An Exact Approach for Resource Virtualization and Job Scheduling in Grid Networks

Rosy Aoun and Maurice Gagnaire *

École Nationale Supérieure des Télécommunications - LTCI - UMR 5141 CNRS,
46, rue Barrault F-75634 Paris Cedex 13 - France,
rosy.aoun | maurice.gagnaire@enst.fr

Abstract. Grid environment aims to facilitate networking and computing resources virtualization in carrier's networks. Considering a set of job demands generated at different nodes of the network, the main problem consists in assigning the larger number of jobs to various remote computers with a dual objective. First, jobs must be served in accordance to their scheduling requirements. Second, the inherent networking and computing costs of the jobs must be minimized. In this paper, we propose an original Integer Linear Programming (ILP) formulation of this problem considering three possible marketing strategies for the carrier. The first one corresponds to an auction sale process while the second one aims to maximize the carriers' benefit. We propose a third alternative strategy inspired by previous works carried out within our team for the routing of scheduled traffic demands. The basic idea of our approach consists in considering the time-space correlation of job demands for the cost evaluation of their associated data-transfers and computation. A set of numerical results outline the benefit of our strategy over the two previous ones.

1 Introduction

According to its initial definition, Grid environment aims to satisfy computer applications (or jobs) requiring massive data storage and high performance processing. Nowadays, Grid environment (Grid networking) refers to the concept of resource virtualization. Ideally, one aims to provide a service to the application layer by using a set of distributed resources like storage and computing devices. The problem of assigning computing resources to different types of jobs has attracted considerable attention and has been studied intensively since the beginning of the decade [1]. Many exact and heuristic approaches have been provided in the literature for scheduling demands in Grid networks. In [2], the authors compare the efficiency of different scheduling heuristics according to the constraints imposed by the application layer. These constraints may correspond for instance to the time window during which each job must be processed or

* This work has been granted by the CARRIOCAS French national research program on Grid networking.

to the network resources consumed by each job. In [3], both an exact and an approximate approach are considered for jobs scheduling. A computing and storage resources allocation strategy is also provided. In [4], an ILP formulation is proposed for the assignment of computing resources to job demands and for the routing of their associated Input/Output (I/O) data. The authors in [5] transfer large data files using a two steps approach. The first one is an offline approach that relies on past profiles of data transfers. Two situations may occur: either the files finish earlier than scheduled, or the time given to them is insufficient to transfer all the data, thus the scheduling is than modified in the second step according to the situation.

In the context of Wavelength Division Multiplexing (WDM) optical networks, the concept of Scheduled Demand (SD) has been introduced to deal with preplanned traffic. The routing of SDs in an optical network may be viewed as an intermediate between static network planning and dynamic traffic engineering. Within our team, multiple exact and approximate routing strategies have been proposed for exploiting the time-space correlation among a set of SDs [6]. An appropriate ILP formulation for the routing of SDs is proposed in [7], where the setup and tear-down dates of the demands are known in advance.

In this paper, we assimilate the networking and computing resources proper to a job as a SD. The setup and tear-down dates of this SD correspond to the ideal time window during which the job must be executed. We then propose an ILP formulation that merges the network awareness scheduling presented in [4] and the SD routing strategy. Our ILP formulation considers three possible objective functions corresponding to three different marketing strategies for the carrier.

This paper is structured as follows. The typical Grid network configuration and job characterization are presented in Section 2. In Section 3, we describe the three objective functions corresponding to different marketing strategies for the carriers. The ILP model is presented in Section 4. Numerical results are presented in Section 5. Finally, we give in Section 6 our conclusions and perspectives.

2 Main parameters considered in Grid networking

2.1 Analogy between a job and a scheduled traffic demand

We assume that each network node may generate job demands and is also equipped with computing facilities. We assume that a job is processed by a single remote computer. We neglect propagation delays. Once the remote computing site has been assigned to a job demand, the client node generating the job transfers input data to this site for processing. The output data obtained at the end of the computation are sent back to the client node where they are stored. One assumes an infinite storage capacity at the client nodes. The routing of I/O data of a job may differ according to their bandwidth requirements. Let us consider a set of $\mathcal{D}$ job demands, where each demand $\delta_d \in \mathcal{D}$ is described by its:

- **Activation time γ_d** and **teardown time η_d:** These two parameters indicate the period of time during which the demand needs to be executed. This period includes the I/O data transfers inherent to the execution of the demand. It may be chosen in order to enable a certain time flexibility in the execution of the job [8].
- **Input/Output (I/O) data d_d^i and d_d^o:** These two parameters correspond to the size in bits required for input and output data transfers respectively.
- **CPU power p_d:** This parameter represents the processing capacity in MIPS required by the demand δ_d during its processing period.
- **Submission site s_d:** The site s_d from which the demand accesses the network. The output data generated after the demand has been processed is then stored in the same site s_d.
- **Budget w_d:** This parameter corresponds to the price a client node is ready to pay for the execution of its job.

2.2 Network parameters

We model a Grid network as a set $\mathcal{C}$ of computing resources, where each site c ($c \in \mathcal{C}$) represents a cluster or a group of computers connected together through a high bandwidth network. Each site is also described by its processing capacity $\mathcal{P}_c$ and can satisfy up to $\mathcal{N}_c$ demands at a time. Independently, a node c may generate job demands to be computed on any available computing site in the network. We also model the network as a collection of links $(u, v) \in \mathcal{E}$, where $(u, v) \in \mathcal{C} \times \mathcal{C}$. Each link is made of two contra-directional fibers of maximum capacity $\mathcal{B}_{(u,v)}$ in bit per second.

3 Three possible marketing strategies for Grid networking operators

Considering the computing and networking costs, three possible marketing strategies are possible for a Grid network operator. The buyer (the client node) who is the initiator of the demand, is ready to invest a certain budget for the satisfaction of his demand. The operator in charge of assigning jobs to remote computers is the seller (the carrier). These strategies are summarized in three distinct objective functions f_i ($i \in 1, 2, 3$) introduced in the ILP formulation described in the next section.

- A first objective function f_1 aims to maximize the number of accepted demands independently from their effective cost. This approach typically corresponds to an auction sale of Grid services. Each client node proposes a budget to obtain the satisfaction of its job. By convention, this budget is chosen randomly from zero up to the highest possible cost. In practice, this highest cost corresponds to the highest networking and computing cost combination.

– A second and intuitive objective function f_2 consists in assigning a job to remote computer only if the budget paid by the buyer is greater than the effective cost of the demand. In this context, f_2 tries to maximize the number of selected demands.
– We propose a third and more original objective function f_3 that aims to maximize the number of accepted demands in considering the impact of network and computing resources reutilization on the total cost, that is the cost for provisioning a set of demands. For instance, let us consider a demand with $\boldsymbol{b_d^i}$ = 10 Mbps on a link (u, v) with $\mathcal{B}_{(u,v)}$ = 2.5 Gbps. If this demand is active from 8 am to 9 am and that another demand had been accepted on this same link (u, v) from 4 am to 5 am requiring 100 Mbps of bandwidth, the carrier may accept the current demand whatever the budget of the buyer is, since it has already installed sufficient resources in his network. In other terms, the current demand may be accepted even if its budget is lower than its effective cost.

In the following, costs are expressed in cost units which are currency agnostic.

4 ILP formulation

Let us consider a set $\mathcal{D}$ of job demands δ_i ($i \in 1, 2, \ldots N_g$) generated at different nodes of a network. The network is characterized by a graph $\mathcal{G}[\mathcal{C},\mathcal{E}]$, where $\mathcal{C}$ is the set of nodes and $\mathcal{E}$ is the set of optical links. In this section, we propose an ILP formulation for the assignment of a set of jobs δ_i to the various computing sites $c \in \mathcal{C}$. An intuitive objective consists in enabling the provision of the highest number of jobs at the lowest cost both in terms of consumed network and computing resources. As mentioned in the previous sections, the time periods during which jobs are served must respect a certain scheduler. For each accepted SD, our formulation provides the location of the computing site of the job and the two routes on which are transmitted the input and the output data between the client node and the computing site. The following parameters are considered as inputs for our ILP formulation:

- $C_{(u,v)}$: cost of link between node u and node v
- C_c: cost of computational resource at the computing site c
- t^m: set of discrete time instants, $t^m \in$ T
- t_d^m: a subset of t^m proper to δ_d where $t_d^m = 1$ if and only if $t^m \in [\gamma_d, \eta_d]$
- δ_d: a set of SDs $\delta_d \in \mathcal{D}$ described by $\delta_d = (b_d^i, b_d^o, p_d, s_d, w_d)$ where, b_d^i and b_d^o are the I/O bandwidth of the demands defined by $b_d^{i/o} = \frac{d_d^{i/o}}{\eta_d - \gamma_d}$. Parameters p_d, s_d, w_d have been described previously in section 2.1.

The variables required for the ILP formulation are listed below. We note that all our variables, common to the three objective functions, described in section 3, are binary.

- $x_d = 1$ if and only if δ_d is accepted for execution

- $I/O_d^{(u,v)} = 1$ if and only if δ_d uses link (u,v) for routing I/O data at any time
- $I/O_{d,m}^{(u,v)} = 1$ if and only if δ_d uses link (u,v) for routing its I/O data at time instant t^m
- $S_d^c = 1$ if and only if δ_d uses resource c for storage
- $y_d^c = 1$ if and only if δ_d uses resource c for processing at any time
- $y_{d,m}^c = 1$ if and only if δ_d uses resource c for processing at t^m.

4.1 Constraints

$$\sum_c y_d^c = x_d \quad \forall d \qquad ; \qquad y_{d,m}^c = y_d^c \cdot t_d^m \quad \forall d, \forall m, \forall c \tag{1}$$

In (1), the first constraint forces an accepted demand to use computational power of a single computing node. The second constraint reserves computational resource of a computing node c only during the active period of the demand. Thus, resources are free again after the demand has been processed.

$$I/O_d^{(u,v)} \leq x_d \quad \forall d, \forall (u,v) \quad ; \quad I/O_{d,m}^{(u,v)} = I/O_d^{(u,v)} \cdot t_d^m \quad \forall d, \forall (u,v), \forall m \tag{2}$$

The first equation in (2) allows any job demand δ_d to use multiple fiber links. The second equation in (2) specifies that I/O bandwidth capacity is reserved on each of these fiber links only over the active periods of δ_d.

$$S_d^c = \begin{cases} x_d & \text{if } c = s_d \\ 0 & \text{otherwise} \end{cases} \quad \forall d, \forall c \tag{3}$$

In (3), we state that the storage site of a demand δ_d is its inherent client site.

$$\sum_d y_{d,m}^c \leq N_c \qquad ; \qquad \sum_d (y_{d,m}^c \cdot p_d) \leq P_c \qquad \forall c, \forall m \tag{4}$$

The two constraints in (4) declare that a processor can be time-shared by $\mathcal{N}_c$ demands at a time and that the processing powers of these demands should not exceed the capacity of the processor at that time.

$$\sum_d [b_d^i \cdot I_{d,m}^{(u,v)} + b_d^o \cdot O_{d,m}^{(u,v)}] \leq B_{(u,v)} \quad \forall (u,v), \forall m \tag{5}$$

In (5) we state that, at any time t_m, the bandwidth of the demands using link (u,v) for their input or their output data at that time should not exceed the bandwidth of the link. Note that because of the time limited computing resource reservation of a scheduled demand (or job), a processor may satisfy globally more than $\mathcal{N}_c$ demands over a day thanks to resource reutilization.

$$\sum_v I_d^{(u,v)} - \sum_v I_d^{(v,u)} = S_d^u - y_d^u \quad \forall d, \forall u \tag{6}$$

$$\sum_{v} O_d^{(u,v)} - \sum_{v} O_d^{(v,u)} = y_d^u - S_d^u \quad \forall d, \forall u \tag{7}$$

The flow conservation in the network for input and output data is formulated in (6) and (7), respectively.

4.2 Objective functions

Additional constraints must be added to two of the three cost functions introduced in section 3.

$$\textit{function 1:} \quad \max \sum_{d} (w_d \cdot x_d) \tag{8}$$

We have described in section 3 the principle of auctions that consists for a carrier to serve in priority the customers who are ready to pay up to a certain limit a higher price than other customers for a similar service. In that sense, function 1 aims to optimize the carrier's benefit.

$$\textit{function 2:} \quad \max \sum_{d} (w_d \cdot x_d - \zeta_d) \tag{9}$$

where:

$$\zeta_d = \alpha \cdot \sum_{c} (C_c \cdot p_d \cdot y_d^c) + \beta \cdot \sum_{(u,v)} \cdot C_{(u,v)} (b_d^i \cdot I_d^{(u,v)} + b_d^o \cdot O_d^{(u,v)}) \tag{10}$$

In (10), ζ_d is an additional variable defining the cost of a demand for using network links and CPU resources. The parameters α and β are introduced in order to balance in the cost ζ_d the fraction due to computation and the fraction due to data transfers. Equation (9) rejects a demand if its cost ζ_d it greater than its budget w_d. It maximizes the benefit of the carrier by maximizing the number of selected demands.

$$\textit{function 3:} \quad \max \sum_{d} (w_d \cdot x_d) - \sum_{(u,v)} Z_{(u,v)} - \sum_{c} Z_c \tag{11}$$

In (11), $Z_{(u,v)}$ and Z_c are two additional variables defined by the following two constraints:

$$Z_{(u,v)} \geq \sum_{d} (b_d^i \cdot I_{d,m}^{(u,v)} + b_d^o \cdot O_{d,m}^{(u,v)}) \qquad ; \qquad Z_c \geq \sum_{d} (p_d \cdot y_{d,m}^c) \tag{12}$$

For a given set of job demands $\mathcal{D}$, our ILP formulation is applied to a finite period of time Δ that covers all the time intervals $[\gamma_d, \eta_d]$ associated to each of these demands. Function $Z_{(u,v)}$ evaluates the upper bound over Δ of the bandwidth consumed on the fiber link between nodes u and v. Similarly, function Z_c evaluates the upper bound over Δ of the computing power consumed on node

c. By commodity, we introduce a total cost ϕ that will be used in section 5.3 to comment our numerical results.

$$\phi = \sum_{(u,v)} Z_{(u,v)} + \sum_{c} Z_c \tag{13}$$

5 Numerical applications

5.1 Network topology

The topology considered in our simulations is depicted in Figure 1. It is made of 6 sites corresponding to 3 core nodes and 3 edge nodes. The core nodes are characterized by a CPU capacity 3 times greater than the CPU capacity of the edge nodes. The storage capacity at each of the 6 sites is assumed to be unlimited. Let us remind that each of the 6 sites also corresponds to a client node able to generate job demands. The core nodes are connected to a bi-directional optical fiber ring with a 5 Gbps capacity. Each edge node is connected to a core node by means of a point-to-point bi-directional fiber link with a 2.5 Gbps capacity. It is important to recall that we neglect propagation delays in our ILP formulation.

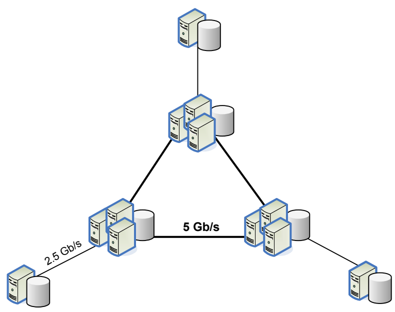


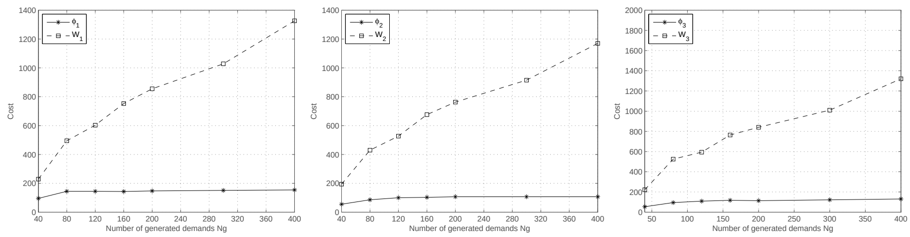
Fig. 1. The adopted Grid network topology

5.2 Traffic scenarios

We assume that for each job, the required input and output bandwidths are identical ($b^i_d = b^o_d$) and equal to 0.5 Gbps. We also assume that each core node and its attached edge node generate job demands during a specific period of time. Thus, each of these three pairs of core and edge nodes generates jobs over a period of 8 successive hours per day. Let us call these three successive time periods P_1 from midnight to 8 am, P_2 from 8 am to 4 pm, and P_3 from 4 pm to midnight. The instant of activation γ_d of a demand generated at a node c is uniformly distributed over the time period P_i associated to the localization of this node. The tear-down time η_d of this demand is randomly distributed over the interval $[\gamma_d, 24 \text{ hours}]$.

5.3 Numerical results

We have underlined in section 2.1 the benefit, in terms of cost, of considering job demands as scheduled traffic demands. Instead of cumulating independently the networking cost of each job, we try to reuse in the best possible way the same network resources for different jobs according to their time-space correlation. Let us introduce the concept of total cost ϕ_i (see equation 13), where i stands for the index of the adopted objective function f_i (see section 3). This total cost represents the summation of two terms referring to the consumed bandwidth and to the consumed CPU. Let N_g and N_a be the total number of generated jobs and the total number of accepted jobs over 24 hours.



(a) (b) (c)

Fig. 2. Global budget and total cost of accepted demands

In Figure 2, we compare in three different sets of curves the global budget $\mathcal{W}_i$ (the total income for the carrier) and the total cost ϕ_i (networking and computing costs) versus N_g for each of the three objective functions f_i. One observes that for the three functions, the benefit of the carrier ($\mathcal{W}_i - \phi_i$) is always positive. Function f_1 aims to maximize the income of the carrier via an auction process in neglecting the effective cost of each demand (Figure 2(a)). Function f_2 aims to maximize the cumulated benefit induced by each demand. In this case, a demand is accepted only if the budget of the buyer is greater than the cost of this demand for the carrier (Figure 2(b)). Function f_3 also aims to maximize the benefit of the carrier. Meanwhile, it does not impose each accepted demand to provide a benefit but simply the combination between all the demands to do so (Figure 2(c)). The rationale behind this rule consists in neglecting the network cost of a demand on a given link if one knows *a priori* that a higher network cost is required by another demand on the same link at a different time period. Numerically, one observes that $\mathcal{W}_i$ is equal to 1326, 1170 and 1320 for $i = 1, 2, 3$ respectively when $N_g = 400$. Similarly, ϕ_i is equal to 154, 107 and 131 for $i = 1, 2, 3$ respectively when $N_g = 400$. In these three cases, $N_a = 186, 153$ and 186 respectively. The budget being proportional to N_a, it is normal that this budget is quite similar for objective functions f_1 and f_3. It is to be noted that f_3 provides a total cost lower than f_1 for the same number of accepted demands thanks to its proper rationale.

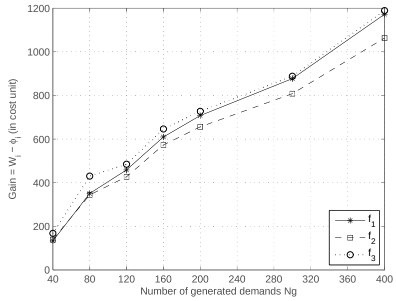
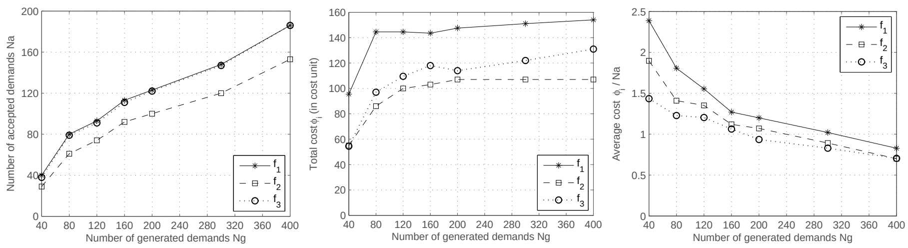


Fig. 3. Gain provided by the three objective functions

In order to outline the differences between the performance of the three objective functions, we have plotted in Figure 3 the gain inherent to each function f_i versus N_g, deducted from the curves plotted in Figure 2. Functions f_1 and f_3 provide a higher gain than function f_2. The reason why f_3 enables a greater gain than f_2 is due to its intrinsic rationale which make it agnostic to the unitary cost of a demand. Thus, a demand δ_d may be accepted under f_3 in spite of having a cost ζ_d higher than its budget w_d. This is possible due to the fact that it can reuse the network and/or the computing resources of other demands.



(a) (b) (c)

Fig. 4. Number of accepted demands (a); Total cost (b); Average cost (c)

Figure 4 illustrates in three successive sets of curves the fluctuations of N_a, ϕ_i and the average cost of an accepted demand $\frac{\phi_i}{N_a}$ versus N_g. It is important to note that, although f_1 and f_3 enable approximately the same number of accepted demands (Figure 4(a)), f_3 provides such a result at a lower total cost than f_1 (Figure 4(b)). Figure 4(b) shows that f_3 induces a higher total cost than f_2; this is simply due to the fact that f_3 accepts a higher number of demands than f_2. The main benefit of objective function f_3 is that it provides the lowest average cost while enabling a benefit comparable to the auction policy (function f_1). Hence, from Figure 4(c) we see that the lowest average cost per accepted demand is provided to us by maximizing f_3.

6 Conclusion and perspectives

In this paper, we have proposed an exact formulation for resource virtualization and job scheduling in Grid networks. For that purpose, we have defined an ILP formulation of the problem enabling to optimize the number of accepted jobs according to three marketing strategies. The simplest strategy is based on the auctions' principle. Another and intuitive strategy consists in maximizing the number of demands that guarantee a benefit for the carrier. We propose a third and original strategy that tries to maximize the number of accepted jobs while taking into account the time-space correlation of their associated I/O data transfers. In order to compare numerically these three marketing strategies, we have considered different traffic scenarios with an increasing number of generated demands. Our numerical results show that our strategy (function f_3) provides a benefit lightly greater than the auctions' principle and clearly superior to the intuitive approach. At the same time, this strategy guarantees the lowest average cost for the provision of a job.

In order to increase the reliability of our results by means of confidence intervals, our coming studies will first consider multiple jobs' generation scenarios for each value of N_g. In a second step, we will propose a heuristic model enabling to consider more realistic scenarios with larger networks and higher number of job demands. In these heuristics, the impact of propagation delays will not be neglected.

References

1. D. de Roure, M. Baker, N. R. Jennings, and N. Shadbolt, "The evolution of the grid," *Berman, F., Fox, G. and Hey, A. J. G., Eds. Grid Computing - Making the Global Infrastructure a Reality*, pp. 65–100, 2003.
2. B. Volckaert, P. Thysebaert, F. De Turck, B. Dhoedt, and P. Demeester, "Application-specific hints in reconfigurable grid scheduling algorithms," *Lecture Notes in Computer Science*, vol. 3038, pp. 149–157, 2004.
3. L. Marchal, P. V.-B. Primet, Y. Robert, and Z. Jingdi, "Optimal bandwidth sharing in grid environments," *2006 15th IEEE International Symposium on High Performance Distributed Computing*, pp. 144–155, June 2006.
4. B. Volckaert, P. Thysebaert, M. De Leenheer, F. De Turck, B. Dhoedt, and P. Demeester, "Network aware scheduling in grids," *NOC2004*, 2004.
5. A. Banerjee, W. Feng, M. Mukherjee, and D. Ghosal, "Routing and scheduling large file transfers over lambda grids," *PFLDNet 2005*, Feb. 2005.
6. M. Gagnaire, M. Koubaa, and N. Puech, "From network planning to traffic engineering in wdm all-optical networks," accepted paper for publication in IEEE Journal on Selected Areas in Communications, last quarter 2007.
7. J. Kuri, N. Puech, M. Gagnaire, E. Dotaro, and R. Douville, "Routing and wavelength assignment of scheduled lightpath demands," *IEEE JSAC*, vol. 21, no. 8, pp. 1231–1240, Oct. 2003.
8. A. Jaekel, "Lightpath scheduling and allocation under a flexible scheduled traffic model," *GLOBECOM 2006, IEEE*, pp. 1–5, Nov. 2006.