Multiobjective Workflow Scheduling in a Federation of Heterogeneous Green-Powered Data Centers

Santiago Iturriaga, Sergio Nesmachnow
Universidad de la República, Uruguay
Email: {siturria, sergion}@fing.edu.uy

Andrei Tchernykh
CICESE Research Center, Mexico
Email: chernykh@cicese.mx

Bernabé Dorronsoro
Universidad de Cádiz, Spain
Email: bernabe.dorronsoro@uca.es

***Abstract*—The energy consumption of large data centers has been increasing for the last decades and currently is a major concern for economic and environmental reasons. Accurate scheduling of the data center operation and use of renewable energy sources present themselves as promising solutions for this problem.**

In this paper we study the problem of scheduling workflows of tasks in distributed heterogeneous data centers which are partially powered by renewable energy sources. This problem takes into account quality of service, infrastructure usage, and power consumption of machines and cooling devices. We propose a mathematical model for accurate scheduling solutions.

***Keywords*-data centers; green energy; scheduling**

I. Introduction

Data centers are facilities hosting hundreds to thousands of computing resources providing compute, network and storage services. These massive computing infrastructures are used to solve large-scale problems in different application domains such as science, industry and commerce. A federation of data centers is a set of geographically distributed data centers which cooperate with each other to solve problems not addressable by a single data center [1]. This is the architecture of modern supercomputing systems such as clouds and grids.

Energy efficiency is critical for the operation of data centers. Their energy consumption rate has been consistently on the raise since 2005. In 2010, it yielded about 1.3% of the total worldwide energy consumption. Such an energy consumption causes economic, environmental, and technical concerns for providers [2]. The use of green energy is a clear alternative to traditional brown energy, reducing the operational budget and the enviromental impact of the data center operation. Nevertheless, further reduction of the energy consumption is important. Green energy is not a solution by itself because of its unreliability and because it does not solve issues such as the increasing heat dissipation.

There are a number of techniques for reducing the energy consumption in a data center, ranging from embeded hardware solutions to more general software-controlled methods [1]. However, reducing energy consumption usually reduces the computing performance and may affect negatively the Quality of Service (QoS). This is a complex scenario which requieres a multiobjective analysis for finding accurate solutions with different trade-offs between energy consumption and QoS [3].

In this work, we tackle the scheduling of High Performance Computing (HPC) workflows with deadlines in a federation of distributed green-powered datacenters for minimizing the energy consumption, maximizing the green energy consumption, and maximizing the QoS. We specifically tackle scenarios with HPC workflows which require high computing usage and low networking and storage usage. Thus, we approximate the energy consumed by each data center with the energy consumed by their Central Processing Units (CPUs). We measure the QoS by considering the number of workflows' deadlines met.

For tackling this problem, we divide it in two smaller scheduling subproblems. A higher-level scheduling problem for allocating workflows to data centers, and a lower-level scheduling problem for scheduling the tasks of each workflow to the computing resources inside each data center. We propose two MultiObjective Evolutionary Algorithms (MOEAs), one for solving each scheduling subproblem. These algorithms take into account not only the execution of the workflows, but also powering servers on/off and controlling cooling devices.

II. Higher-level scheduling problem in federated data centers

The higher-level scheduler allocates workflows to geographically distributed data centers in a federation of data centers, and determines the order in which each datacenter executes assigned workflows. We define the accuracy of the schedule by considering the time required to complete the execution of all workflows (makespan), the energy consumed during their execution, and the number of deadlines violated. Figure 1 shows an overview of the problem tackled by the higher-level scheduler.

The model of the meta scheduling problem is composed of the following elements.

- A the federation of data centers comprised by k heterogeneous datacenters $DC = \{dc_1, \ldots, dc_k\}$. Each data

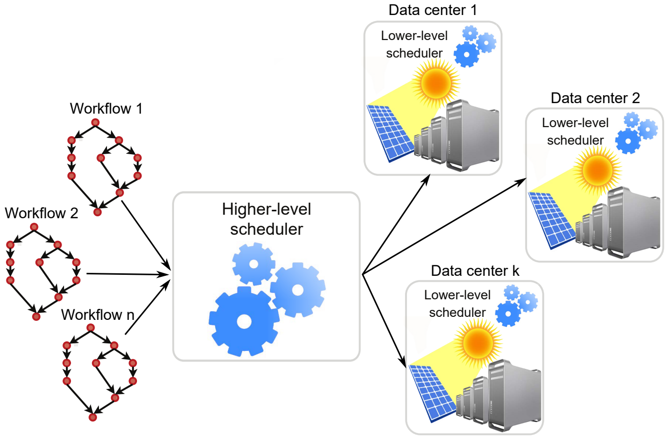

Figure 1. Overview of the higher-level scheduling problem.

center dc_r is comprised of a set of heterogeneous multi-core machines $S_r = \{s_1, \ldots, s_s\}$, each machine s_j characterized by its number of cores c_j, its performance in FLoating-point Operations Per Second (FLOPS) ops_j, and its energy consumption at idle usage e_j^{idle} and peak usage e_j^{max}.

- A set of n independent heterogeneous workflows $Q = \{q_1, \ldots, q_n\}$. Each workflow q has an associated soft deadline d_q which indicates its desired completion time. Each workflow q is a parallel application composed of a set of tasks $WT_q = \{wt_1, \ldots wt_m\}$ with dependencies among them.
- Each task α is characterized by its number of required floating point operations o_α, and its number of required cores nc_α.

Each workflow is represented as a *Directed Acyclic Graph* (DAG), i.e. a precedence task graph $q = (V, E)$, where the nodes of V are tasks α $(1 \leq \alpha \leq m)$ of the workflow. The set of directed edges E represents the dependencies between tasks. A partial order $\alpha \prec \beta$ models the precedence constraints: an edge $e_{\alpha\beta} \in E$ means that task β cannot start its execution before task α is completed.

We consider the multiobjective problem $\min\ (f_M, f_E, f_S)$, that proposes the simultaneous optimization of the makespan (f_M), energy consumption (f_E), and deadline violations (f_D).

The makespan evaluates the total time required to execute a set of workflows, as shown in Eq. (1), where $\vec{x}$ represents a scheduling solution, k is the number of data centers, and CT_r is the completion time of dc_r.

$$f_M(\vec{x}) = \max_{0 \leq r \leq k} CT_r \tag{1}$$

The energy consumption function is defined in Eq. (2). We use the energy model for multi-core architectures proposed in [4], where f_1 is the higher-level scheduling function, and

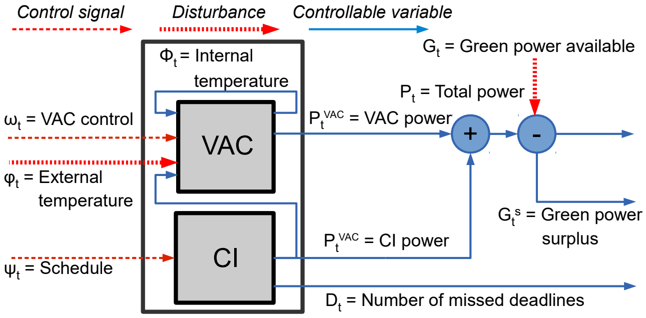

Figure 2. Schema of the green-powered data center's model.

f_2 is the lower-level scheduling function. The total energy consumption takes into account both the energy required to execute the tasks assigned to each computing resource within a data center, and the energy that each resource consumes in idle state.

$$f_E(\vec{x}) = \sum_{r \in DC} \sum_{\substack{q \in Q: \\ f_1(q)=r}} \sum_{\substack{wt_\alpha \in WT_q: \\ f_2(wt_\alpha)=s_j}} \frac{o(wt_\alpha)}{ops(s_j)} \times e_{s_j}^{max} + \sum_{s_j \in S_r} e_{s_j}^{idle} \tag{2}$$

The deadlines violations function is the total number of workflows that violate their deadline. Function f_D is defined in Eq. (3), where $violated(q) = 1$, if the deadline of workflow q is violated and 0, otherwise.

$$f_D(\vec{x}) = \sum_{q \in Q} violated(q) \tag{3}$$

We deal with scenarios composed of thousands of workflows (i.e. hundreds of thousands of tasks) to be scheduled onto a federation of several data centers comprised of hundreds to thousands of machines.

III. Lower-level scheduling problem in green-powered data centers

The lower-level algorithm schedules the execution of tasks of the workflows assigned to each data center. The data center model used for the local scheduling algorithm is based on the model proposed in [5]. Figure 2 shows a schema of the green-powered data center.

In our model, we consider two power consuming components: cooling devices or Ventilation-Air Conditioning (VAC), and computing infrastructure (CI). The thin-dotted arrows represent control signals; the fat-dotted arrows represent external disturbances; and the solid arrows represent controllable variables.

Control signals are controlled by the scheduling algorithm to modify the state of the data center: ω_t determines the operation of the VAC system, while the schedule ψ_t determines which servers are on and off, and where and when to execute each task.

Disturbances are variables which are non controllable: external temperature φ_t represents the temperature outside the data center, and available green power G_t represents the power generated by renewable energy sources.

Controllable variables are output variables whose values are determined by control signals and disturbances: the QoS variable D_t is the number missed workflow deadlines; the internal temperature Φ_t is the temperature inside the data center; P_t^{VAC} is the sum of air conditioning power and fan ventilation power; P_t^{CI} is the power consumed by the data center's machines; the green power surplus G_t^s determines the excedent of renewable energy not used by the data center, and finally, brown power B_t represents the amount of brown energy consumed from the power grid.

All variables defines the state of the data center at time t. Our objective is to schedule the control signal during for a sufficiently large planning horizon (K) so that the total power consumption P_t does not exceed a given power profile R_t, while the brown energy budget and QoS degradation are minimized, subject to maintaining the internal data center temperature Φ_t below its maximum operative value.

Formally, we want to simultaneously *minimize*:

$$f_P = \sum_{t=1}^{K} \begin{cases} (P_t - R_t)/\max(R_t) & \text{if } P_t > R_t \\ 0 & \text{if } P_t \leq R_t \end{cases} \tag{4a}$$

$$f_B = \sum_{t=1}^{K} B_t \times M_t^b \tag{4b}$$

$$f_D = \sum_{i=1}^{N} \begin{cases} FT(i) - D(i) & \text{if } FT(i) > D(i) \\ 0 & \text{if } FT(i) \leq D(i) \end{cases} \tag{4c}$$

Eq. (4a) specifies the power consumption of the system should not be above the reference power profile. Eq. (4b) is the total monetary cost of the energy consumption of the system. Eq. (4c) represents the total time of the deadline violations.

The data center in our model is comprised of two subsystems: VAC and CI. CI power P_t^{CI} is calculated as the sum of the total power consumption of all machines that are executing, idle and sleep at time t. We consider three VAC modes: air conditioning cooling, ventilation cooling, or no cooling. In the air conditioning mode, we use a conventional direct expansion air conditioner, which can be on and off. In ventilation cooling mode, the air conditioner is turned off and outside air is blown into the data center by a fan. The value of P_t^{VAC} and the cooling mode directly affect the temperature Φ_t in the datacenter. As shown in [5], the temperature follows an Auto-Regressive eXogenous (ARX) model, where the inputs are the air conditioning state, fan speed, outside temperature, machines load, free cooling damper state and temperature setpoint.

IV. MOEA SCHEDULING METHODS

MultiObjective Evolutionary Algorithms (MOEAs) are non-deterministic population-based metaheuristics inspired by biological evolution mechanisms for solving optimization, search, and learning problems [6].

In this work, we propose two MOEAs for solving the higher- and lower-level scheduling problems. Both MOEAs are based on the Non-dominated Sorting Genetic Algorithm, version II (NSGA-II) [7], a state-of-the-art MOEA that includes crowding features for preserving diversity in the population. Next, we describe each algorithm in detail.

A. Higher-level scheduler implementation

The higher-level scheduler assigns workflows to data centers. Schedules are encoded using a permutation of integers with length $n+k-1$, being n the number of workflows and k the number of data centers. Values from 0 to $n-1$ represent each workflow, while values from n to $n+k-1$ represent workflow groups, one group for each data center. Initial population is constructed using randomly created permutations using a uniform distribution. Selection is performed using the binary tournament method, crossover uses the Partially Matched Crossover (PMX) method, and mutation uses a simple swap method [6]. An additional evolutionary operator is included for repairing non-feasible solutions resulting from applying the crossover and mutation operator. This repair operator simply verifies, for each datacenter, if it can execute its assigned workflows. If not, then all the infeasible workflows are reassigned to the next datacenter which is able to execute them.

This algorithm is configured with a population size of 100 solutions, a crossover probability of 0.9, a mutation probability of 0.001, and a stopping criterion of 25,000 evaluations.

B. Lower-level scheduler implementation

The lower-level scheduler controls the energy consumption of the computing infrastructure, HVAC components, and the execution of the computing tasks. To tackle this efficiently, the NSGA-II is hybridized with a Local Search (LS) [8]. First, the NSGA-II solves the computing and HVAC energy consumption problem, and then the LS algorithm schedules the execution of the computing tasks subject to the energy constraints.

The NSGA-II solutions are encoded by an integer vector with $2K$ elements, where K is the number of time steps. The first K integers encode the cooling power for each time step, while the second K elements encode the server state for each time step. The server state determines which machines are on or off. The cooling power determines the heating dissipation capabilities by defining: *i)* when free cooling is applied and its fan speed, *ii)* when the air conditioning unit is operating, and *iii)* when both air conditioning and free cooling are off. For further details on the proposed encoding, please see [8]. Initial population is constructed using randomly created permutations using a uniform distribution function, and selection is performed using the binary tournament method. A three-point crossover method is used with three points p_1, p_2 and p_3; where p_1 is randomly selected in $(1, K)$, p_2 is K, and p_3 is $K+p_1$. This guarantees to produce feasible solutions. The mutation operator performs a genwise uniformly distributed random mutation. This algorithm is configured with a population of 50 solutions, a crossover probability of 0.9, a mutation probability of 0.01, and a stopping criterion of 500 generations. .

Finally, LS task-scheduler starts by generating an initial solution using the Best Fit Hole (BFH) algorithm proposed in [5]. The neighbourhood for the LS is constructed using a simple task moving operation which moves one task from its current machine to a new machine in some arbitrary position. We use dominance as an acceptance criterion, that is, we accept a new best solution only when it dominates the current best solution. Finally, we use a stopping criterion of 4000 iterations.

V. Preeliminary results

Currently, the proposed higher- and lower-level schedulers have been evaluated separately. Results show the proposed higher-level scheduler computes accurate schedules respecting all SLA agreements, with average makespan improvements of 20.3% and energy consumption improvements of 41.6%, when compared to a list-scheduling greedy algorithm [9]. Furthermore, preliminary results show average respected power profile improvements of 83.5%, budget improvements of 30.4%, and deadline improvements of 42.2%, when comparing the proposed lower-level scheduler with the scenario described in [8].

Results are encouraging and we expect that both schedulers will be able compute accurate solutions when working together.

VI. Conclusions

In this paper, we have proposed a hierarchical model for the workflow scheduling problem in distributed green-powered heterogeneous datacenters. We introduced two problem formulations, one for each level of the model hierarchy. We evaluate the accuracy of the proposed model and scheduling algorithms. In future work, we integrate both scheduling algorithms and evaluate a complete scheduling solution.

We expect to contribute to knowledge with a state-of-the-art model for the workflow scheduling problem in real-world distributed infrastructures. We will construct realistic instances of this problem and make them freely available for downloading. Finally, we will propose evolutionary schedulers for solving these problem instances accurately.

Acknowledgment

This work is partially funded by Universidad de la Repblica, ANII and PEDECIBA of Uruguay.

References

[1] A. Zomaya and S. Khan, *Handbook on Data Centers*. Springer, 2014.

[2] J. Koomey, "Growth in data center electricity use 2005–2010," 2011, Analytic Press.

[3] A. Tchernykh, L. Lozano, U. Schwiegelshohn, P. Bouvry, J. E. Pecero, S. Nesmachnow, and A. Y. Drozdov, "Online bi-objective scheduling for iaas clouds ensuring quality of service," *Journal of Grid Computing*, pp. 1–18, 2015.

[4] B. Dorronsoro, S. Nesmachnow, J. Taheri, A. Zomaya, E.-G. Talbi, and P. Bouvry, "A hierarchical approach for energy-efficient scheduling of large workloads in multicore distributed systems," *Sustainable Computing*, vol. 4, no. 4, pp. 252–261, 2014.

[5] S. Nesmachnow, C. Perfumo, and I. Goiri, "Controlling datacenter power consumption while maintaining temperature and QoS levels," in 3^{rd} *IEEE International Conference on Cloud Networking*, 2014, pp. 242–247.

[6] T. Bäck, D. Fogel, and Z. Michalewicz, Eds., *Handbook of evolutionary computation*. Oxford University Press, 1997.

[7] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. J. Wiley & Sons, Chichester, 2001.

[8] S. Iturriaga and S. Nesmachnow, "Multiobjective scheduling of green-powered datacenters considering QoS and budget objectives," in *IEEE Innovative Smart Grid Technologies in Latin America*, 2015, pp. 570–573.

[9] S. Iturriaga, B. Dorronsoro, and S. Nesmachnow, "Multiobjective evolutionary algorithms for energy and service level scheduling in a federation of distributed datacenters," *International Transactions in Operational Research*, 2016, submitted on 10-Nov-2015, awaiting response.

[10] F. Pinel, B. Dorronsoro, J. Pecero, P. Bouvry, and S. Khan, "A two-phase heuristic for the energy-efficient scheduling of independent tasks on computational grids," *Cluster Computing*, vol. 16, no. 3, pp. 421–433, 2013.

[11] I. Goiri, M. Haque, K. Le, R. Beauchea, T. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Matching renewable energy supply and demand in green datacenters," *Ad Hoc Networks*, vol. 25, Part B, no. 0, pp. 520–534, 2015.