Optimal and Distributed Task Scheduling in Volunteer Clouds

Stefano Sebastio
LIMS London Institute of Mathematical Sciences, London, UK
Email: stefano.sebastio@alumni.imtlucca.it
Giorgio Gnecco
IMT Institute for Advanced Studies, Lucca, Italy
Email: giorgio.gnecco@imtlucca.it
Alberto Bemporad
IMT Institute for Advanced Studies, Lucca, Italy
Email: alberto.bemporad@imtlucca.it

Abstract

The ever increasing request of computational resources has shifted the computing paradigm towards solutions where less computation is performed locally. The most widely approach adopted nowadays is represented by cloud computing. With the cloud, users can transparently access to virtually infinite resources with the same aptitude of using any other utility. Next to the cloud, the volunteer computing paradigm has gained attention in the last decade, where the spared resources on each personal machine are shared thanks to the users willingness to cooperate. In the volunteer paradigm each user shares a quote of its unused resources (e.g., during night or web-browsing activity) with other users, and receives other shared resources when he needs more than the local resources at his disposal. Cloud and volunteer paradigms have been recently seen as companion technologies to better exploit the use of local resources, also in perspective of green computing. Conversely, this scenario places complex challenges in managing such a large-scale environment, as the resources available on each node and the presence of the nodes online are not known a-priori. The complexity further increases in presence of tasks that have an associated Service Level Agreement specified, e.g., through a deadline. Distributed management solutions have then be advocated as the only approaches that are realistically applicable.

In this paper, we propose a framework to allocate tasks according to different policies, defined by suitable optimization problems. Then, we provide a distributed optimization approach relying on the Alternating Direction Method of Multipliers (ADMM) algorithm for one of these policies, and we compare it with a centralized approach. Results show that, when a centralized approach can not be adopted in a real environment, it could be possible to rely on the good suboptimal solutions found by the ADMM algorithm.

Keywords

cloud computing; distributed optimization; integer programming; combinatorial optimization; ADMM

I. Introduction

The ever growing demand of computational resources has shifted users from local computation on personal devices towards the adoption of centralized *cloud computing* technologies. Using the virtualized resources available on remote data centers, the cloud allows the access to virtually unlimited computational resources with the same aptitude of using any other utility service.

Together with the latest advance on virtualization technologies adopted by the cloud, in the last decade also the CPU manufacturers brought a great leap forward in performance starting the *multi-core era* even for the personal devices. Nowadays, desktop and laptop devices have resources largely unused for great part of the time (e.g., during web-browsing or text-editing activities) while having possibly scarce resources for other activities (e.g., for large-scale simulations or graphic-editing). These scenarios have opened the door to the growth of another ICT trend of the last years: the *volunteer computing* paradigm. Such a paradigm foresees the use of the spared resources on personal devices by all other network participants, relying on the users' willingness to cooperate. Therefore, the volunteer network turns out to be a heterogeneous (in terms of shared resources) and highly dynamic (not knowing a-priori the presence online of the volunteers) large-scale system (usually constituted from hundreds to thousands of nodes).

The cloud can then be enhanced through the combination with the volunteer paradigm. The combination is referred as the *volunteer cloud* [1]. Indeed, in specific application domains, such as latency dependent environments, the cloud still suffers its centralized nature where, obviously, its data-centers can not be located close to all the users. Examples of such domains have been described by Microsoft [2] as *cloudlets*, where the computational burden on mobile devices running virtual- or augmented- reality applications can be alleviated relying on low-latency (i.e., closely located) more powerful devices. The growing interest and success of the volunteer cloud is witnessed by the large number of existing platforms and projects based on such a paradigm, e.g., BOINC [3], HTCondor [4], OurGrid [5], Seattle [6] and SETI@home [7].

Typically, the volunteer nodes are organized and can communicate with each other through a Peer-to-Peer (P2P) overlay network. An example of the volunteer cloud network is depicted in Figure 1. The inherent characteristics of the volunteer cloud network makes impossible to rely on centralized solutions to effectively decide on the distribution of execution-oriented applications. Indeed, in such scenario a central node can constitute a bottleneck without being able to have updated knowledge on the nodes characteristics and their actual load in order to perform a global optimization. Distributed managing solutions have then been advocated as the only practical solution. Several distributed computing approaches [8]–[10] based on autonomic (i.e., self-managing) agents [11] have been explored in the literature. We only mention some works without dwelling on them since all the proposed solutions are *pure heuristic* solutions e.g., relying on Ant Colony Optimization [12], [13], Evolutionary Algorithm [14], Spatial computing [15] or Reputation-based [16] techniques.

The present work aims at proposing a framework to distribute task execution requests, according to different policies, each formalized as a mathematical optimization problem solved in a distributed fashion. In particular, we focus on the Alternating Direction Method of Multipliers (ADMM) to decompose the optimization problem, which

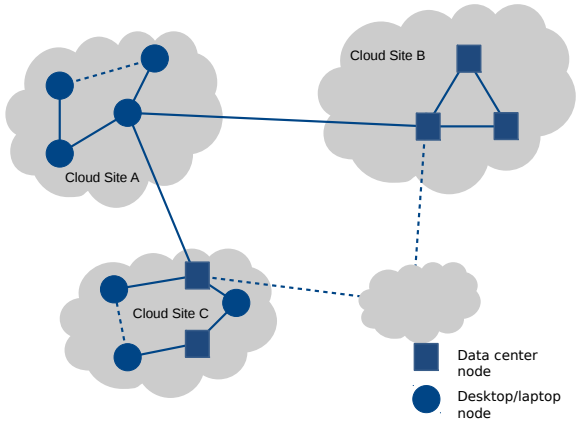


Fig. 1. Example of a volunteer cloud network: each cloud site can be constituted by data centers, personal devices or both.

is then distributed to and independently solved by the volunteer nodes.

Synopsis: The paper is structured as follows. Section II presents related work and particularly Section II-A briefly presents the ADMM algorithm. Our volunteer cloud model is presented in Section III while the optimization problem is formalized in Section IV. One of the policies allowed by our framework is accurately illustrated in Section V and solved with a centralized solution and with two different distributed ADMM variants. The considered scenario, its Matlab implementation relying on *CVX* [17] and *Gurobi* [18], is discussed in Section VI together with results and a numerical comparison of the proposed solutions. The description of the objective functions for the other policies of our framework is presented in Section VII. Finally, Section VIII concludes the work with some final remarks and outlines other current and future research efforts.

II. Related Work

Recently, the volunteer cloud has been modeled in [19] (called the cloud-assisted approach therein). Criticality and infeasibility of a centralized solution, to determine if a subset of tasks would be better executed by the cloud or by the volunteers, are highlighted in that paper referring to the problem as a *cloudy knapsack problem* i.e., an online knapsack problem but in presence of a limited knowledge on the current status (*cloudy view*). Despite in the paper motivation the authors argue for the need of a distributed decision making, later in the work the focus is only on the characterization of the knapsack problem where the available budget (corresponding to the available resources) is known only in terms of a probability distribution. Authors prove an upper bound on the performance of their knapsack algorithm where a probability distribution models the budget uncertainty.

In [20] authors accurately model the hybrid cloud environment (i.e., composed by on-promise and off-promise machines) as a mixed-integer nonlinear programming problem specified in AMPL [21]. The authors' goal is the minimization of the monetary cost to run a set of both compute- and data- intensive tasks. Actually, their implementation takes even the deadline into account in the cost function, although not in a explicit way. Indeed, their

cost function mixes the time for completing the tasks with the actual price for using a given service. Although there are similarities with our model, the major differences are represented by their indifference to distributed solutions and by not considering the actual load on machines.

The above mentioned model is further extended in [22] to deal with workflows instead of independent tasks. The authors define the workflow as a directed acyclic graph and then group the tasks into levels. In each level, tasks are independent among them and intermediate deadlines are assigned.

DONAR [23] is a distributed system that aims at mapping replica selection in the cloud, i.e., the choice of a particular location based on performance, load and cost in presence of geo-replicated services. Its optimization problem is designed in a way to be easily decomposed into independent subproblems that can be solved locally (since even the constraints can be split on the nodes). The global constraint (related to the bandwidth cap) could be temporarily violated by the intermediate solutions of the algorithm. Moreover, the problem turns out to be a convex optimization problem, where the optimization variables are constituted by the portions of traffic load that should be mapped from each client to each node. A similar problem is dealt in [24]. There a parallel implementation of ADMM is adopted to solve a convex optimization problem that has been validated on a trace-driven simulation relying on the Wikipedia requests trace.

DONAR is extended in [25] considering energy efficient solutions as a "first-class" design constraint (i.e., a priority) alongside performance, while distributing tasks in the cloud. In particular, the optimization function takes into account the energy costs in the presence of data-intensive services deployed in the cloud. The objective function of the convex optimization problem in [25] is constituted by the sum of local objective functions, and the problem is solved using the dual decomposition and the consensus-based distributed projected subgradient method (CDPSM) [26].

The task allocation problem in presence of hard deadlines and classes of priorities has been studied in [27] for powertrain automotive applications. Their problem formulation turns out to be a mixed integer linear programming problem solved using CPLEX [28]. The different characteristics of their domain of application makes difficult a direct comparison with our approach, despite it is possible to recognize that their solution assumes a centralized approach where an updated global knowledge on the load of the nodes is available (since their problem does not need to deal with a large-scale system).

Main Contributions: All in all, our paper contributes to the existing literature related to the volunteer cloud in that it proposes:

- a framework to accommodate different task distribution policies in large-scale distributed cloud environments;
- a mathematical formulation of the optimization problem associated with such a framework, which is driven by real system requirements of the volunteer cloud and takes into account various issues, such as, FIFO queue, tasks with deadlines, the actual load on the machines, the need to adopt a distributed solution;
- the application of the distributed ADMM algorithm for solving the above optimization problem.

A. The ADMM: Alternating Direction Method of Multipliers

The increasing size and complexity of many datasets pose challenges to analyze such problems especially when the data under analysis originate from decentralized sources. Distributed solutions are thus advocated as promising and desirable to work in such large-scale domains.

Recently the *Alternating Direction Method of Multipliers (ADMM)* is gaining momentum [29] in the field of distributed optimization, despite the method has been developed almost fifty years ago. The ADMM is an algorithm to solve *convex optimization* [30] problems in a distributed / parallelized fashion adopting a *decomposition-coordination* method, in which the problem is subdivided in small problems that can be solved independently, under the "coordination" of a central node, to solve the original problem. The method has its roots on the methods of dual decomposition and augmented Lagrangian.

The method is usually applied to convex optimization problems of the form:

$$minimize\ f(x) + g(z)\ , \tag{1a}$$

$$subject\ to\ Ax + Bz = c\ , \tag{1b}$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ are convex functions, and $x \in \mathbb{R}^n, z \in \mathbb{R}^m, A \in \mathbb{R}^{p\times n}, B \in \mathbb{R}^{p\times m}, c \in \mathbb{R}^p$. Then, given the *penalty parameter* $\rho > 0$, the augmented Lagrangian is expressed as:

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2 \tag{2}$$

and, finally, each iteration of the ADMM algorithm is defined as follows:

$$x^{k+1} := \underset{x}{\operatorname{argmin}}\, L_\rho(x, z^k, y^k)\ , \tag{3a}$$

$$z^{k+1} := \underset{z}{\operatorname{argmin}}\, L_\rho(x^{k+1}, z, y^k)\ , \tag{3b}$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)\ , \tag{3c}$$

where the penalty parameter in the augmented Lagrangian constitutes the *step size* while updating the *dual variable* y. ADMM is often transformed in the equivalent *scaled* form for convenience of implementation, resulting in a more compact expression. The scaled form of ADMM is expressed scaling the dual variable as $u = \frac{1}{\rho}y$.

We refer the interested reader to [29] for the proof of convergence (under mild conditions) and methods to compute the stopping criteria of the ADMM algorithm. A method for the optimal choice of the ρ parameter in ADMM that ensure the smallest possible convergence factor for quadratic problems is provided in [31].

If the problem is not convex, it is still possible to apply ADMM, despite this should be just considered as a heuristic in that case. An extension of the ADMM algorithm for nonconvex problems through the use of heuristics is proposed in [32]. There, integer and continuous variables of the problem are identified, and the source of non

convexity are the first variables. The continuous relaxation of the problem is initially considered obtaining bounds on the optimal solution value of the original problem and thus using these hints to reduce the search space. The basic idea is to start with ADMM and integer variables, but once these variables are observed to do not change values for some iterations, their value is fixed (so actually excluding them from the variables set and changing them to constants), to simplify the following ADMM iterations that could refine the solution for the continuous variables. Their *Release and Fix* approach swings from the original to the simplified problem (i.e., without the integer variables) at each iteration. Numerical results demonstrate that their distributed approach can reach a 1% relative optimality gap with a superior solution time compared to a centralized solution, while it is executed on a computer cluster.

Another interesting application of ADMM for nonconvex problems in computer vision is studied in [33]. There, the authors adopt a tree-based decomposition of discrete labeling problems in large-scale random fields. The decomposition exploits the submodularity of the non-convex subproblem functions allowing to solve them exactly. Despite the lack of guarantees on convergence for the original problem, their empirical results are encouraging. A parallel implementation relying on both CPU and GPU cores allowed the authors of [33] to obtain a significant speed-up while benchmarking low-level vision problems such as stereo correspondence, image segmentation and denoising.

III. Model

The volunteer cloud environment is constituted by a large number of nodes that share their computational and/or storage resources with the other network participants. Usually the cloud architecture relies on a centralized load balancer which collects information from and distributes commands to the nodes. Unfortunately, in the volunteer cloud, the large-scale nature of such a kind of network (in the order of hundreds or thousands of nodes) and the volatility of the nodes participation (i.e., it is not possible to know a-priori the time a node will join or leave the network), makes the dissemination of resources and load characteristics to a central load balancer a hard, if not impossible, job. Consequently, tasks execution requests cannot be distributed by solving a completely centralized global optimization problem.

Distributed optimization techniques are then advocated as promising alternatives. In this work we focus on such a technique, to optimize the task allocation among the volunteer nodes. This section presents the volunteer cloud computing model that will be used in Section IV to define the optimization problem of interest. Since our focus is on the distribution of tasks execution requests, the main component of our model are the tasks (Section III-A) and the resources (Section III-B). In such a large-scale decentralized environment each node acts as both producer and consumer of tasks execution requests.

A. Tasks

Application programs are modeled as sets of independent *tasks*. Each task is defined by its duration, the degree of parallelism that it is able to exploit, and its minimum memory requirement. Our model assumes that the task

duration can be predicted with accuracy. Moreover, we assume a perfect linear speed-up for the parallelism.

Definition 3.1 (task): A *task* is a tuple $\langle \delta, \rho, \mu \rangle$, where $\delta \in \mathbb{R}^+$ is the task duration (expressed in clock cycles), $\rho \in \mathbb{N}^+$ is the degree of parallelism of the task, and $\mu \in \mathbb{N}^+$ is the memory requirement of the task.

Adopting a perfect linear speed-up model, a degree of parallelism ρ means that a task with duration δ can be ideally executed with a CPU that has ρ computational units (e.g., cores) in time δ/ρ. If less than ρ computational units are available, the time to execute the task will be bounded in terms of the number of available units (as expressed in the following Equation (4)).

Definition 3.2 (task execution request): A *task execution request* is a tuple $\langle \delta, \rho, \mu, \tau_a, \tau_d \rangle$, where $\langle \delta, \rho, \mu \rangle$ is a task, $\tau_a \in \mathbb{R}^+$ is the task arrival date, and $\tau_d \in \mathbb{R}^+$ is its termination deadline. The deadline can constitute one of the parameters specified in the task Service Level Agreement (SLA) [34].

We assume that for each task a single execution is required, i.e., it is enough that a single node takes the execution request in charge.

B. Virtual Resources

The volunteer nodes provide an isolated environment for each task execution request through a Virtual Machine (VM), modeled by memory and processing units. In our scenario, the main feature of interest of the VMs is the latter since our main concern is the running time of tasks, which we consider to be computation-intensive (rather than data- or communication- intensive). We assume that, given the voluntary participation of the nodes and the not dedicate use of the physical machine (only spare resources are shared by users), each node runs a single VM. Therefore, to provide an isolated environment for each user, tasks execution requests are processed one at a time. For the sake of simplicity we consider that VMs have homogeneous processing units i.e., all the cores in the same VM have the same clock frequency.

Definition 3.3 (Virtual Machine, VM): A *VM* is a tuple $\langle \kappa, \phi, \nu \rangle$ where $\kappa \in \mathbb{N}^+$ is the number of cores, $\phi \in \mathbb{N}^+$ is their frequency, and $\nu \in \mathbb{N}^+$ is the amount of memory.

Definition 3.4 (Execution time): The execution time $e(T, VM)$ required for completing a task $T = \langle \delta, \rho, \mu \rangle$ on a virtual machine $VM = \langle \kappa, \phi, \nu \rangle$ whose cores have frequency ϕ is defined by the following equation:

$$e(T, VM) = \frac{\delta}{\phi \cdot min\{\rho, \kappa\}} \ , \tag{4}$$

i.e., the task duration on a single core which is equal to δ/ϕ is divided by the maximum degree of parallelism that can be exploited, which is bounded by both the amount of available cores (κ) and the degree of parallelism of the task (ρ).

A task $T = \langle \delta, \rho, \mu \rangle$ can be executed on a virtual machine $VM = \langle \kappa, \phi, \nu \rangle$ only if the following memory requirement constraint is satisfied:

$$\mu \leq \nu \ . \tag{5}$$

Tasks accepted for execution are added to the node execution queue and executed according to a FIFO policy. If not otherwise specified, a machine VM accepts a task T in its execution queue only if it can respect the task deadline.

IV. The Optimization Problem

In our model, resource capacities (i.e., memory and cores characteristics) and tasks already on the node execution queue pose limits on the following task execution requests that a node can accept respecting the associated SLAs.

In a first approximation, task execution requests can be dispatched periodically according to an *allocation period*. In each period, task requests should be distributed optimizing with respect to an allocation policy defined by the volunteer cloud manager. All the tasks generated after an allocation period are assigned to the subsequent *task set* and dispatched in the next allocation period. The period itself constitutes a parameter which can vary the performance of the system. A short period could allow a more timely allocation of the tasks while resulting more computationally expensive. Conversely, a large period can possibly reduce the chance to satisfy task requests that have more stringent deadlines, despite they could be less computational demanding.

This section presents the data structures (Section IV-A) used in the definition and in the resolution of our optimization problem. Moreover a brief description of some allocation policies that can be implemented within our framework is presented at the end of the section (Section IV-B).

A. Data structures

In each allocation period, having K tasks and N nodes, the data structures used hereinafter (following the notations used in the definitions in Sections III-A and III-B) are as follows:

- $taskSet = \begin{pmatrix} \tau_{a1} & \tau_{d1} & \rho_1 & \delta_1 & \mu_1 \\ \vdots & & \cdots & & \vdots \end{pmatrix} \in \mathbb{R}^{K \times 5}$, where the tasks are sorted according to their deadline τ_d, from closest to farthest (this sorting should increase the chance to execute more tasks respecting their deadline, see the notes at the end of this section).
- $nodeSet = \begin{pmatrix} \phi_1 & \kappa_1 & \nu_1 \\ \vdots & \cdots & \vdots \end{pmatrix} \in \mathbb{R}^{N \times 3}$, with the characteristics of the VMs.
- $freeTimes_t = \begin{pmatrix} free_1^t \\ \vdots \end{pmatrix} \in \mathbb{R}^N$, with the times at which each node completes the execution of all the tasks that are already in its execution queue (i.e., execution requests that have been assigned to the nodes in the previous allocation periods but that have not yet been completed).
- $X(:= taskAlloc) \in \mathbb{Z}_2^{K \times N}$ is the *task allocation matrix*, a binary matrix ($\mathbb{Z}_2 = \{0, 1\}$) where the element x_{ij} is equal to 1 if the task i is executed on node j and 0 otherwise. Thus, reading it by column one gets the tasks accepted by each node in the current allocation period; while, reading it by row one gets the node(s) that has (have) accepted the task for execution, if any. This matrix represents the *decision variable* of our optimization problem.
- $execEnd_X \in \mathbb{R}^K$, collects the times at which each task execution is completed, taking into account the subset of tasks each node has accepted for execution in the current allocation X (a more formal definition and a

step-to-step computation is given in the following, see Section V). The components of $execEnd_X$ assume value 0 for the tasks that are not accepted by any node.

- $executedTasks := \sum_{l=0}^{K-1} X(l,j) \in \mathbb{R}^N$, collects the number of tasks, belonging to the current allocation period, that each node has taken in charge to execute.

In each allocation period, tasks are sorted and evaluated for execution according to their deadline, thus the considered allocation scheduling policy is the *EDF* (Earliest Deadline First). While, as stated before, once the tasks have been accepted are executed according to a FIFO policy. This choice is dictated by the willingness to increase the chance to accept as many tasks as possible (respecting their deadline), while maintaining a low complexity, avoiding to rearrange the execution queue every time new tasks are accepted for execution in the next allocation period.

It is worth noting that the choice of the scheduling policies (for both allocation and execution) affects the selection of tasks that are assigned to the nodes. Indeed, a different sorting in the tasks to be executed or allocated could bring to a different time for completing the tasks ($execEnd_X$) and in turn to a different allocation of the tasks (X). In other terms, with a different sorting of the execution requests a *feasible* allocation could become *unfeasible* and vice-versa.

For instance, let us consider a long-running task with a relaxed deadline that is evaluated for allocation before other small-running tasks but with high demand on deadline: the first task can constitute a bottleneck for all other tasks, which could not be executed respecting their SLAs. A simple tasks sorting (where the tasks with a high demand on deadline are evaluated first) could allow to execute more tasks, while the tasks with a relaxed deadline could be executed later without any effect on their SLA. Obviously, other, more sophisticated and effective scheduling disciplines could be implemented and evaluated.

B. Allocation policies

The optimization function considered in our framework could take into account different performance metrics while allocating the tasks to the nodes. In this section we briefly describe five different allocation policies:

- **Number of executed task**: it attempts to maximize the number of executed tasks (Section V).
- **Termination time**: it attempts to minimize the time at which the execution ends for the entire task set (Section VII-A).
- **Response time**: it attempts to minimize the sum of the response times for each task in the set, even ignoring the SLA if needed. Thus, all the execution requests are assigned (Section VII-B).
- **Fair strategy**: it attempts to evenly distribute the tasks among the participant nodes (Section VII-C).
- **Green strategy**: it attempts to consolidate the tasks among few participant nodes (Section VII-D).

Each of the above mentioned policies, during the formalization of the optimization problem, will require one or a subset of the *constraints* which are listed here:

1) each node can execute a task only if it has enough memory:

$$\mu \leq \nu \quad \forall\ task\ executed\ on\ a\ node\ ; \tag{6}$$

2) each task must be assigned to one and only one node, even if there is no node that can satisfy its deadline requirement:

$$\sum_{j=0}^{N-1} X(l,j) \;=\; 1 \quad \forall\, l \;\in\; \{0,1,...K-1\}\ . \tag{7}$$

If it is impossible to satisfy all the tasks deadlines, other node policies will discern which tasks should be executed first and which ones discarded or completed lately (i.e., on a *best-effort* basis not fulfilling the deadline specified in the SLA);

3) each task should be executed respecting its deadline:

$$execEnd_X(l) \leq \tau_d \quad \forall\ task\ l\ executed\ on\ a\ node\ j\ ; \tag{8}$$

4) each task can be executed by only one node:

$$\sum_{j=0}^{N-1} X(l,j) \;\leq 1 \quad \forall\, l \;\in\; \{0,1,...K-1\}\ . \tag{9}$$

V. Maximize the number of executed tasks

In this section we focus on the accurate description of one of the policies of our framework, namely, the maximization of the number of executed tasks. Without loss of generality other policies can be implemented straightforwardly as exemplified later in Section VII.

The policy described in this section aims at optimizing the number of tasks that are executed respecting their deadline. The associated optimization problem can then be written as the maximization of:

$$\sum_{j=0}^{N-1} \sum_{l=0}^{K-1} taskAlloc(l,j)\ , \tag{10}$$

where the inner summation counts the tasks executed on the node j. The constraints 1, 3 and 4 must be met.

Thus, more formally, the optimization problem is:

$$maximize \sum_{j=0}^{N-1} \sum_{l=0}^{K-1} X(l,j)\ , \tag{11}$$

subject to:

$$\sum_{j=0}^{N-1} (diag(\mu)\cdot X)(l,j) \leq X\cdot\nu \qquad \forall\, l \;\in\; \{0,...K-1\}\ , \tag{12a}$$

$$\sum_{j=0}^{N-1} X(l,j) \;\leq 1_K \qquad \forall\, l \;\in\; \{0,...K-1\}\ , \tag{12b}$$

$$execEnd_X(l) \leq \tau_d \qquad \forall\ task\ l \in \{0,...K-1\}\ executed\ on\ a\ node\ j \in \{0,...N-1\}\ , \tag{12c}$$

where all the inequalities are component-wise comparison and the $\cdot$ symbol denotes the matrix product. The left side of Equation (12c), with the implicit dependency from X, expresses the time required to complete the execution of each task l according to the node j that has taken in charge its execution, if any.

Throughout the paper, if not otherwise specified, the index l refers to a task, j refers to a volunteer node, while i refers to one of the constraints related to the optimization problem under study.

The above mentioned optimization problem has been solved in *Matlab* relying on the *CVX* modeling system, which is based on the Disciplined Convex Programming ruleset (DCP). The DCP ruleset [35] allows one to describe problems that are convex by construction. Objective functions and constraints are expressed relying on a set of basic atoms and a restricted set of operation rules that preserve convexity. In this way, DCP is able to keep track of affine, convex and concave expressions. If a problem does not adhere to the DCP it is rejected even if it is convex, the converse instead can not happen (i.e., a problem recognized as convex by DCP is always convex). *CVX* [17], [35] is a modeling system that, relying on DCP, is able to simplify the task of specifying the problem for: linear (LPs), quadratic (QPs), second-order cone (SOCPs) and semidefinite (SDPs) programs. Recently, *CVX* has been extended to support *mixed-integer* models where one or more variables have integer or binary values (as X in our problem). These models are defined as *Mixed-Integer Disciplined Convex Problems (MIDCPs)*. Strictly speaking, these are nonconvex optimization problems for which, if the integrality constraint is removed, one obtains a Disciplined Convex Problem.

While all other constraints can be written straightforwardly, the implementation of Equation (12c) in *CVX* requires some computations to respect the DCP, since the ruleset implies that convexity should be an explicit goal of the modeling process itself. In the following we describe the computations required to express the last constraint in *CVX*. Recalling that:

$$e(\delta, \phi, \rho, \kappa) := \frac{\delta}{\phi \cdot min\{\rho, \kappa\}} \in \mathbb{R}^{K \times N} \ , \tag{13}$$

in the full matrix in Equation (13) each element e_{lj} represents the computational time required to execute the task l on node j.

A vector $initTime$ with the times at which each node can start the execution in the current allocation period, can be built considering the current time t and the residual computation from the previous time periods ($freeTimes_t$):

$$initTime := max\{t, freeTimes_t\} \in \mathbb{R}^{1 \times N} \ . \tag{14}$$

Thus, in order to evaluate the time required to execute the tasks in the current period, the node should consider the computation in progress and the tasks already in its queue (and accepted for execution in the previous allocation periods). An intermediate step of computation requires adding the $initTime$ to the tasks execution times:

$$timeToExecTask = e + \begin{pmatrix} initTime \\ \\ 0 \end{pmatrix} = \begin{pmatrix} e_{11} + tt_1 & e_{12} + tt_2 & \cdots \\ \\ e_{21} & e_{22} & \cdots \\ \\ \vdots & \ddots & \vdots \end{pmatrix} \ . \tag{15}$$

Then, assuming a FIFO policy during the task execution, if every task was executed by each node, one would

obtain:

$$execCum = cumulativeSum(timeToExecTask) = \begin{pmatrix} e_{11} + tt_1 & e_{12} + tt_2 & \cdots \\ e_{21} + e_{11} + tt_1 & e_{22} + e_{12} + tt_2 & \cdots \\ \vdots & \ddots & \vdots \end{pmatrix} , \tag{16}$$

where the *cumulativeSum* function (*cumsum* in Matlab) returns a matrix containing the cumulative sums for each column starting from the first row.

Defining $\overline{X}$ as the *complement matrix* of the binary matrix X with the tasks that are not executed by the nodes, with further computation it is possible to compute the time at which the execution of each task is completed ($execEnd_X$) according to the node that has taken it in charge, if any:

$$execRemovedNotExec = execCum - \overline{X}. * execCum \ , \tag{17a}$$

$$finishingTimeDirty = execRemovedNotExec - cumulativeSum(\overline{X}. * e) \ , \tag{17b}$$

$$execEnd_X = finishTime = \max_{\forall \ row \ l} \{finishingTimeDirty\} = \sum_{by \ row} (finishingTimeDirty^+) \ , \tag{17c}$$

where $.*$ is the element-by-element product of the matrices. Equation (17a) sets to zero the cells for the tasks that are not executed by the corresponding node in Equation (16). However, this operation is not enough to obtain the time needed to complete the tasks executions. Indeed, observing Equation (16), it is possible noting that since we are considering a FIFO policy, setting to zero the cells for the tasks not executed is not enough. For instance, if the first node executes the second but not the first task, with Equation (17a) the first cell is set to zero but the element in the second row first column still maintains a dependency from a task that is not actually executed (e_{11} in this example). Equation (17b) is introduced to remove these undesired dependencies. Conversely, this operation builds a *dirty* matrix, since its matrix elements could have negative values while subtracting from cells that where already set to zero from the first operation (Equation (17a)). Finally, Equation (17c) builds the tasks finishing times for the tasks that are actually executed, considering the nodes that have accepted their executions (where the f^+ operator chops off the negative values). Note that the last equality in Equation (17c) is obtained considering that for each row there can be only one positive element.

Other simpler operations could have led to the same result but requiring, at some point, the product of the optimization variable with itself (i.e., $X. * X$). For instance, a possible easily attained computation of $execEnd_X$ could be given by $cumulativeSum(e. * X). * X + tmp. * X$, where the $cumulativeSum$ is needed by the FIFO queue and the second element-by-element product sets to zero the matrix elements corresponding to the tasks not executed by a node. Unfortunately, despite in our definition X is binary, this operation brings the expression to be not convex as well as not comply with the DCP ruleset that considers the operation just as the square of the optimization variable.

Note on our optimization problem: The problem formulated above (Equations (10)-(12c)) can be seen as an *Integer Programming (IP)* problem (since X assumes integer values) whose goal consists in finding an optimal allocation of the tasks. With a significant implementation effort, particularly when K and / or N are large, Equation (17c) can be reformulated substituting the max operator with linear inequalities, therefore obtaining an Integer Linear Programming (ILP) problem. ILP problems are well known to be *NP-complete* in literature. Moreover, in our formulation the optimization variable is binary. The *0-1 ILP* (also referred as *Binary ILP*, BILP) expressed as a decision problem constituted by only restrictions without an optimization function, has been classified as one of the *Karp's 21 NP-complete problems* [36].

Relaxing the X variable in the real domain, a convex problem could be obtained which could provide an useful upper bound for the optimal solution value of the original problem.

A. ADMM: unscaled form and augmented Lagrangian

The optimization problem, as expressed in the previous section, allows *CVX* and *Gurobi* to solve the problem, but it is still formulated as a completely centralized approach. As stated above, such an approach can be hard to be applied in practice in the domain of our interest, since distributing resource characteristics and load of each node (with hundreds or thousands of nodes) is challenging. A possible way to solve the problem in a distributed fashion can be constituted by the iterative ADMM algorithm [29].

The original optimization problem is reported here:

$$maximize \sum_{j=0}^{N-1} \sum_{l=0}^{K-1} X(l,j) \ , \tag{18}$$

subject to:

$$\mu X \leq X\nu \ , \tag{19a}$$

$$\sum_{j=0}^{N-1} X(l,j) \ \leq 1_K \ , \tag{19b}$$

$$execEnd(X) \leq \tau_d \ . \tag{19c}$$

For ease of exposition, we rename the functions as: $f(X) = -\sum_{j=0}^{N-1} \sum_{l=0}^{K-1} X(l,j)$, $\mu X = m_r(X)$ (the requested memory), $X\nu = m_a(X)$ (the memory available) and then $m_r(X) - m_a(X) = m(X)$ (memory constraint), $\sum_{j=0}^{N-1} X(l,j) = s(X)$ (number of executions / nodes for each task), $execEnd(X) = e(X)$. Finally, the problem in Equations (18)-(19) can be rewritten as a minimization problem as follows:

$$minimize \quad f(X) \ , \tag{20}$$

subject to:

$$m(X) \leq 0_K \ , \tag{21a}$$

$$s(X) \ \leq 1_K \ , \tag{21b}$$

$$e(X) \leq \tau_d \ . \tag{21c}$$

The ADMM method (and more in general the augmented Lagrangian) does not explicitly allow the use of inequality constraints, thus the use of a *slack vector* z for each constraint is required [37]. The inequality is implicitly considered in the optimization function requiring z to be positive through the indicator function $I_+(z)$ (that can be seen as an infinite penalty on the optimizing function when z is negative, and is zero otherwise).

$$minimize \quad f(X) + I_+(z_1) + I_+(z_2) + I_+(z_3) \ , \tag{22}$$

subject to:

$$h_1(X, z_1) = m(X) + z_1 = 0 \ , \tag{23a}$$

$$h_2(X, z_2) = s(X) - 1 + z_2 = 0 \ , \tag{23b}$$

$$h_3(X, z_3) = e(X) - \tau_d + z_3 \ . \tag{23c}$$

The augmented Lagrangian used in the ADMM algorithm can thus be easily written as:

$$L_\rho(X, z, y) = f(X) + \sum_{i=1}^{3} y_i^T h_i(X, z_i) + \sum_{i=1}^{3} \frac{\rho}{2} \|h_i(X, z_i)\|_2^2 + I_+(z_1) + I_+(z_2) + I_+(z_3) \ , \tag{24}$$

and the ADMM (in unscaled form) is expressed as:

$$X^{k+1} = \underset{X}{\operatorname{argmin}} \, L_\rho(X, z_i^k, y_i^k) \ , \tag{25a}$$

$$z_i^{k+1} = max\{\underset{z_i}{\operatorname{argmin}} \, L_\rho(X^{k+1}, z_i, y_i^k), \ 0\} \ , \tag{25b}$$

$$y_i^{k+1} = y_i^k + \rho(h_i(X^{k+1}, z_i^{k+1})) \ , \tag{25c}$$

with $i = 1, 2, 3$ and $\rho > 0$ (the augmented Lagrangian parameter).

It is worth noting that the update of z_i in Equation (25b) has required the use of the max operator to take into account that z_i cannot be negative (recalling the introduction of the I_+ function in the optimization problem to deal with the inequality constraints).

B. ADMM: scaled form and augmented Lagrangian

For ease of implementation the ADMM can be rewritten in scaled form. The terms in the augmented Lagrangian containing $r_i = h_i(X, z_i)$ can be transformed as:

$$\sum_{i=1}^{3} y_i^T r_i + \sum_{i=1}^{3} \frac{\rho}{2} \|r_i\|_2^2 = \sum_{i=1}^{3} \frac{\rho}{2} \|r_i + \frac{1}{\rho} y_i\|_2^2 - \sum_{i=1}^{3} \frac{1}{2\rho} \|y_i\|_2^2 \underset{u_i = \frac{1}{\rho} y_i}{=} \sum_{i=1}^{3} \frac{\rho}{2} \|r_i + u_i\|_2^2 - \sum_{i=1}^{3} \frac{\rho}{2} \|u_i\|_2^2 \ , \tag{26}$$

becoming:

$$L_\rho(X,z,u) = f(X) + I_+(z_1) + I_+(z_2) + I_+(z_3) + \sum_{i=1}^{3} \frac{\rho}{2}\|h_i(X,z_i) + u_i\|_2^2 - \sum_{i=1}^{3} \frac{\rho}{2}\|u_i\|_2^2 \ , \tag{27}$$

and finally the ADMM expressed in scaled form is:

$$X^{k+1} = \underset{X}{\operatorname{argmin}} \Big(f(X) + \sum_{i=1}^{3} \frac{\rho}{2}\|h_i(X,z_i) + u_i^k\|_2^2 \Big) \ , \tag{28a}$$

$$z_i^{k+1} = max\Big\{ \underset{z_i}{\operatorname{argmin}} \Big(\sum_{i=1}^{3} \frac{\rho}{2}\|h_i(X,z_i) + u_i^k\|_2^2 \Big), 0 \Big\} \ , \tag{28b}$$

$$u_i^{k+1} = u_i^k + h_i(X^{k+1}, z_i^{k+1}) \ . \tag{28c}$$

The scaled form makes easier the implementation in *Matlab* expressing the ADMM in a more compact and convenient form. Obviously, the two forms (scaled and unscaled) are equivalent.

C. Distributed ADMM using Global Consensus

The ADMM formulations defined in the previous Sections V-A and V-B allow only a marginal parallelization of the algorithm on z_i and y_i (u_i), where $i = \{1,2,3\}$. Reasoning about a real implementation of one of these formulations in the volunteer cloud environment, the ADMM turns out to be parallelized in only three computational threads but definitely not distributable on all the volunteers. To better exploit the distribution capabilities of the ADMM algorithm on all the volunteers, it is required to have X partitionable and f separable with respect to this partition. The original problem can be rewritten as a *Global Consensus* problem where the N nodes want to converge towards an optimal shared solution X. Expressing as $f_j(X) = -\sum_{l=0}^{K-1} X(l,j)$ the number of tasks per node (the "minus" sign is used to express the optimization problem as a minimization one), the problem in Equations (22)-(23) has the form:

$$minimize \quad \sum_{j=0}^{N-1} f_j(X) + \sum_{i=1}^{3} I_+(z_i) \ , \tag{29}$$

subject to:

$$h_i(X,z_i) = 0 \quad \forall i \ , \tag{30}$$

that can be distributed over the nodes introducing the local variables X_j, one for each node. Each X_j is a *copy* of the original X, and is updated locally and independently by each node. The global consensus variable Z is added. For ease of presentation the range of j index is shifted from 1 to N in the following:

$$minimize \quad \sum_{j=1}^{N} f_j(X_j) + \sum_{i=1}^{3} I_+(z_i) \ , \tag{31}$$

subject to:

$$X_j - Z = 0 \quad \forall j = 1, \dots, N \ , \tag{32a}$$

$$h_i(Z, z_i) = 0 \quad \forall i = 1, 2, 3 \ . \tag{32b}$$

The augmented Lagrangian is:

$$\begin{aligned} &L_\rho(X_1, \dots, X_j, \dots, X_N, Z, z_1, z_2, z_3, Y_1, \dots, Y_N, y_{N+1}, y_{N+2}, y_{N+3}) \\ &= \sum_{j=1}^{N} \Big(f_j(X_j) + Y_j(X_j - Z) + \frac{\rho}{2}\|X_j - Z\|_2^2 \Big) + \sum_{i=1}^{3} \Big(y_{N+i} h_i(Z, z_i) + \frac{\rho}{2}\|h_i(Z, z_i)\|_2^2 + I_+(z_i) \Big) \ . \end{aligned} \tag{33}$$

The ADMM for the global consensus problem becomes:

$$X_j^{k+1} = \underset{X_j}{\operatorname{argmin}} \Big(f_j(X_j) + Y_j^k (X_j - Z^k) + \frac{\rho}{2}\|X_j - Z^k\|_2^2 \Big) \quad \forall j = 1, \dots, N \ , \tag{34a}$$

$$Z^{k+1} = \underset{Z}{\operatorname{argmin}} \left(\sum_{j=1}^{N} \Big(- Y_j^k Z + \frac{\rho}{2}\|X_j^{k+1} - Z\|_2^2 \Big) + \sum_{i=1}^{3} \Big(y_{i+N}^k h_i(Z, z_i^k) + \frac{\rho}{2}\|h_i(Z, z_i^k)\|_2^2 \Big) \right) \ , \tag{34b}$$

$$z_i^{k+1} = max\Big\{ \underset{z_i}{\operatorname{argmin}} \Big(y_i^k h_i(Z^{k+1}, z_i) + \frac{\rho}{2}\|h_i(Z^{k+1}, z_i)\|_2^2 \Big), 0 \Big\} \quad \forall i = 1, 2, 3 \ , \tag{34c}$$

$$Y_j^{k+1} = Y_j^k + \rho(X_j^{k+1} - Z^{k+1}) \quad \forall j = 1, \dots, N \ , \tag{34d}$$

$$y_i^{k+1} = y_i^k + \rho(h_{i-N}(Z^{k+1}, z_i^{k+1})) \quad \forall i = N+1, N+2, N+3 \ . \tag{34e}$$

Thus, the steps for updating X_j and Y_j can be executed locally on each node, while Z performs the role of the central collector for the optimization problems solved locally. Summing up, the ADMM formulation of the problem expressed in this section, compared to the solution proposed in the previous section, is characterized by an increase in the computation performed locally by the nodes. Some steps of the algorithm still need to be executed in a centralized fashion (i.e., Z, z_i, y_i), while others could be executed locally (i.e., X_j, Y_j).

D. Distributed ADMM using Global Consensus, scaled form

Similarly to what has been done in Section V-B, substituting $R_j = X_j - Z$, $r_i = h_i(Z, z_i)$ and $u_i = \frac{1}{\rho} y_i$, $U_j = \frac{1}{\rho} Y_j$, the augmented Lagrangian can be rewritten as:

$$\begin{aligned} &L_\rho(X_1, \dots, X_j, \dots, X_N, Z, z_1, z_2, z_3, U_1, \dots, U_N, u_{N+1}, u_{N+2}, u_{N+3}) \\ &= \sum_{j=1}^{N} \Big(f_j(X_j) + \frac{\rho}{2}\|R_j + U_j\|_2^2 - \frac{\rho}{2}\|U_j\|_2^2 \Big) + \sum_{i=1}^{3} \Big(\frac{\rho}{2}\|r_i + u_{i+N}\|_2^2 - \frac{\rho}{2}\|u_{i+N}\|_2^2 + I_+(z_i) \Big) \ . \end{aligned} \tag{35}$$

Then, the scaled ADMM for the global consensus problem becomes:

$$X_j^{k+1} = \underset{X_j}{\operatorname{argmin}} \left(f_j(X_j) + \frac{\rho}{2} \|X_j - Z^k + U_j^k\|_2^2 \right) \quad \forall j = 1, \ldots, N \ , \tag{36a}$$

$$Z^{k+1} = \underset{Z}{\operatorname{argmin}} \left(\sum_{j=1}^{N} \left(\frac{\rho}{2} \|X_j^{k+1} - Z + U_j^k\|_2^2 \right) + \sum_{i=1}^{3} \left(\frac{\rho}{2} \|h_i(Z, z_i^k) + u_{i+N}^k\|_2^2 \right) \right) \ , \tag{36b}$$

$$z_i^{k+1} = max\Big\{ \underset{z_i}{\operatorname{argmin}} \left(\frac{\rho}{2} \|h_i(Z^{k+1}, z_i) + u_{i+N}^k\|_2^2 \right), 0 \Big\} \quad \forall i = 1, 2, 3 \ , \tag{36c}$$

$$U_j^{k+1} = U_j^k + X_j^{k+1} - Z^{k+1} \quad \forall j = 1, \ldots, N \ , \tag{36d}$$

$$u_i^{k+1} = u_i^k + h_{i-N}(Z^{k+1}, z_i^{k+1}) \quad \forall i = N+1, N+2, N+3 \ . \tag{36e}$$

Note: the fairness and distributed capabilities of this approach (problem formulated as global consensus and solved with ADMM) relies on the fact that each node can take its decision autonomously without a central decision system. Conversely, this approach compared to Section V-B has incremented the number of the constraints: from the initial 3 up to $N + 3$ with the distributed (global consensus) approach.

E. A More Distributed ADMM using Global Consensus

Observing the global consensus problem in Equation (31) and the ADMM steps in Equations (34), it is worth noting that each y_i depends on the constraint h_i, which in turn depends on the global variable Z. This dependency does not allow to execute the step to update y_i in a distributed fashion. It is thus possible to define an equivalent problem considering the X_j local on each node even for which concern the constraints h_i. Using X_j even for the constraints it is thus possible to increase the degree of distribution of the ADMM, applying it to the following equivalent optimization problem:

$$minimize \quad \sum_{j=1}^{N} f_j(X_j) + \sum_{i=1}^{3} I_+(z_i) \ , \tag{37}$$

subject to:

$$X_j - Z = 0 \quad \forall j = 1, \ldots, N \ , \tag{38a}$$

$$h_i(X_j, z_i) = 0 \quad \forall i = 1, 2, 3 \wedge j = 1, \ldots, N \ . \tag{38b}$$

Thus the original constraints are distributed on each node becoming a total of $3 \cdot N$. In this way one obtains the augmented Lagrangian:

$$\begin{aligned} &L_\rho(X_1, \ldots, X_j, \ldots, X_N, Z, z_1, z_2, z_3, Y_1, \ldots, Y_j, \ldots, Y_N, y_{11}, \ldots, y_{ij}, \ldots) \\ &= \sum_{j=1}^{N} \left(f_j(X_j) + Y_j(X_j - Z) + \frac{\rho}{2} \|X_j - Z\|_2^2 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{3} \left(y_{ij} h_i(X_j, Z_i) + \frac{\rho}{2} \|h_i(X_j, z_i)\|_2^2 + I_+(z_i) \right) \right) \\ &= \sum_{j=1}^{N} \left(f_j(X_j) + Y_j(X_j - Z) + \frac{\rho}{2} \|X_j - Z\|_2^2 + \sum_{i=1}^{3} \left(y_{ij} h_i(X_j, Z_i) + \frac{\rho}{2} \|h_i(X_j, z_i)\|_2^2 + I_+(z_i) \right) \right) \ , \end{aligned} \tag{39}$$

and then, the ADMM steps (where, through the y_{ij} each node j can manage its own subset i of constraints):

$$X_j^{k+1} = \underset{X_j}{\operatorname{argmin}} \left(f_j(X_j) + Y_j^k (X_j - Z^k) + \frac{\rho}{2} \|X_j - Z^k\|_2^2 + \sum_{i=1}^{3} \left(y_{ij}^k h_i(X_j, Z_i^k) + \frac{\rho}{2} \|h_i(X_j, Z_i^k)\|_2^2 \right) \right) \quad \forall j = 1, \ldots, N \ , \tag{40a}$$

$$Z^{k+1} = \underset{Z}{\operatorname{argmin}} \left(\sum_{j=1}^{N} \left(- Y_j^k Z + \frac{\rho}{2} \|X_j^{k+1} - Z\|_2^2 \right) \right) \ , \tag{40b}$$

$$z_i^{k+1} = max \left\{ \underset{z_i}{\operatorname{argmin}} \left(\sum_{j=1}^{N} \left(y_{ij}^k h_i(X_j^{k+1}, z_i) + \frac{\rho}{2} \|h_i(X_j^{k+1}, z_i)\|_2^2 \right), 0 \right\} \quad \forall i = 1, 2, 3 \ , \tag{40c}$$

$$Y_j^{k+1} = Y_j^k + \rho (X_j^{k+1} - Z^{k+1}) \quad \forall j = 1, \ldots, N \ , \tag{40d}$$

$$y_{ij}^{k+1} = y_{ij}^k + \rho \Big(h_i(X_j^{k+1}, z_i^{k+1}) \Big) \quad \forall ij = \{i = 1, 2, 3\} \times \{j = 1, \ldots, N\} = \{i \times j\} \ . \tag{40e}$$

This new version of the distributed ADMM algorithm has more steps that can be executed in parallel (i.e., the updates of the y_{ij}) than the approach presented in Section V-C.

F. About distributing the optimization problem with ADMM

Despite the ADMM algorithm does not allow to obtain a fully distributed approach, due to the presence of the *central collector*, great part of the problem is solved locally on the nodes, while the centralized effort is significantly reduced. In Table I we report the steps and the parallelism achievable with the three formulations of the ADMM algorithm. The steps identify the *barriers* (synchronization points) required by the parallelization, namely, the points in time where one variable depends on one or more variables computed in previous steps.

TABLE I
PARALLELISM CAPABILITIES OF THE ADMM FORMULATIONS

Step	ADMM centralized		ADMM distributed			ADMM more distributed		
	Variable	Parallelism	Variable	Parallelism		Variable	Parallelism	
I	X	1	X_j	N		X_j	N	
II	z_i	3	Z	1		Z	1	4
						z_i	3	
III	y_i	3	z_i	3	$3 + N$	Y_j	N	$N + 3 \cdot N$
			Y_j	N		y_{ij}	$3 \cdot N$	
IV			y_i	3				

In the two distributed ADMM approaches described above (see Sections V-C and V-E), for each iteration of the algorithm it is required to exchange a total of $2N$ messages among the nodes: N to distribute the local computation

TABLE II
NODE ATTRIBUTES

CPU frequency (GHz)	CPU (cores)	RAM (GBs)
$1-3$	$1-4$	$0.5-8$

of X_j and Y_j to the central collector (*distribution phase*), and N to receive from the central collector the computation of Z^k (*gathering phase*).

It is worth noting that it could be possible to further parallelize the algorithm in Section V-E acting on z_i and introducing separate variables z_{ij} for each node, similarly to what has been done for y_{ij} in Section V-E.

VI. IMPLEMENTATION AND NUMERICAL RESULTS

The methods have been implemented in *Matlab* with the *CVX toolbox* for the definition of the optimization problems and using the *Gurobi solver* which allows one to use integer variables [*]. This section presents the evaluated scenario (Section VI-A) and some numerical results (Section VI-B) comparing the two different versions of the ADMM (presented in Sections V-B and V-D) with the initial IP formulation (see the beginning of Section V). The IP formulation provides the global maximum, while, as shown in this section, the ADMM formulations provide good suboptimal solutions.

A. Scenario

We focus on two main aspects: the network and the workload models. As a matter of fact each simulation run has two stages: one for generating the network configuration (cloud participants), and another one for evaluating the actual activity period.

The nodes characteristics are showed in Table II, where values are uniformly distributed within the specified intervals. We consider the communication overhead for transferring data as negligible. Since our problem formulation assumes a task distribution through allocation periods, the dynamism in the online presence of the volunteer nodes can be managed similarly. Evaluating allocation periods with a reasonable short duration, the optimization problem, in a first approximation, could consider only the nodes actually online at the beginning of such period. If a node correctly leaves the network (informing the other participants of its action), the tasks in its queue could be evaluated for re-execution in the subsequent allocation period. While, if a node abruptly leaves the network, all the tasks in its execution queue are lost.

For the workload characterization, our main reference is the Google Cloud Backend described in [38]. There, tasks are characterized according to their duration, CPU and memory requirements, each abstracted as either *small* (s) or *large* (l). For confidentiality reasons, the Google Cluster dataset [39] provides obfuscated information about

[*]For the convenience of the reviewers the implemented *Matlab* code is available at https://www.dropbox.com/s/j2r043wu3kmpfzt/optimization_codeSubmitted_150912.zip?dl=0

TABLE III

TASK ATTRIBUTES

size	duration (hours)	CPU (cores)	RAM (GBs)	Deadline offset (percentage on duration)	Poisson mean arrival (ms)
large	$1-12$	$1-4$	$1-4$	0.4	600

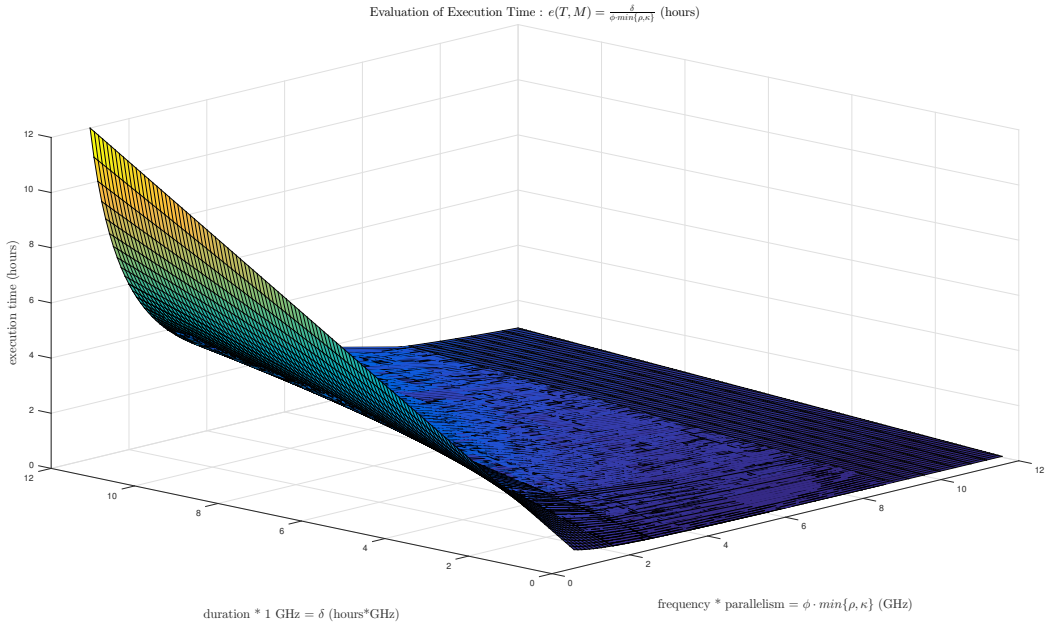


Fig. 2. Required execution time with varying tasks / nodes characteristics

the real hardware characteristics of the Google cluster nodes: every reported value is normalized to the capacity of the best cluster node. Another obfuscated information relevant for the purpose of our work regards the QoS properties such as deadlines. For these reasons we have made some assumptions. We consider that the CPUs of the cluster have a frequency of 1 GHz. In this work we have considered only the tasks of type *large*, whose attributes are uniformly distributed within the intervals (grouped by qualitative coordinates), as reported in Table III.

Evaluating the task execution time as defined in Equation (4), while using the characteristics of nodes and tasks considered during our experiments and defined above, it is possible to plot the time required to execute the tasks. In Figure 2, the z and x axes represent, respectively, the task duration and the execution capability that the node can exploit while the y-axis shows the required execution time. From the plot is possible noting that the task duration is largely affected by the node capabilities. In fact observing Table III the considered tasks have a good degree of parallelism that can be exploited to reduce the required execution time (the blue region in the plot).

B. Results

In this section, we compare the optimal solution value of the original IP optimization problem with values obtained applying the two versions of ADMM discussed in Section V. In particular, the quality of the ADMM solutions has been evaluated. Since the ADMM algorithm, in presence of nonconvex problem, should be considered as a heuristic without any guarantee on convergence, we assume as stopping criteria 1000 iterations of ADMM.

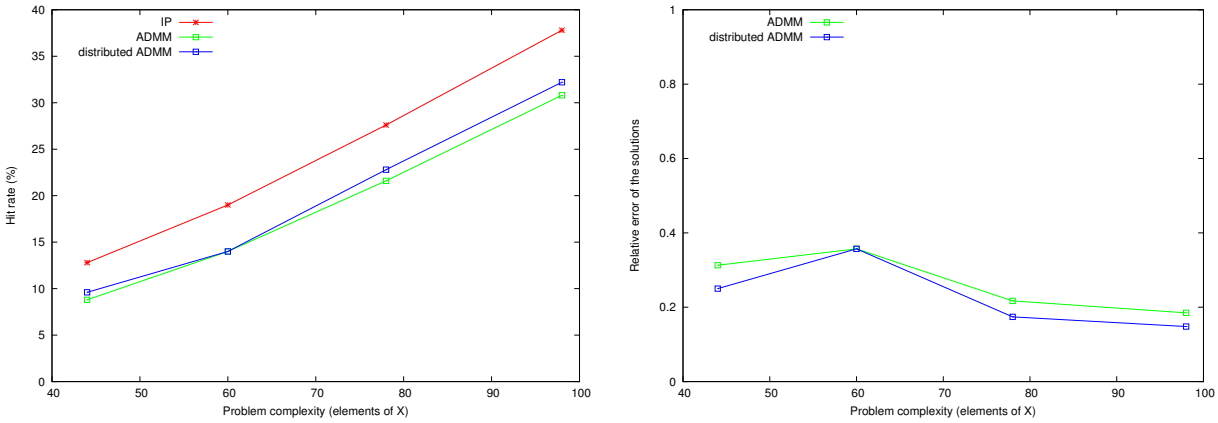
All the experiments have been performed on a machine running Linux 3.16.0-46 64-bit equipped with an Intel Core i5-4460 and 32 GB of RAM, using *Matlab* 2015a with *CVX* 2.1 and the *Gurobi* solver version 6.0.4. A single run, with the largest problem, considering that our Matlab implementation is not able to exploit the parallelism of the algorithm, has required around one hour of execution to perform 1000 iterations of the ADMM algorithm (recalling that no other stopping criteria has been adopted). Increasing the parallelism, the subproblems for the computation of the ADMM variables become easier, and so the computational time for each iteration decreases.

To evaluate the scalability of the approach, we have considered different sizes of the problem (X) varying K (number of tasks) and N (number of nodes). The results shown in this section are the averages of 5 runs, where the random seed changes the characteristics of nodes and tasks at each run according to the previous two Tables. *CVX* currently supports three solvers for problems that make use of integer variables, namely, *Gurobi, MOSEK and GLPK*. Unfortunately, none of the these solvers compatible with *CVX* allow obtaining all the solutions in a straightforward way (it is only possible to apply some tricks to force the solver to skip the solutions already found, e.g., applying cuts that make the solutions already found unfeasible). Thus, comparing the approaches, not all the solutions of X found by the two ADMM implementations are identical. The main metric of interest during our experiments is the *hit rate*, defined as the number of tasks successfully executed (i.e., the value found for the optimization function normalized over K). This is evaluated at the best solution found by each method.

Figure 3 shows the numerical results of our comparisons. Increasing the size of the problem the relative number of tasks that can be executed increase (scaling X, the relative number of K and N is kept constant at about $1:2$). The global maximum (red line) and the best value found by the ADMM formulations for the hit rate are shown in Figure 3a. It is worth noting that, for each size of the problem, the ADMM formulations have found the global maximum in $2/3$ out of the 5 runs. The relative error, defined as the relative difference of the ADMM solutions with the IP one, is reported in Figure 3b.

Despite the ADMM is not always able to find the global optimum, in relation to the goodness of the local maximum and considering its ability to fairly distribute the execution on all the volunteer nodes, it should be considered a compelling approach in a real distributed implementation, when the centralized approach can not be practically applied.

It is worth observing that the size of the optimization variable could be reduced through a simple *pre-filtering* on the nodes' memory. Indeed, the memory is a hard-constraint (Equation (12a)) that must be satisfied by each node independently from the choices of all other nodes and from the other accepted tasks. Thus, it is possible that, for each optimization period, all the nodes perform an initial bid with the set of tasks for which they can satisfy the



(a) (b)

Fig. 3. Numerical results on the ADMM formulations: hit rate (left) and relative error (right).

memory constraint. The original problem can then be subdivided in simpler problems, where the memory constraint can be removed.

VII. The Optimization Framework

In this section, we briefly formulate other task distribution policies which can be modeled according to our framework. The definition of these policies has proved to be straightforward and only minor changes in the optimization function have been required while the ADMM implementation can be basically used *as-is*, changing the $f(X)$ in Section V, as described in this section. The implementation of all the following policies has been necessary to us to obtain an upper bound for the global optimum while evaluating our other simulation-based studies e.g., [10], [13], [16], [40], [41].

A. Minimize the termination time for the task set

In this scenario, we want to minimize the time needed to complete all the tasks in the set. With this approach, the nodes become sooner free, and thus can more easily execute future requests in the subsequent allocation period.

The first step is the construction of the $execEnd_X$ vector as described in Equation (12c), according to the actual choice of the executor nodes. Since the execution end time is zero for the tasks that are not executed, a trivial solution could be constituted by no task executed by any node. This misleading behavior can be eliminated stimulating the execution by adding a penalty factor for the tasks that are not actually executed:

$$penalty := c \times \left(K - \sum_{j=0}^{N-1} executedTasks(j) \right) , \tag{41}$$

where c is the penalty factor for non executing a task. The penalty factor can be varied according to the importance given to the execution of the tasks.

The optimization problem can then be expressed as the minimization of the sum of the total time needed to execute the tasks and the penalty term mentioned above:

$$minimize \sum_{l=0}^{K-1} execEnd_X(l) + penalty \ , \tag{42}$$

subject to the constraints 1, 3 and 4.

In the penalty factor assignment, a more sophisticated strategy could take into account the class s of "importance" for the tasks that are not executed. This point will be considered in a future work.

B. Minimize the response time

The minimization of the response time proceeds similarly to the minimization of the termination time for the task set (Section VII-A). The main difference relies on the choice of the constraints. In this problem, each task is assigned to a node, even if is not possible to satisfy the deadline requirement. This constraint makes useless the inclusion of a penalty factor and thus the problem is formulated as:

$$minimize \sum_{l=0}^{K-1} execEnd_X(l) \ , \tag{43}$$

subject to the constraints 1 and 2.

C. Maximize nodes fairness

In this case, the goal is the maximization of the distribution of tasks among nodes. Thus, the highest number of nodes should participate in the execution of the task set. The objective function is the difference of tasks executed by the node that executes more tasks to the node that executes less:

$$\Delta_{task} = max(executedTasks) - min(executedTasks) \ . \tag{44}$$

This strategy, if implemented trivially brings to a situation where none of the nodes execute tasks since this situation allows to have $\Delta_{task} = 0$. To prevent this vicious behavior from behalf of the nodes, the number of executed tasks is checked and for each not executed task a penalty is assigned (Equation (41)). It is possible to express this problem as a maximization problem: $maximize(node\_fairness(\cdots))$, or dually as a minimization one: $minimize(node\_unfairness(\cdots)$. Our implementation follows the latter approach:

$$minimize \quad \Delta_{task} + penalty \ , \tag{45}$$

subject to the constraints 1, 3 and 4.

D. Maximize nodes greenness

In this case, the goal is the reduction of the number of nodes performing computation in a given allocation period, while taking into account the nodes that are already active and running tasks belonging to previous periods. The underlying idea is that in this way the nodes that do not execute any task can shift towards an *energy-saving* state.

The executed tasks must still respect the deadline constraints, but solutions that use a lower number of nodes are preferred.

This green policy consolidates the tasks in the smallest possible number of nodes but, to prevent that no tasks will be executed, a penalty factor is added (Equation (41)). Maximizing the greenness (i.e., maximizing the number of nodes *off*) is analogous to minimizing the *green cost* (i.e., minimizing the number of nodes *on*). We have followed the latter approach. The computation of the green cost is articulated in a few steps:

$$activeNodes_t = min\{\sum_{l=0}^{K-1} X(l,j), 1_N\} \ , \tag{46a}$$

$$switched = activeNodes_{t-1} - activeNodes_t \ , \tag{46b}$$

$$greenCost = c_2 \times \sum_{j=0}^{N-1} |min\{switched(j), 0_N\}| - b \times \sum_{j=0}^{N-1} max\{switched(j), 0_N\} + penalty \ . \tag{46c}$$

The active nodes are evaluated in Equation (46a), where the min function is used to count a node as active disregarding the number of tasks it is executing (since just one task is enough to require the node to be running). Then, in Equation (46b) the nodes that should change their status in the current allocation period t are evaluated, i.e., each element of the vector $switched \in \mathbb{R}^N$ is 1 if a node can be switched to an energy-saving state, -1 if a new node needs to be activated, and 0 if no change of the status occurs. It follows that the undesired situations occur for the -1 elements of $switched$, while 1 represents the preferred value. Finally, the $greenCost$ is computed in Equation (46c), where the penalty factor c_2 takes into account the nodes that are required to wake-up, while b is a bonus for the nodes switched to energy-saving status.

In fact, the green optimization problem as formulated in Equations (46), turns out to be a multi-criteria optimization problem, where a trade-off among the benefit of executing more tasks (last terms of Equation (46c)) and the use of fewer nodes (the middle term) should be weighted with the need to wake-up further nodes. The optimization problem considers the minimization of the $greenCost$ (Equation (46c)) and it is subject to the constraints 1, 3 and 4 of our framework.

VIII. Conclusions and Future Work

The volunteer cloud computing is characterized by a large-scale heterogeneous and dynamic environment. Such a complex environment makes it hard, if not impossible, to perform global optimization algorithms, and to distribute tasks execution requests (often with associated deadline requirements), in a centralized fashion. Distributed algorithms are thus advocated, as the only solution methods applicable in a real environment.

In this work, we propose a distributed framework defining the optimization problem related to the allocation of the tasks execution requests according to different policies. Our problem formulation has been driven by the requirements of a real environment that we considered in our other previous simulative works [10], [13], [16], [40], [41] based on pure heuristic solutions. At the best of our knowledge, those requirements are instead often neglected

in the literature related to the volunteer cloud, simplifying the domain specification while formalizing the problem as an optimization problem. Example of the key elements of our modeling are: execution queue with FIFO policy, tasks with associated deadline, and evaluation of the actual load on the machines. Without loss of generality in the context of our framework, one of these policies has been implemented relying on the ADMM algorithm, with various parallelism capabilities. The numerical results show that, despite the ADMM is just a heuristic in presence of a non-convex problem (as in our formulation), it can still be used successfully. Up to the authors' knowledge, the application of ADMM for task scheduling in the context of the volunteer cloud is novel.

We conclude by mentioning that, in a real domain, a single policy could be driven by multiple goals. We are investigating how to fruitfully integrate *multi-criteria* optimization techniques [42] in our framework. A further point that can be investigated to increase the realism of our model is related to the time required to transfer the task towards the executor node (recalling that, in the volunteer cloud, each node acts as both task producer and consumer). E.g., evaluating the $execTime$ is possible to build a "local-view" for each node. In each allocation period the tasks could be sorted (preferred) by each node according to the above mentioned transmission cost.

ACKNOWLEDGMENT

This research was partially supported by the EU through the HOME/2013/CIPS/AG/4000005013 project CI2C. The contents of the paper do not necessarily reflect the position or the policy of funding parties.

REFERENCES

[1] V. Cunsolo, S. Distefano, A. Puliafito, and M. Scarpa, "Volunteer computing and desktop cloud: The cloud@home paradigm," in *Network Computing and Applications, 2009. NCA 2009. Eighth IEEE International Symposium on*, July 2009, pp. 134–139.

[2] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, oct.-dec. 2009.

[3] D. P. Anderson, "Boinc: A system for public-resource computing and storage," in *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, ser. GRID '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 4–10. [Online]. Available: http://dx.doi.org/10.1109/GRID.2004.14

[4] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: the Condor experience." *Concurrency - Practice and Experience*, vol. 17, no. 2-4, pp. 323–356, 2005.

[5] F. Brasileiro, E. Araujo, W. Voorsluys, M. Oliveira, and F. Figueiredo, "Bridging the high performance computing gap: The ourgrid experience," in *Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid*, ser. CCGRID '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 817–822. [Online]. Available: http://dx.doi.org/10.1109/CCGRID.2007.28

[6] J. Cappos, I. Beschastnikh, A. Krishnamurthy, and T. Anderson, "Seattle: A platform for educational cloud computing," *SIGCSE Bull.*, vol. 41, no. 1, pp. 111–115, Mar. 2009. [Online]. Available: http://doi.acm.org/10.1145/1539024.1508905

[7] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, "Seti@home: An experiment in public-resource computing," *Commun. ACM*, vol. 45, no. 11, pp. 56–61, Nov. 2002. [Online]. Available: http://doi.acm.org/10.1145/581571.581573

[8] O. Babaoglu, M. Marzolla, and M. Tamburini, "Design and implementation of a p2p cloud system," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12. New York, NY, USA: ACM, 2012, pp. 412–417. [Online]. Available: http://doi.acm.org/10.1145/2245276.2245357

[9] E. Di Nitto, D. J. Dubois, and R. Mirandola, "On exploiting decentralized bio-inspired self-organization algorithms to develop real systems," in *Proceedings of the 2009 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems*, ser. SEAMS '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 68–75. [Online]. Available: http://dx.doi.org/10.1109/SEAMS.2009.5069075

[10] M. Amoretti, A. Lafuente, and S. Sebastio, "A cooperative approach for distributed task execution in autonomic clouds," in *Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on*, Feb 2013, pp. 274–281.

[11] D. Talia, "Cloud computing and software agents: Towards cloud intelligent services," in *Proceedings of the 12th Workshop on Objects and Agents, Rende (CS), Italy, Jul 4-6, 2011*, ser. CEUR Workshop Proceedings, G. Fortino, A. Garro, L. Palopoli, W. Russo, and G. Spezzano, Eds., vol. 741. CEUR-WS.org, 2011, pp. 2–6. [Online]. Available: http://ceur-ws.org/Vol-741/INV02_Talia.pdf

[12] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem," *IEEE Trans. Evolutionary Computation*, vol. 1, no. 1, pp. 53–66, 1997.

[13] S. Sebastio, M. Amoretti, and A. Lluch Lafuente, "A computational field framework for collaborative task execution in volunteer clouds," in *Proceedings of the 9th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, ser. SEAMS 2014. New York, NY, USA: ACM, 2014, pp. 105–114. [Online]. Available: http://doi.acm.org/10.1145/2593929.2593943

[14] J.-T. Tsai, J.-C. Fang, and J.-H. Chou, "Optimized task scheduling and resource allocation on cloud computing environment using improved differential evolution algorithm," *Computers & Operations Research*, vol. 40, no. 12, pp. 3045 – 3055, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030505481300169X

[15] F. Zambonelli and M. Mamei, "Spatial Computing: An Emerging Paradigm for Autonomic Computing and Communication," in *Autonomic Communication*, ser. Lecture Notes in Computer Science, M. Smirnov, Ed., vol. 3457. Springer Berlin Heidelberg, pp. 44–57. [Online]. Available: http://dx.doi.org/10.1007/11520184_4

[16] A. Celestini, A. Lluch Lafuente, P. Mayer, S. Sebastio, and F. Tiezzi, "Reputation-Based Cooperation in the Clouds," in *Trust Management VIII*, ser. IFIP Advances in Information and Communication Technology, J. Zhou, N. Gal-Oz, J. Zhang, and E. Gudes, Eds. Springer Berlin Heidelberg, 2014, vol. 430, pp. 213–220. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-43813-8_15

[17] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[18] I. Gurobi Optimization, "Gurobi optimizer reference manual," 2015. [Online]. Available: http://www.gurobi.com

[19] H. Haridas, S. Kailasam, and J. Dharanipragada, "Cloudy knapsack problems: An optimization model for distributed cloud-assisted systems," in *Peer-to-Peer Computing (P2P), 14-th IEEE International Conference on*, Sept 2014, pp. 1–5.

[20] M. Malawski, K. Figiela, and J. Nabrzyski, "Cost minimization for computational applications on hybrid cloud infrastructures," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1786–1794, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X13000186

[21] R. Fourer, D. M. Gay, and B. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, 2nd ed., S. W. Wallace, Ed. Cengage Learning, 2002.

[22] M. Malawski, K. Figiela, M. Bubak, E. Deelman, and J. Nabrzyski, "Scheduling Multilevel Deadline-Constrained Scientific Workflows on Clouds Based on Cost Optimization," *Scientific Programming*, vol. 2015, pp. 680 271:1–680 271:13, 2015. [Online]. Available: http://dx.doi.org/10.1155/2015/680271

[23] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford, "DONAR: decentralized server selection for cloud services," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. –, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043164.1851211

[24] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *INFOCOM, 2013 Proceedings IEEE*, April 2013, pp. 854–862.

[25] B. Li, S. Song, I. Bezakova, and K. Cameron, "EDR: An energy-aware runtime load distribution system for data-intensive applications in the cloud," in *Cluster Computing (CLUSTER), 2013 IEEE International Conference on*, Sept 2013, pp. 1–8.

[26] A. Nedic, A. Ozdaglar, and P. Parrilo, "Constrained Consensus and Optimization in Multi-Agent Networks," *Automatic Control, IEEE Transactions on*, vol. 55, no. 4, pp. 922–938, April 2010.

[27] Q. Zhu, H. Zeng, W. Zheng, M. D. Natale, and A. Sangiovanni-Vincentelli, "Optimization of Task Allocation and Priority Assignment in Hard Real-time Distributed Systems," *ACM Trans. Embed. Comput. Syst.*, vol. 11, no. 4, pp. 85:1–85:30, Jan. 2013. [Online]. Available: http://doi.acm.org/10.1145/2362336.2362352

[28] IBM, "ILOG CPLEX optimizer."

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[31] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal Parameter Selection for the Alternating Direction Method of Multipliers (ADMM): Quadratic Problems," *Automatic Control, IEEE Transactions on*, vol. 60, no. 3, pp. 644–658, March 2015.

[32] M. J. Feizollahi, M. Costley, S. Ahmed, and S. Grijalva, "Large-scale decentralized unit commitment," *International Journal of Electrical Power & Energy Systems*, vol. 73, no. 0, pp. 97–106, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014206151500191X

[33] O. Miksik, V. Vineet, P. Pérez, and P. Torr, "Distributed Non-convex ADMM-based inference in large-scale random fields," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[34] "The Service Level Agreement," Sep. 2015. [Online]. Available: http://www.sla-zone.co.uk

[35] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[36] R. M. Karp, "Reducibility among Combinatorial Problems," in *Complexity of Computer Computations*, ser. The IBM Research Symposia Series, R. Miller, J. Thatcher, and J. Bohlinger, Eds. Springer US, 1972, pp. 85–103. [Online]. Available: http://dx.doi.org/10.1007/978-1-4684-2001-2_9

[37] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.

[38] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards Characterizing Cloud Backend Workloads: Insights from Google Compute Clusters," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 34–41, 2010.

[39] J. L. Hellerstein, "Google cluster data," Google research blog, Jan. 2010, posted at http://googleresearch.blogspot.com/2010/01/google-cluster-data.html.

[40] S. Sebastio, M. Amoretti, and A. Lluch Lafuente, "AVoCloudy: a simulator of volunteer clouds," *Software: Practice and Experience*, vol. In Press. [Online]. Available: http://dx.doi.org/10.1002/spe.2345

[41] S. Sebastio and A. Scala, "A Workload-Based Approach to Partition the Volunteer Cloud," in *Proceedings of the 1st IEEE International Conference on Collaboration on Internet Computing*, ser. CIC'15. to appear in IEEE Computer Society, 2015.

[42] M. Ehrgott, *Multicriteria Optimization (2. ed.)*. Springer, 2005. [Online]. Available: http://dx.doi.org/10.1007/3-540-27659-9