures. The specific internal aspects of the datacenter decisions about the workloads dispatch and the associated power consumption were deeply studied in [5]. These platforms are conceived as examples of planned systems that have emerged in modern societies, linked to the smart grid paradigm. We focus in *colocation datacenters* which represent almost 40% of datacenters in USA [6]. In colocation datacenters, multiple tenants deploy and keep full control of their own physical servers in a shared space, while the datacenter operator provides facility support (e.g. high-availability power, bandwidth and cooling). Data center and super-computing facilities can adjust power consumption to help the electric network to fulfill specific goals, either by consuming available surplus of energy to execute complex tasks, or by deferring or discarding activities (i.e., tasks execution) in case of an expected event happens (lost of a generation unit, line congestion in transportation o distribution grid, etc).

This research studies the participation of a data center in an emergency demand response, i.e., a type of program in which participants sign contracts and are obliged to reduce their load for a short period of time, when requested. The article is structured as follows. Section II develops different demand response schemes. In Section III, a mathematical model is introduced, together with a demand respond strategy. Finally, Sections IV and V present simulations results, along with the respective analysis and some conclusions.

II. Demand response schemes

Power grids must always match power supply and demand. A stable AC frequency can be thought of as an equivalent physical condition for this constant supply-demand balance. Secondary frequency control occurs after primary frequency control (i.e. instantaneous and autonomous operating through automatic generation control) has already stabilized the initial frequency fluctuation to a certain level and its goal is to subsequently manage frequency back to the target level (see Figure 1) [7].

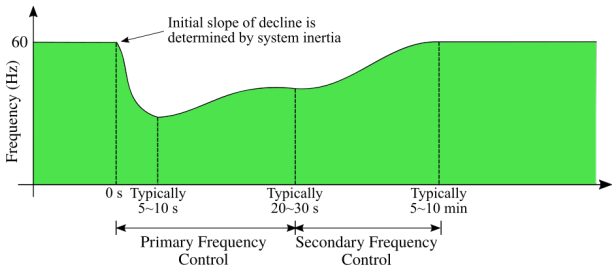

Fig. 1: Time vs. frequency after a grid disturbance event causing a drop in grid frequency.

In case of an emergency event is triggered by the grid operator, the datacenter must reduce its load in response to the demand response event either through extracting IT energy reductions (re-scheduling workflow) or by turning on an on-site generator. Since the mandatory demand response reduction target is fixed during the event, the datacenter operator must balance between IT energy reductions and using on-site generation in order to minimize the datacenter operational cost.

A. Wholesale Market - Ancillary Services

In open deregulated electricity markets, demand response is considered a supply-side resource in capacity and ancillary services. In particular, in Pennsylvania-New Jersey-Maryland Interconnection (PJM) electricity market, demand response services are concentrated in the synchronized reserves for frequency restoration and the capacity markets through the emergency and economic load program [7]. In other electricity markets (e.g., UK or Germany) there are similar demand response programs for supply-side resource, mainly focusing on frequency restoration or for demand turn up services [7], [8]. Secondary frequency reserves, such us frequency containment reserve in the European Electricity Balancing Market or synchronized reserves in PJM markets, are operating reserves necessary for constant containment of frequency deviations (fluctuations) from nominal value in order to constantly maintain the power balance in the whole synchronously interconnected system.

These demand response programs are effective for customers which can manage their power profile in the short time and can adapt its consumption level quickly if needed. In the other hand, this programs allows the wholesale market to incorporate load-reduction actions. From the perspective of the grid operator, the flexible power demand of datacenters serves as a valuable energy buffer, helping balance the grid power supply and demand at runtime. In this article we focuses on the PJM electricity market, particularly on the synchronized reserve market and the emergency load program. In both cases, if the grid operator anticipates an emergency (e.g., wrong forecast of demand, extreme weather, or a generation unit out of work), participants are notified, usually at least 10 minutes

in advance, and obliged to fulfill their contracted amounts of energy reduction during the event, which may span a few minutes to a few hours [9], [10].

The main differences between the two DR services are the time horizon of the offers and the mechanism to participate. For the emergency and economic load program, participants typically sign contracts with a load serving entity in advance (e.g., three years ahead in PJM) and receive financial rebates for their committed energy reduction.In the synchronized reserve market, the demand resource participates in a real time market through auctions for each time gap of the market, defining in the offer the price and the capacity to be reduced if necessary.

B. Demand Response program

Demand response (DR) is a reduction in the consumption of electric energy by customers from their expected consumption in response to an increase in the price of electric energy or to incentive payments designed to induce lower consumption of electric energy [7]. DR programs come in a variety of types. Some of them are created and run by utilities or the grid operator who work directly with customers to form and execute curtailment plans. Other utility-based DR programs are delivered in the form of dynamic pricing tariffs that encourage customers to reduce their loads during peak times. So, DR programs have different incentive schemes and program objectives. Two of the primary types of incentives are capacity payments (basically for reduction of load) and energy payments. A capacity program provides payments to customers to stand by to be ready to help the grid, either to reduce peak demand or to stabilize the grid during an emergency and prevent blackouts. Energy payments are provided based on the actual energy supplied (e.g., not consumed) by a customer over a set period of time during a demand response event. For customers participating in dynamic pricing programs, the incentives are typically represented by rate discounts during the off-peak periods that more than offset the significantly higher rates during the critical peak periods.

In this work, we focus on the first type of program, where the grid operator ask for a reduction in the actual load to the customer through a DR call event. For this reason the measurement and verification of DR is a critical component of any program. The baseline is the primary tool for measuring curtailment during a DR event. A baseline is an estimate of the electricity that would have been consumed by a customer in the absence of a DR event [11].

A DR event has three phases of curtailment, as is shown in Figure 2 [11].

- Phase 1 – The ramp period, which begins with deployment, is when sites begin to curtail.
- Phase 2 – The sustained response period, which is the time period bounded by the reduction deadline and the release/recall, is the time in which the DR resources are expected to have arrived and to stay at their committed level of curtailment.
- Phase 3 – The recovery period, which occurs after customers have been notified that the event has ended, is the period when customers begin to resume normal operations.

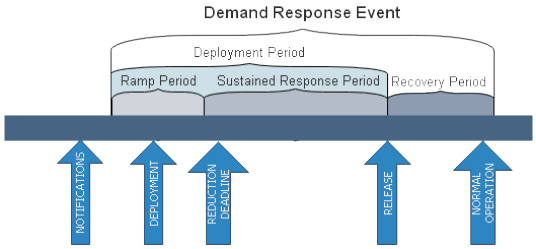

Fig. 2: Timing of a DR event.

Generally, an ancillary services event is intended to reduce load on the grid at that moment, for a short period of time, rather than to reduce a dynamic load profile likely to fluctuate over time. As a result, the most common measurement methodology is Meter Before/Meter After, which can be defined as: "A performance evaluation methodology where electricity consumption or demand over a prescribed period of time prior to Deployment is compared to similar readings during the Sustained Response Period" [11]. In ancillary services, the duration of the event is usually ten minutes to two hours. For this reason ancillary programs typically use Meter Before-Meter After baselines.

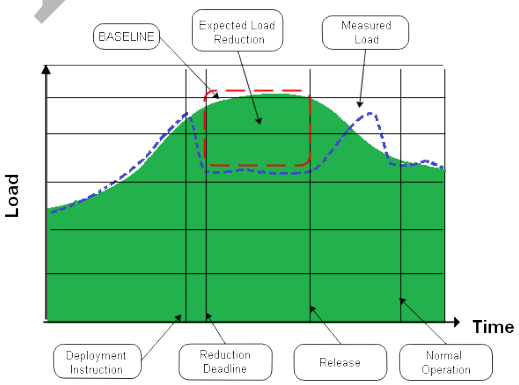

Fig. 3: Meter before/meter after baseline for ancillary service demand DR program.

Figure 3 illustrates the basic idea behind baselines [7]. On a day without a dispatch signal, the customer would have used electricity over time along the green line labeled "baseline." Given the dispatch signal, the customer will reduce their load during few minutes to hours, by the amount of the gap between the baseline and the dotted line labeled "Measured Load." As indicated above, knowledge of load over time requires time-interval meters and a method for recording their output for reporting to the grid operator or other authority.

III. System model and proposed algorithm

In this Section we present the DR scheme for the colocation datacenter. First, we briefly describe how the internal negotiation between the datacenter operator and the tenants works. After that, we present the main idea and the respective algorithm that is run by the datacenter operator.

A. Energy consumption model for datacenters and supercomputing

The datacenter follows a *colocation* model with a set of tenants, each with a subset of the total computing resources of the datacenter. Computing resources of any single tenant are considered to be homogeneous in the proposed model, however resources are heterogeneous when considering multiple tenants. The energy consumed by the datacenter is determined by a set of workload schedules, one for each tenant. The algorithm for computing each schedule is specific for each tenant and solves an underlying multi-objective optimization to minimize energy budget, violation of due dates and execution time, among other goals. Figure 4 presents a scheme of the proposed model. Details on how the optimization of each tenant is carried out, including a detailed power consumption model for a given workload, can be found in [5], [12] and references therein. Recall that the datacenter operator is responsible for non-IT facility support (e.g., high-availability power, cooling). We capture the non-IT energy consumption using Power Usage Effectiveness (PUE) γ, which is the ratio of the total data center energy consumption to the IT energy consumption. Typically, γ ranges from 1.1 to 2.0, depending on factors such as outside temperature.

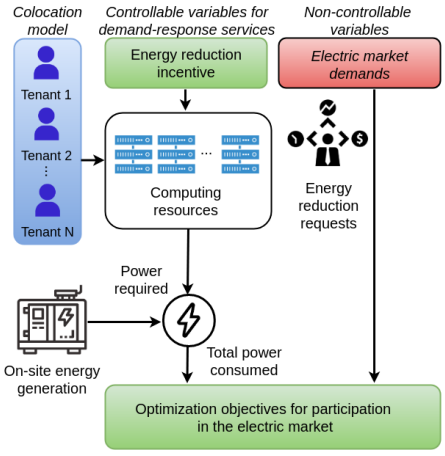

Fig. 4: Schema of the proposed model for energy consumption in datacenters and super-computing facilities [5].

When the electric market operator requests the datacenter to reduce its energy consumption by a target $D * \gamma$, where D is the IT-energy consumption, the datacenter initiate a negotiation phase with its tenants. During this negotiation, the datacenter offers a monetary incentive per each unit of energy reduced. Considering this incentive, each tenant may choose to modify its planned scheduling and compute a new one by postponing or even cancelling the execution of some of its tasks to reduce its energy consumption. In addition, an on-site energy (e.g., fossil fuel) generator is considered, with a monetary cost per unit of energy generated. This generator can be used to reduce the energy the datacenter consumes from the grid, in case the energy reduction from its tenants is not enough to meet the reduction target D from the market operator. The on-site generator is controlled by the datacenter owner.

B. Client offer evaluation.

To evaluate the monetary offer of the datacenter administrator and determine the amount of power to be reduced, tenants simulate the execution of their workload, applying an energy optimization strategy (Sec. IV in [5]). The offer evaluation function computes its power consumption reduction according to the monetary offer received. The monetary offer of the datacenter administrator is accepted if the net income obtained from the energy reduction minus the loss the tenant must pay in case not complying with the Service Level Agreements (SLA) with his users, is greater than zero. In any case, different trade-offs are obtained for different monetary offers from the negotiation. These trade-offs can be considered if the datacenter cannot meet the desired power consumption reduction, to account for different compromises between the problem objectives (energy reduction and cost).

C. Proposed tenant-negotiation scheme

The proposed market mechanism for colocation datacenters follows the main idea from Chen et al. [13], characterizing the Nash optimum of the resulting non-cooperative game as an optimization problem known as allocation problem.In this approach, the operator can induce a reduction on tenant's power consumption, diminishing the need of brown energy (on-site diesel generator), using a parameterized supply function represented in Eq. 1, where: r_i is the power reduction for tenant i, D is data center's power reduction target, b_i is the tenant offer for reducing the power consumption by r_i and p is the market clearing price determined by the operator.

$$r_i(b_i, p) = D - \frac{b_i}{p} \tag{1}$$

The market mechanism for reducing D amount of energy is exercised in four steps in an iterative approach:

(i) The datacenter broadcasts the supply function to the clients, $r_i(b_i, p) = D - \frac{b_i}{p}$.

(ii) Each tenant i bids a reward b_i for reducing r_i amount of power, in order to maximize its utility and can be interpreted as the IT revenue that tenant is willing to give up.

(iii) The datacenter determines the market clearing price p and the amount of energy to produce via on-site generation y (with generation cost α) by minimizing the total cost, represented by the cost of generation and the rewards payed to the tenants.

$$p(b_i, y) = \frac{\sum_i b_i}{(N-1)D + y} \quad (2)$$

$$y = \arg\min_{0 \le y \le D} (D - y)p + \alpha y \quad (3)$$

The first-order optimality condition for Eq. 3 gives the value for y:

$$y = \sqrt{\frac{(\sum_{i=1}^{N} b_i) N D}{\alpha}} - (N-1)D \quad (4)$$

(iv) If p and y converges, latest bids are accepted and energy reduction is scheduled by the clients, else the operator broadcast the new supply function with the updated value for p.

Algorithm 1 Datacenter market mechanism

INPUT: D (power reduction target), $price_0$
OUTPUT: $price$, *on-site-generation*

$k \leftarrow 0$ ▷ iteration step
$price_k \leftarrow price_0$
while $\epsilon \geq \epsilon_{min}$ **do**
    end
    j=1 to N
$reduction[j] \leftarrow client\_evaluation(price, j)$
$bid[j] \leftarrow (D - reduction[j]) \times price_k$
$y_k \leftarrow \max(\sqrt{(\sum bid) N.D/\alpha} - (N-1)D, 0)$
$price_k \leftarrow \sum_j bid/((N-1)D + y_k)$
$\epsilon \leftarrow \|(y_k + \sum_j reduction - D)/D\|$
$k \leftarrow k + 1$

on-site_generation $\leftarrow y_k$

Algorithm 1 describes the strategy used by the datacenter, by implementing a solution of the allocation problem based on a proximal method [5]. A distributed solution is generated for each agent. These solutions are coupled by the power balance equation $D = \sum_{i=1}^{N} reduction[i] + y_k$. This equality constraint is relaxed in the proximal method. In the algorithm, D is the power reduction target, $price$ is the market clearing price per Watt, N is the number of tenants and j is the tenant id. The function $client\_evaluation(price, j)$ corresponds to the offer evaluation of the tenant j, considering its SLAs. This function returns the energy reduction committed by the tenant ($reduction[j]$), according to the price, *bid*[*j*] is the offer of tenant j for reducing the power consumption, y_k is the iteration variable, which at the end of the negotiation corresponds to the power generated by the on-site generator. The cost of generate one Watt using the generator is denoted α. The parameter ϵ is a measure of the compliance of the coupling restriction.

IV. Simulations and results analysis

Strategies for smart planning of tasks execution and management of energy utilization are proposed for the National Supercomputing Center in Uruguay (Cluster-UY) [14], taking into account the energy consumption and the Quality of Service (QoS) provided to users. Another feature of the proposed scheme is that, unlike other works in the related literature, real tasks data and real energy consumption evaluation are considered for the planning instead of using theoretical models. The experimental evaluation is performed through simulations that consider realistic workloads, high-end servers, and a power consumption model built from real data. Results suggest the effectiveness of the proposed strategies to implement demand response techniques for datacenters and provide ancillary services under the smart grid paradigm.

Tenants are assumed to be focused on executing scientific applications, which are the ones that demand significant energy consumption [15]. Applications are modeled as *computing tasks*. Two types of tasks are considered: CPU-intensive and memory-intensive, which accounts for the most common types of scientific applications, according to the related literature [12].

A. Numerical results

Problem instances for the tenants were created considering real data from both workloads executed and computing resources available on nowadays datacenters and supercomputing facilities. Further details can be found in [5]. Regarding simulation parameters, the following values were set: the simulation time horizon for task planning is T=60 minutes. The time duration of the DR call event is 10 minutes and the demanded IT-power reduction D is $3kW$, and $100kW$ for small and large instances, respectively; a price per Watt for the on-site power generator of 2 monetary units; an error threshold of 0,005; and an initial offer price of 0,01 monetary units.

The files describing the instances and the source code of the instances generator implemented are publicly available at https: //www.fing.edu.uy/inco/grupos/cecal/hpc/DRAS/.

In Figure 5 we present simulation results of the offer for three sample scheduling strategies for small instances. Results show that the power optimization strategy manages to reduce the power consumption of the tenants during the DR event duration. In blue line we show the IT-power baseline load executed by the datacenter without a DR call. In red we show an strategy where the data center incentives tenants to reduce consumption during the sustained response period but put no constraints on the ramp up limit after the DR event ends. As a results, tenants scheduling planers execute as many task as possible using all the resource available during the recovery periods. This effect could have a negative impact for the grid operator. Nevertheless, we show in yellow and grid lines two different ramp up control strategies, constraining the power resources available (set of servers available for the tenants) in the recovery periods. These strategies delay in time the execution of tasks in the workflow obtaining a smooth power consumption profile.

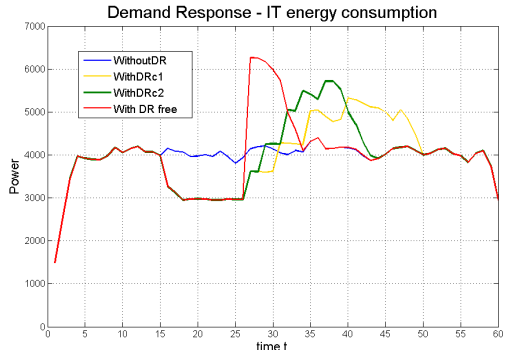

Fig. 5: DR re-scheduling for small instances.

In Figure 6 we can see the different phases of curtailment response of the datacenter to a DR event and the extracted load reduction from the tenants during the event. We can see that the performance of the datacenter accomplish the three phases of curtailment established for a DR event.

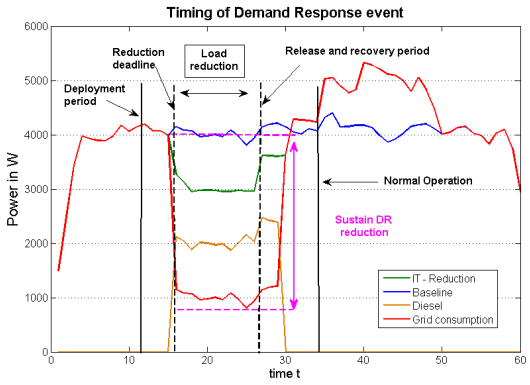

Fig. 6: Timing of a DR event.

In Figure 7 we can see also different strategies of ramp up control executed for large workflows of several tenants, achieving positives results.

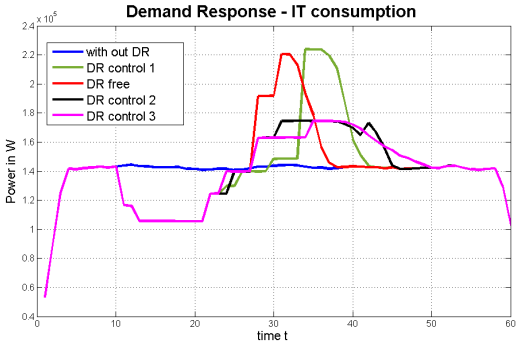

Fig. 7: DR re-scheduling for large instances.

V. Conclusion and Future Work

This article studied a negotiation approach for the participation of datacenters and super-computing facilities in smart electricity markets providing ancillary services, an important problem in modern smart grid systems. A specific case of demand response strategy was studied for colocation datacenters to commit power reductions during a sustained period, according to offers proposed to tenants. The negotiation algorithm and a heuristic planning method for energy reduction optimization were experimentally validated over realistic problem instances that model different problem dimensions and flexibility of the datacenter clients. The obtained results indicate that the proposed approach is effective to provide appropriate frequency reserves control according to monetary incentives.

Acknowledgments

The work was partially supported by Agencia Nacional de Investigación e Innovación (FSE-2017-1-144789). The work of S. Nesmachnow and S. Iturriaga has been partly funded by ANII and PEDECIBA, Uruguay. The authors also want to thank the Centro Nacional de Supercomputación (Cluster.Uy).

References

[1] Steven Stoft. *Power System Economics: Designing Markets for Electricity*. Wiley-IEEE Press, 1 edition, 2002.

[2] James Momoh. *Smart Grid: Fundamentals of Design and Analysis*. Wiley-IEEE Press, 2012.

[3] Federal Energy Regulatory Commission. Assessment of demand response & advanced metering. Technical Report AD-06-2-00, 2006.

[4] European Automotive Research Partner Association. Smart Grids European Technology Platform. January 2020.

[5] J. Muraña, S. Nesmachnow, S. Iturriaga, S. Montes de Oca, G. Belcredi, P. Monzón, V. Shepelev, A. Tchernykh. Negotiation approach for the participation of datacenters and supercomputing facilities in smart electricity markets. *Programming and Computer Software*, 2020.

[6] Pierre Delforge and Josh Whitney. Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers. Technical report, Natural Resources Defense Council and Anthesis, 2014.

[7] . D. Hurley, P. Peterson, M. Whited *Demand Response as a Power System Resource* Synapse, Energy Economics,Inc. May2013

[8] National Grid PLC. Product roadmap for frequency response and reserve, December 2017.

[9] PJM Day-Ahead and Real-Time Market Operations Division. *Manual 11: Energy & Ancillary Services Market Operations*. PJM Interconnection LLC, 108 edition, 3 2018.

[10] PJM Dispatch Operations Division. *Manual 12: Balancing Operations*. PJM Interconnection LLC, 39 edition, 2 2019.

[11] Rossetto, Nicolò *An overview of the methodologies for calculating customer baseline load in PJM*, European University Institute, 2018,

[12] Jonathan Muraña, Sergio Nesmachnow, Fermín Armenta, and Andrei Tchernykh. Characterization, modeling and scheduling of power consumption of scientific computing applications in multicores. *Cluster Computing*, 22(3):839–859, Sep 2019.

[13] Niangjun Chen and Xiaoqi Ren and Shaolei Ren and Adam Wierman. Greening multi-tenant data center demand response. *Performance Evaluation*, 91:229–254, 2015.

[14] Sergio Nesmachnow and Santiago Iturriaga. Cluster-UY: scientific HPC in Uruguay. In *International Supercomputing in México*, 2019.

[15] Sergio Nesmachnow, Cristian Perfumo, and Íñigo Goiri. Holistic multiobjective planning of datacenters powered by renewable energy. *Cluster Computing*, 18(4):1379–1397, 2015.