Scheduling tasks sharing files on heterogeneous clusters

Arnaud Giersch,
Yves Robert,
Frédéric Vivien

May 2003

Research Report N^o 2003-28

Scheduling tasks sharing files on heterogeneous clusters

Arnaud Giersch, Yves Robert, Frédéric Vivien

May 2003

Abstract

This paper is devoted to scheduling a large collection of independent tasks onto heterogeneous clusters. The tasks depend upon (input) files which initially reside on a master processor. A given file may well be shared by several tasks. The role of the master is to distribute the files to the processors, so that they can execute the tasks. The objective for the master is to select which file to send to which slave, and in which order, so as to minimize the total execution time. The contribution of this paper is twofold. On the theoretical side, we establish complexity results that assess the difficulty of the problem. On the practical side, we design several new heuristics, which are shown to perform as efficiently as the best heuristics in [4, 3] although their cost is an order of magnitude lower.

Keywords: Scheduling, heterogeneous clusters, grid, independent tasks, file-sharing, heuristics.

Résumé

Cet article est consacré à l'ordonnancement d'un grand ensemble de tâches indépendantes sur clusters hétérogènes. Les tâches dépendent de fichiers (d'entrée) qui résident initialement sur un processeur maître. Un fichier donné peut être partagé par plusieurs tâches, Le rôle du maître est de distribuer les fichiers aux processeurs de manière à ce qu'ils puissent exécuter les tâches. L'objectif, pour le maître, est de sélectionner quel fichier envoyer à quel esclave, et dans quel ordre, afin de minimiser le temps d'exécution total. La contribution de cet article est double. D'un point de vue théorique, nous établissons des résultats de complexité évaluant la difficulté du problème. D'un point de vue pratique, nous concevons plusieurs nouvelles heuristiques qui se montrent aussi performantes que les meilleures heuristiques de [4, 3] bien que leur coût soit d'un ordre de grandeur inférieur.

Mots-clés: Ordonnancement, clusters hétérogène, grilles de calcul, tâches indépendantes, partage de fichiers, heuristiques.

1 Introduction

In this paper, we are interested in scheduling independent tasks onto heterogeneous clusters. These independent tasks depend upon files (corresponding to input data, for example), and difficulty arises from the fact that some files may well be shared by several tasks.

This paper is motivated by the work of Casanova et al. [4, 3], who target the scheduling of tasks in APST, the AppLeS Parameter Sweep Template [2]. APST is a grid-based environment whose aim is to facilitate the mapping of application to heterogeneous platforms. Typically, an APST application consists of a *large* number of independent tasks, with possible input data sharing. By *large* we mean that the number of tasks is usually at least one order of magnitude larger than the number of available computing resources. When deploying an APST application, the intuitive idea is to map tasks that depend upon the same files onto the same computational resource, so as to minimize communication requirements. Casanova et al. [4, 3] have considered three heuristics designed for completely independent tasks (no input file sharing) that were proposed in [8]. They have modified these three heuristics (originally called Min-min, Max-min, and Sufferage in [8]) to adapt them to the additional constraint that input files are shared between tasks. As was already pointed out, the number of tasks to schedule is expected to be very large, and special attention should be devoted to keeping the cost of the scheduling heuristics reasonably low.

In this paper, we restrict to the same special case of the scheduling problem as Casanova et al. [4, 3]: we assume the existence of a master processor, which serves as the repository for all files. The role of the master is to distribute the files to the processors, so that they can execute the tasks. The objective for the master is to select which file to send to which slave, and in which order, so as to minimize the total execution time. This master-slave paradigm has a fundamental limitation: communications from the master may well become the true bottleneck of the overall scheduling scheme. Allowing for inter-slave communications, and/or for distributed file repositories, should certainly be the subject of future work. However, we believe that concentrating on the simpler master-slave paradigm is a first but mandatory step towards a full understanding of this challenging scheduling problem.

The contribution of this paper is twofold. On the theoretical side, we establish complexity results that assess the difficulty of the problem. On the practical side, we design several new heuristics, which are shown to perform as efficiently as the best heuristics in [4, 3] although their cost is an order of magnitude lower.

The rest of the paper is organized as follows. The next section (Section 2) is devoted to the precise and formal specification of our scheduling problem, which we denote as TasksSharingFiles. Next, in Section 3, we state complexity results, which include the NP-completeness of the very specific instance of the problem where all files and tasks have the same size. Then, Section 4 deals with the design of low-cost polynomial-time heuristics to solve the TasksSharingFiles problem. We report some experimental data in Section 5. Finally, we state some concluding remarks in Section 6.

2 Framework

In this section, we formally state the optimization problem to be solved.

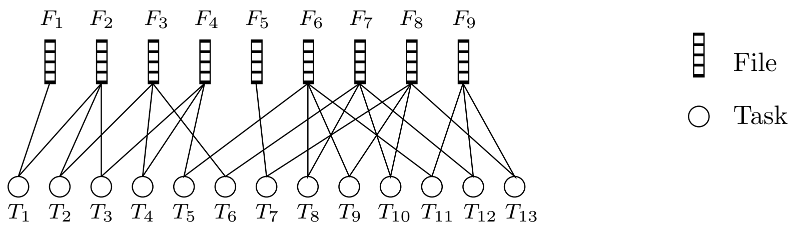

F_1 F_2 F_3 F_4 F_5 F_6 F_7 F_8 F_9
File
Task
T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} T_{11} T_{12} T_{13}

Figure 1: Bipartite graph gathering the relations between the files and the tasks.

2.1 Tasks and files

The problem is to schedule a set of n tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$. Theses tasks have different sizes: the weight of task T_j is t_j, $1 \leq j \leq n$. There are no dependence constraints between the tasks, so they can be viewed as independent.

However, the execution of each task depends upon one or several files, and a given file may be shared by several tasks. Altogether, there are m files in the set $\mathcal{F} = \{F_1, F_2, \ldots, F_m\}$. The size of file F_i is f_i, $1 \leq i \leq m$. We use a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the relations between files and tasks. The set of nodes in the graph $\mathcal{G}$ is $\mathcal{V} = \mathcal{F} \cup \mathcal{T}$, and there is an edge $e_{i,j} : F_i \rightarrow T_j$ in $\mathcal{E}$ if and only if task T_j depends on file F_i. Intuitively, files F_i such that $e_{i,j} \in \mathcal{E}$ correspond to some data that is needed for the execution of T_j to begin. The processor that will have to execute task T_j will need to receive all the files F_i such that $e_{i,j} \in \mathcal{E}$ before it can start the execution of T_j. See Figure 1 for a small example, with $m = 9$ files and $n = 13$ tasks. For instance[1], task T_1 depends upon files F_1 and F_2.

To summarize, the bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node in $V = \mathcal{F} \cup \mathcal{T}$ is weighted by f_i or t_j, and where edges in $\mathcal{E}$ represent the relations between the files and the tasks, gathers all the information on the application.

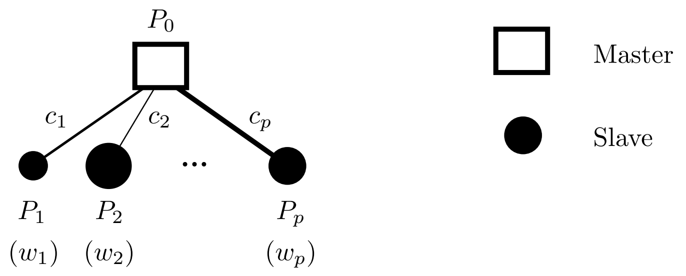


Figure 2: Heterogeneous fork-graph.

2.2 Platform graph

The tasks are scheduled and executed on a master-slave heterogeneous platform. We consider a fork-graph (see Figure 2) with a master-processor P_0 and p slaves P_i, $1 \leq i \leq p$. Each slave P_q has a (relative) computing power w_q: it takes $t_j.w_q$ time-units to execute task T_j on processor P_q. We let $\mathcal{P} = \{P_0, P_1, P_2, \ldots, P_p\}$ denote the platform graph.

The master processor P_0 initially holds all the m files in $\mathcal{F}$. The slaves are responsible for executing the n tasks in $\mathcal{T}$. Before it can execute a task T_j, a slave must have received from the

[1] In this example all tasks depend upon two files exactly. This is because we will re-use the example later in Section 3.2. In the general case, each task depends upon an arbitrary number of files.

master all the files that T_j depends upon. For communications, we use the one-port model: the master can only communicate with a single slave at a given time-step. We let c_q denote the inverse of the bandwidth of the link between P_0 and P_q, so that $f_i.c_q$ time-units are required to send file F_i from the master to slave P_q. We assume that communications can overlap computations on the slaves: a slave can process one task while receiving the files necessary for the execution of another task.

Coming back to the example of Figure 1, assume that we have a two-slave schedule such that tasks T_1 to T_6 are executed by slave P_1, and tasks T_7 to T_{13} are executed by slave P_2. Overall, P_1 will receive six files (F_1 to F_4, F_6 and F_7), and P_2 will receive five files (F_5 to F_9). In this schedule, files F_6 and F_7 must be sent to both slaves.

To summarize, we assume a fully heterogeneous master-slave paradigm: slaves have different speeds and links have different capacities. Communications from the master are serial, and may well become the major bottleneck.

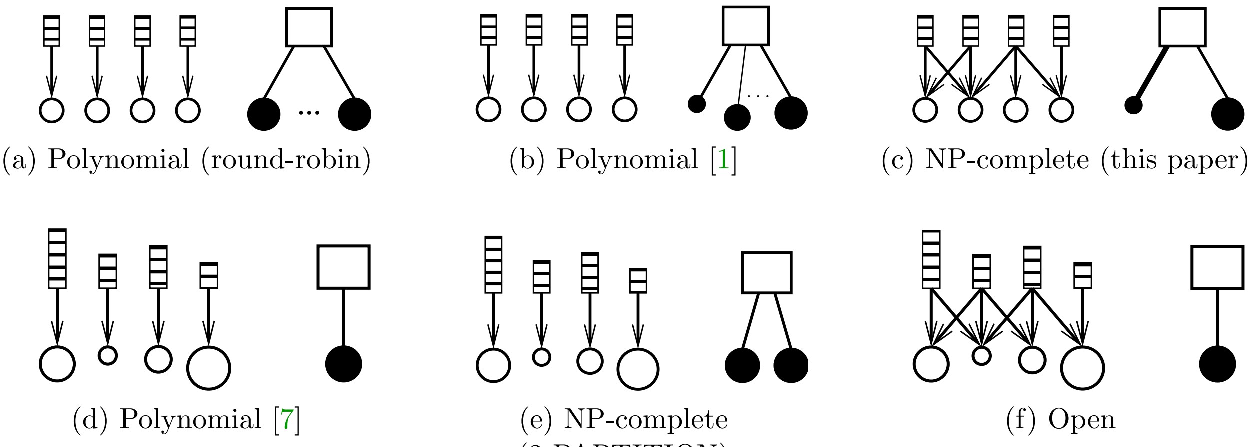
2.3 Objective function

The objective is to minimize the total execution time. The execution is terminated when the last task has been completed. The schedule must decide which tasks will be executed by each slave. It must also decide the ordering in which the master sends the files to the slaves. We stress two important points:

- some files may well be sent several times, so that several slaves can independently process tasks that depend upon these files

- a file sent to some processor remains available for the rest of the schedule. If two tasks depending on the same file are scheduled on the same processor, the file must only be sent once.

We let TasksSharingFiles($\mathcal{G}, \mathcal{P}$) denote the optimization problem to be solved.


(a) Polynomial (round-robin) (b) Polynomial [1] (c) NP-complete (this paper)

(d) Polynomial [7] (e) NP-complete (2-PARTITION) (f) Open

Figure 3: Complexity results for the problem of scheduling tasks sharing files.

3 Complexity

Most scheduling problems are known to be difficult [9, 5]. However, some particular instances of the TasksSharingFiles optimization problem have a polynomial complexity, while the decision problems associated to other instances are NP-complete. We outline several results in this section, which are all gathered in Figure 3. In Figure 3, the pictographs read as follows: for each of the six case studies, the leftmost diagram represents the application graph, and the rightmost diagram represents the platform graph. The application graph is made up of files and tasks which all have the same sizes in situations (a), (b) and (c), while this is not the case in situations (d), (e) and (f). Tasks depend upon a single (private) file in situations (a), (b), (d), (e), which is not the case in situations (c) and (f). As for the platform graph, there is a single slave in situations (d) and (f), and several slaves otherwise. The platform is homogeneous in cases (a) and (e), and heterogeneous in cases (b) and (c). The six situations are discussed in the text below.

3.1 With a single slave

The instance of TasksSharingFiles with a single slave turns out to be more difficult than we would think intuitively. In the very special case where each task depends upon a single non-shared file, i.e. $n = m$ and $\mathcal{E}$ reduces to n edges $e_{i,i} : F_i \rightarrow T_i$, the problem can be solved in polynomial time (this is situation (d) of Figure 3). Indeed, it is equivalent to the two-machine flow-shop problem, and the algorithm of Johnson [7] can be used to compute the optimal execution time. According to Johnson's algorithm we first schedule the tasks whose communication time (the time needed to send the file) is smaller than (or equal to) the execution time in increasing order of the communication time. Then we schedule the remaining tasks in decreasing order of their execution time.

At the time of this writing, we do not know the complexity of the general instance with one slave (situation (f) of Figure 3). Because Johnson's algorithm is quite intricate, we conjecture that the decision problem associated to the general instance, where files are shared between tasks, is NP-complete. We do not even know what the complexity is when files are shared between tasks, but all tasks and files have the same size.

3.2 With two slaves

With several slaves, some problem instances have polynomial complexity. First of all, a greedy round-robin algorithm is optimal in situation (a) of Figure 3: each task depends upon a single non-shared file, all tasks and files have the same size, and the fork platform is homogeneous. If we keep the same hypotheses for the application graph but move to heterogeneous slaves (situation (b) of Figure 3), the problem remains polynomial, but the optimal algorithm becomes complicated: see [1] for a description and proof.

The decision problem associated to the general instance of TasksSharingFiles with two slaves writes as follows:

Definition 1 (TSF2-Dec($\mathcal{G}$,$\mathcal{P}$,$p = 2$,K)). *Given a bipartite application graph $\mathcal{G}$, a heterogeneous platform $\mathcal{P}$ with two slaves ($p = 2$) and a time bound K, is it possible to schedule all tasks within K time-steps?*

Clearly, TSF2-Dec is NP-complete, even if there are no files at all: in that case, TSF2-Dec reduces to the scheduling of independent tasks on a two-processor machine, which itself reduces to the 2-PARTITION problem [6] as the tasks have different sizes. This corresponds to situation (e) in Figure 3, where we do not even need the private files. However, this NP-completeness result does not hold in the strong sense: in a word, the size of the tasks plays a key role in the proof, and there are pseudo-polynomial algorithms to solve TSF2-DEC in the simple case when there are no files (see the pseudo-polynomial algorithm for 2-PARTITION in [6]).

The following theorem states an interesting result: in the case where all files and tasks have unit size (i.e. $f_i = t_j = 1$), the TSF2-DEC remains NP-complete. Note that in that case, the heterogeneity only comes from the computing platform. This corresponds to situation (c) in Figure 3.

Theorem 1. TSF2-DEC*($\mathcal{G},\mathcal{P}$,$p=2$,$f_i = t_j = 1$,K) is NP-complete.

Proof. Obviously, TSF2-DEC($\mathcal{G},\mathcal{P}$,$p=2$,$f_i = t_j = 1$,K) belongs to NP. To prove its completeness, we use a reduction from CLIQUE, which is NP-complete [6]. Consider an arbitrary instance $\mathcal{I}_1$ of CLIQUE: given a graph $G_c = (V_c, E_c)$, and a bound B, is there a clique in G_c (i.e. a fully connected sub-graph) of size B? Without loss of generality, we assume that $|V_c| \geq 9$ and $6 \leq B(B-1) < |E_c|$.

We construct the following instance $\mathcal{I}_2$ of TSF2-DEC($\mathcal{G},\mathcal{P}$,$p = 2$,$f_i = t_j = 1$,K). We let $\mathcal{F} = V_c \cup X$ and $\mathcal{T} = E_c \cup \{T_y\}$ (see Figure 4), which defines $\mathcal{V} = \mathcal{F} \cup \mathcal{T}$. Here, X is a collection of $x = |X|$ additional files, so there is a total of $|V_c| + x$ files, one per node in the original graph G_c and one per new file in X. As for tasks, there are as many tasks as edges in the original graph G_c, plus an additional task T_y.

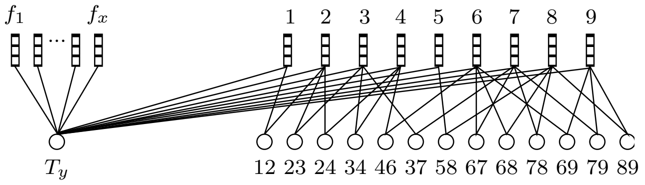


Figure 4: The bipartite application graph used in the proof.

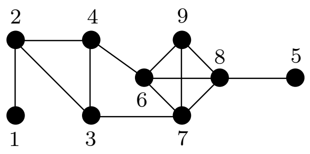


Figure 5: The original graph G_c used to build the bipartite graph.

The relations between tasks and files are defined as follows. First, there is an edge from each file in $\mathcal{F}$ to task T_y; as a consequence, the slave processor that will execute T_y will need to have received all the files in $\mathcal{F}$ from the master before it can begin the execution of T_y. Second, there is an edge from a node (file) $v \in V_c \subset \mathcal{F}$ to a node (task) $e \in E_c \subset \mathcal{T}$ if and only if v was one of the two end-points of edge e in G_c. In the rightmost part of Figure 4, we recognize the bipartite graph of Figure 1, with different labels. In fact, the latter bipartite graph has been obtained from the original graph G_c shown in Figure 5. The files are the nodes in G_c, and the tasks are the edges in G_c. this explains why each task (edge) exactly depends upon two files (the end-points of the edge). We see that G_c has a clique of size $B = 4$ (nodes 6 to 9).

As specified in the problem, all files and tasks have unit size. To complete the description of the application, we let $s = \frac{B(B-1)}{2}$, $r = |E_c| - s$ (note that $s < r$ by hypothesis), and we define $x = (3r-1).|V_c| - 2B + 2$. We check that $x \geq 1$: indeed, $s \geq 3$, $B \geq 3$, $|V_c| \geq B$ and $r \geq 4$; we derive $x \geq 9B + 2$.

There remains to describe the computing platform. The characteristics of the two slave processors are: $w_1 = 3|V_c|$, $w_2 = \frac{3(r+1)|V_c|-4}{s}$, $c_1 = 1$ and $c_2 = 2$. Note that $w_2 > w_1$, because

$w_2 - w_1 = 3(\frac{r}{s} - 1)|V_c| + \frac{3|V_c|-4}{s} > 0$. Thus, $w_2 > 2$. Finally, we define the scheduling bound: $K = 2 + 3(r+1)|V_c|$. Note that $K = 6 + s.w_2$.

Clearly, the instance $\mathcal{I}_2$ can be constructed in time polynomial in the size of $\mathcal{I}_1$. Now we have to show that $\mathcal{I}_2$ admits a solution if and only if $\mathcal{I}_1$ has one.

Assume first that $\mathcal{I}_1$ has a solution, i.e. that G_c possesses a clique of size B. Let $\mathcal{C} = \{v_1, v_2, \ldots, v_B\}$ denote the B vertices in the clique of G_c. The intuitive idea is the following: after sending to slave P_2 the B files corresponding to the B nodes in $\mathcal{C}$, P_2 will be able to process the s tasks that correspond to the edges connecting the nodes of $\mathcal{C}$ without receiving any extra file from the master. The schedule is defined as follows:

- First, at time-steps $t = 0$ and $t = 1$, two files are sent by the master to P_1. These two files are chosen so that they correspond to any two nodes v_a and v_b of V_c that are connected in G_c (i.e. the edge (v_a, v_b) belongs to E_c) and that do not both belong to the clique $\mathcal{C}$: at time-step $t = 2$, P_1 is able to start the execution of the task that corresponds to the edge (v_a, v_b). P_1 terminates this task at time-step $2 + w_1 = 2 + 3|V_c|$.

- Next, the B files that correspond to the clique $\mathcal{C}$ are sent to P_2. As soon as it has received two files, P_2 can start executing one task (the two files correspond to two connected nodes, therefore the task that represents the edge between them is ready for execution). P_2 has received the B files at time-step $2c_1 + Bc_2 = 2 + 2B$, i.e. before it completes the execution of the first task, at time-step $2c_1 + 2c_2 + w_2 = 6 + w_2 > 6 + w_1 = 6 + 3|V_c| > 6 + 3B$, because $B \leq |V_c|$. Therefore, P_2 can process the s tasks corresponding to edges in the clique $\mathcal{C}$ without interruption (i.e. without waiting to receive more files), until time-step $2c_1 + 2c_2 + s.w_2 = 6 + s.w_2 = K$.

- Finally, after sending the B files to P_2, all files but two are sent to P_1: we do not re-send the first two files, but we send all the others, i.e. $|V_c| - 2 + x$ files. We send the $|V_c| - 2$ files corresponding to nodes in V_c before the x files corresponding to nodes in X. When P_1 terminates its first task, at time-step $2 + 3|V_c|$, it has already received the first $|V_c| - 2$ files (the last one is received at time-step $2c_1 + Bc_2 + (|V_c| - 2)c_1 = |V_c| + 2B$). P_1 can process the r tasks corresponding to edges in G_c that do not belong to the clique $\mathcal{C}$ without interruption, until time-step $2c_1 + rw_1 = K - w_1$. At that time-step, P_1 has just received the x last files, because $(|V_c| + x)c_1 + Bc_2 = K - w_1$. P_1 processes then the last task T_y, and the scheduling terminates within K times-steps.

We have derived a valid solution to our scheduling instance $\mathcal{I}_2$.

Assume now that $\mathcal{I}_2$ has a solution. We proceed in several steps:

1. Necessarily, P_1 executes task T_y. Otherwise, P_2 would execute it, but T_y requires $|V_c| + x$ files, and the time needed by P_2 would be at least $(|V_c| + x)c_2 + w_2 = 2(K - w_1 - 2B) + w_2 > 2(K - 5|V_c|) > K$ (because $K \geq 15|V_c|$), a contradiction.

2. P_1 cannot execute more than $\frac{K-2c_1}{w_1} = r + 1$ tasks.

3. All files sent by the master after time-step $K - w_1$ are useless, because the tasks that they might free for execution will not be terminated at time-step K, neither by P_1 nor by P_2 (remember that $w_2 > w_1$). Because P_1 executes T_y, it receives $|V_c| + x$ files. But $K - w_1 = (|V_c| + x)c_1 + Bc_2$, so that P_2 cannot receive more than B tasks from the master.

4. P_2 cannot execute more than s tasks, because $\frac{K-2c_2}{w_2} = \frac{K-6}{w_2} + \frac{2}{w_2} = s + \frac{2}{w_2} < s + 1$.

Overall, a total of $r + s + 1$ tasks are executed. Since P_1 cannot execute more than $r + 1$, and P_2 more than s, they do execute $r + 1$ and s tasks respectively. But P_2 executes s tasks and receives no more than B files: these files define a clique of size B in G_c, thereby providing a solution to $\mathcal{I}_1$. □

4 Heuristics

In this section, we first recall the three heuristics used by Casanova et al. [4, 3]. Next we introduce several new heuristics, whose main characteristic is a lower computational complexity.

4.1 Reference heuristics

Because our work was originally motivated by the paper of Casanova et al. [4, 3], we have to compare our new heuristics to those presented by these authors, which we call *reference heuristics*. We start with a description of these reference heuristics.

Structure of the heuristics. All the reference heuristics are built on the model presented in Figure 6.

```
1: $\mathcal{S} \leftarrow \mathcal{T}$        $\mathcal{S}$ *is the set of the tasks that remain to be scheduled*
2: **while** $\mathcal{S} \neq \emptyset$ **do**
3:   **for** each task $T_j \in \mathcal{S}$ and each processor $P_i$ **do**
4:     Evaluate OBJECTIVE($T_j$, $P_i$)
5:   Pick the "best" couple of a task $T_j \in \mathcal{S}$ and a processor $P_i$ according to OBJECTIVE($T_j$, $P_i$)
6:   Schedule $T_j$ on $P_i$ as soon as possible
7:   Remove $T_j$ from $\mathcal{S}$
```

Figure 6: Structure of the reference heuristics.

Objective function. For all the heuristics, the objective function is the same. OBJECTIVE(T_j, P_i) is indeed the minimum completion time (MCT) of task T_j if mapped on processor P_i. Of course, the computation of this completion time takes into account:

1. the files required by T_j that are already available on P_i (we assume that any file that once was sent to processor P_i is still available and do not need to be resent);
2. the time needed by the master to send the other files to P_i, knowing what communications are already scheduled;
3. the tasks already scheduled on P_i.

Chosen task. The heuristics only differ by the definition of the "best" couple (T_j, P_i). More precisely, they only differ by the definition of the "best" task. Indeed, the "best" task T_j is always mapped on its most favorable processor (denoted $P(T_j)$), i.e. on the processor which minimizes the objective function:

$$\textsc{Objective}(T_j, P(T_j)) = \min_{1 \leq q \leq p} \textsc{Objective}(T_j, P_q).$$

Here is the criterion used for each reference heuristic:

Min-min: the "best" task T_j is the one minimizing the objective function when mapped on its most favorable processor:

$$\textsc{Objective}(T_j, P(T_j)) = \min_{T_k \in \mathcal{S}} \min_{1 \leq l \leq p} \textsc{Objective}(T_k, P_l).$$

Max-min: the "best" task is the one whose objective function, on its most favorable processor, is the largest:

$$\textsc{Objective}(T_j, P(T_j)) = \max_{T_k \in \mathcal{S}} \min_{1 \leq l \leq p} \textsc{Objective}(T_k, P_l).$$

Sufferage: the "best" task is the one which will be the most penalized if not mapped on its most favorable processor but on its second most favorable processor, i.e. the "best" task is the one maximizing:

$$\min_{P_q \neq P(T_j)} \textsc{Objective}(T_j, P_q) - \textsc{Objective}(T_j, P(T_j)).$$

Sufferage II and **Sufferage X:** these are refined version of the **Sufferage** heuristic. The penalty of a task is no more computed using the second most favorable processor but by considering the first processor inducing a significant increase in the completion time. See [4, 3] for details.

Computational complexity. The loop on Step 3 of the reference heuristics computes the objective function for any pair of processor and task. For each processor, this computation has a worst case complexity of $O(|\mathcal{S}| + |\mathcal{E}|)$, where $\mathcal{E}$ is the set of the edges representing the relations between files and tasks (see Section 2.1). Indeed, each remaining task must be considered and in the worst case computing the objective function for a single task has a cost proportional to $|\mathcal{E}|$ (see for example task T_y in Figure 4). Hence, the overall complexity of the heuristics is: $O(p.n^2 + p.n.|\mathcal{E}|)$. The complexity is even worse for **Sufferage II** and **Sufferage X**, as the processors must be sorted for each task, leading to a complexity of $O(p.n^2. \log p + p.n.|\mathcal{E}|)$.

4.2 Structure of the new heuristics

When designing new heuristics, we took special care to decreasing the computational complexity. The reference heuristics are very expensive for large problems. We aimed at designing heuristics which are an order of magnitude faster, while trying to preserve the quality of the scheduling produced.

In order to avoid the loop on all the pairs of processors and tasks of Step 3 of the reference heuristics, we need to be able to pick (more or less) in constant time the next task to be scheduled. Thus we decided to sort the tasks *a priori* according to an objective function. However, since our platform is heterogeneous, the task characteristics may vary from one processor to the other. For example, Johnson's [7] criterion which splits the tasks into two sets (communication time smaller than, or greater than, computation time) depends on the processors characteristics. Therefore, we compute one sorted list of tasks for each processor. Note that this sorted list is computed *a priori* and is not modified during the execution of the heuristic.

Once the sorted lists are computed, we still have to map the tasks to the processors and to schedule them. The tasks are scheduled one-at-a-time. When we want to schedule a new task, on each processor P_i we evaluate the completion time of the first task (according to the sorted list) which has not yet been scheduled. Then we pick the pair task/processor with the lowest completion time. This way, we obtain the structure of heuristics presented in Figure 7.

We still have to define the objective functions used to sort the tasks. This is the object of the next section.

```
 1: for any processor $P_i$ do
 2:   for any task $T_j \in \mathcal{T}$ do
 3:     Evaluate $\textsc{Objective}(T_j, P_i)$
 4:   Build the list $L(P_i)$ of the tasks sorted according to the value of $\textsc{Objective}(T_j, P_i)$
 5: while there remain tasks to schedule do
 6:   for any processor $P_i$ do
 7:     Let $T_j$ be the first unscheduled task in $L(P_i)$
 8:     Evaluate $\textsc{CompletionTime}(T_j, P_i)$
 9:   Pick the couple of a task $T_j$ and a processor $P_i$ minimizing $\textsc{CompletionTime}(T_j, P_i)$
10:   Schedule $T_j$ on $P_i$ as soon as possible
11:   Mark $T_j$ as scheduled
```

Figure 7: Structure of the new heuristics.

4.3 The objective functions

The intuition behind the following four objective functions is quite obvious:

Duration: we just consider the overall execution time of the task as if it was the only task to be scheduled on the platform:

$$\textsc{Objective}(T_j, P_i) = t_j.w_i + \sum_{e_{k,j} \in \mathcal{E}} f_k.c_i.$$

The tasks are sorted by increasing objectives, which mimics the Min-min heuristic.

Payoff: when mapping a task, the time spent by the master to send the required files is payed by all the (waiting) processors as the master can only send files to a single slave at a time, but the whole system gains the completion of the task. Hence, the following objective function encodes the payoff of scheduling the task T_j on the processor P_i:

$$\textsc{Objective}(T_j, P_i) = \frac{t_j}{\sum_{e_{k,j} \in \mathcal{E}} f_k}.$$

The tasks are sorted by decreasing payoffs. Furthermore, the order of the tasks does not depend on the processor, so only one sorted list is required with this objective function. Note that the actual objective function to compute the payoff of scheduling task T_j on processor P_i would be: $\textsc{Objective}(T_j, P_i) = \frac{t_j.w_i}{\sum_{e_{k,j} \in \mathcal{E}} f_k.c_i}$; as the factors w_i and w_i do not change the relative *order* of the tasks on a given processor, we just dropped these factors.

Advance: to keep a processor busy, we need to send it all the files required by the next task that it will process, before it ends the execution of the current task. Hence the execution of the current task must be larger than the time required to send the files. We tried to encode this requirement by considering the difference of the computation- and communication-time of a task. Hence the objective function:

$$\textsc{Objective}(T_j, P_i) = t_j.w_i - \sum_{e_{k,j} \in \mathcal{E}} f_k.c_i.$$

The tasks are sorted by decreasing objectives.

Johnson: we sort the tasks on each processor as Johnson does for a two-machine flow shop (see Section 3.1).

Communication: as the communications may be a bottleneck we consider the overall time needed to send the files a task depends upon as if it was the only task to be scheduled on the platform:

$$\textsc{Objective}(T_j, P_i) = \sum_{e_{k,j} \in \mathcal{E}} f_k.$$

The tasks are sorted by increasing objectives, like for *Duration*. As for *Payoff*, the sorted list is processor independent, and only one sorted list is required with this objective function. This simple objective function is inspired by the work in [1] on the scheduling of homogeneous tasks on an heterogeneous platform.

Computation: symmetrically, we consider the execution time of a task as if it was not depending on any file:

$$\textsc{Objective}(T_j, P_i) = t_j.$$

The tasks are sorted by increasing objectives, like for *Duration*. As for *Payoff*, the sorted list is processor independent, and only one sorted list is required with this objective function. This simple objective function is the counterpart (for computations) of the previous one (for communications).

4.4 Additional policies

In the definition of the previous objective functions, we did not take into account the fact that the files are potentially shared between the tasks. Some of them will probably be already available on the processor where the task is to be scheduled, at the time-step we would try to schedule it. Therefore, on top of the previous objective functions, we add the following additional policies. The goal is (to try) to take file sharing into account.

Shared: In the evaluation of the communication times performed for the objective functions, we replace the sum of the file sizes by the weighted sum of the file sizes divided by the number of tasks depending on these files. In this way, we obtain new objective functions which have the same name than the previous one plus the tag "shared". Here are the mathematical expressions for the first three objectives:

Duration-shared:

$$\textsc{Objective}(T_j, P_i) = t_j.w_i + \sum_{e_{k,j} \in \mathcal{E}} \frac{f_k}{|\{T_l \mid e_{k,l} \in \mathcal{E}\}|}.c_i.$$

Payoff-shared:

$$\textsc{Objective}(T_j, P_i) = \frac{t_j}{\sum_{e_{k,j} \in \mathcal{E}} \frac{f_k}{|\{T_l \mid e_{k,l} \in \mathcal{E}\}|}}.$$

Advance-shared:

$$\textsc{Objective}(T_j, P_i) = t_j.w_i - \sum_{e_{k,j} \in \mathcal{E}} \frac{f_k}{|\{T_l \mid e_{k,l} \in \mathcal{E}\}|}.c_i.$$

We do not give the mathematical expression for **Johnson-shared**, because it is cumbersome (but not difficult): as for **Johnson**, we have to split the tasks onto two subsets.

Readiness: for a given processor P_i, and at a given time, the "ready" tasks are the ones whose files are already all on P_i. Under the *Readiness* policy, if there is any ready task on processor P_i at Step 7 of the heuristics, we pick one ready task instead of the first unscheduled task in the sorted list $L(P_i)$.

Locality: in order to try to decrease the amount of file replication, we (try to) avoid mapping a task T_j on a processor P_i if some of the files that T_j depends upon are already present on another processor. To implement this policy, we modify Step 7 of the heuristics. Indeed, we no longer consider the first unscheduled task in $L(P_i)$, but the next unscheduled task which does not depend on files present on another processor. If we have scanned the whole list, and if there remains some unscheduled tasks, we restart from the beginning of the list with the original task selection scheme (first unscheduled task in $L(P_i)$).

Finally, we obtain as many as 48 variants, since any combination of the three additional policies may be used for the six base objective functions.

4.5 Computational complexity

Computing the value of an objective function for all tasks on all processors has a cost of $O(p.(n + |\mathcal{E}|))$. So the construction of all the sorted lists has a cost of $O(p.n.\log n + p.|\mathcal{E}|)$, except for **Payoff** and **Payoff-shared** which only require a single sorted list and whose complexity is thus $O(n.\log n + |\mathcal{E}|)$. The execution of the loop at Step 5 of the heuristics (see Figure 7) has an overall cost of $(p.n.|\mathcal{E}|)$. Hence the overall execution time of the heuristics is:

$$O(p.n.(\log n + |\mathcal{E}|))$$

We have replaced the term n^2 in the complexity of the reference heuristics by the term $n \log n$. The experimental results will assert the gain in complexity. Note that all the additional policies can be implemented without increasing the complexity of the base cases.

5 Experimental results

In order to compare our heuristics and the reference heuristics, we have simulated their executions on randomly built platforms and graphs. We have conducted a very large number of experiments, which we summarize in this section.

5.1 Experimental platforms

Processors: we have recorded the computational power of the different computers used in our laboratories (in Lyon and Strasbourg). From this set of values, we randomly pick values whose difference with the mean value was less than the standard deviation. This way we define a realistic and heterogeneous set of **20 processors**.

Communication links: the **20 communication links** between the master and the slave are built along the same principles as the set of processors.

Communication to computation cost ratio: The absolute values of the communication link bandwidths or of the processors speeds have no meaning (in real life they must be pondered by application characteristics). We are only interested by the relative values of the processors speeds, and of the communication links bandwidths. Therefore, we normalize processor and communication characteristics. Also, we arbitrarily impose the communication-to-computation cost ratio, so as to model three main types of problems: computation intensive (ratio=0.1), communication intensive (ratio=10), and intermediate (ratio=1).

5.2 Tasks graphs

We run the heuristics on the following four types of tasks graphs. In each case, the size of the files and tasks are randomly and uniformly taken between 0.5 and 5.

Two-one: each task depends on exactly two files: one file which is shared with some other tasks, and one un-shared file.

Random: each task randomly depends on 1 up to 50 files.

Partitioned: this is a type of graph intermediate between the two previous ones; the graph is divided into 20 chunks of 100 tasks, and on each chunk each task randomly depends on 1 up to 10 files. The whole graph contains at least 20 different connected components.

Forks: each graph contains 100 fork graphs, where each fork graph is made up of 20 tasks depending on a single and same file.

Each of our graphs contains 2000 tasks and 2500 files, except for the fork graphs which also contain 2000 tasks but only 100 files.

In order to avoid any interference between the graph characteristics and the communication-to-computation cost ratio, we normalize the sets of tasks and files so that the sum of the task sizes equals the sum of the file sizes.

5.3 Results

Heuristic	Relative performance	Standard deviation	Relative cost	Standard deviation
Sufferage	1.110	0.1641	376.7	153.4
Min-min	1.130	0.1981	419.2	191.7
Computation+readiness	1.133	0.1097	1.569	0.4249
Computation+shared+readiness	1.133	0.1097	1.569	0.4249
Duration+locality+readiness	1.133	0.1295	1.499	0.4543
Duration+readiness	1.133	0.1299	1.446	0.3672
Payoff+shared+readiness	1.138	0.126	1.496	0.6052
Payoff+readiness	1.139	0.1266	1.246	0.2494
Payoff+shared+locality+readiness	1.145	0.1265	1.567	0.5765
Payoff+locality+readiness	1.145	0.1270	1.318	0.2329

Table 1: Relative performance and cost of the best ten heuristics.

Table 1 summarizes all the experiments. In this table, we report the best ten heuristics, together with their cost. This is a summary of $12,000$ random tests ($1,000$ tests over all four graph types and three communication-to-computation cost ratios). Each test involves 53 heuristics (5 reference heuristics and 48 combinations for our new heuristics). For each test, we compute the ratio of the performance of all heuristics over the best heuristic. The best heuristic differs from test to test, which explains why no heuristic in Table 1 can achieve an average relative performance exactly equal to 1. In other words, the best heuristic is not always the best of each test, but it is closest to the best of each test in the average. The optimal relative performance of 1 would be achieved by picking, for any of the $12,000$ tests, the best heuristic for this particular case.

We see that **Sufferage** gives the best results: in average, it is within 11% of the optimal. The next nine heuristics closely follow: they are within 13% to 14.5% of the optimal. Out of these nine heuristics, only **Min-min** is a reference heuristic. Clearly, the readiness policy has a major impact on the results.

In Table 1, we also report computational costs (CPU time needed by each heuristic). The theoretical analysis is confirmed: our new heuristics are an order of magnitude faster than the reference heuristics.

As a conclusion, given their good performance compared to **Sufferage**, we believe that the eight new variants listed in Table 1 provide a very good alternative to the costly reference heuristics.

We report more detailed performance data in Figures 8 and 9, which represent a dual view of the experiments. We selected 24 representative variants of the 48 heuristics, and we report the totality of the corresponding data in the Appendix. In Figure 8, we report the relative performance of these variants, averaging on all graphs: this is to show the impact of the communication-to-computation cost ratio. In Figure 9, we report the relative performance of the 24 variants, averaging on communication-to-computation cost ratios: this is to show the impact of the graph type. We state from Figure 8 that:

- the **Advance** heuristic and its variants perform badly
- when the communication-to-computation cost ratio is low, many heuristics are not within 30% of the optimal; when it is high, the performance of all heuristics is globally better (most of them are within 15% of the optimal). In the latter case, it is likely that the communications from the master become the true bottleneck of all scheduling strategies.

Similarly, we see from Figure 9 that:

- the **Computation+readiness** outperforms by 20% all the other heuristics, including the reference heuristics, for the graphs of type Forks.
- however, the **Duration+readiness** gives more stable results for all types of graphs.

Overall, **Computation+readiness** and **Duration+readiness** are the recommended heuristics.

6 Conclusion

In this paper, we have dealt with the problem of scheduling a large collection of independent tasks, that may share input files, onto heterogeneous clusters. On the theoretical side, we have shown a new complexity result. On the practical side, we have improved upon the heuristics proposed by Casanova et al. [4, 3]. We have succeeded in designing a collection of new heuristics which have similar performances but whose computational costs are an order of magnitude lower.

This work, as the one of Casanova et al., was limited to the master-slave paradigm. It is intended as a first step towards addressing the challenging situation where

- input files are distributed among several file servers (several masters) rather than being located on a single master,
- communication can take place *between* computational resources (slaves) in addition to the messages sent by the master(s): some slave may well propagate files to another slave while computing.

We hope that the ideas introduced when designing our heuristics will prove useful for this difficult scheduling problem.

References

[1] O. Beaumont, A. Legrand, and Y. Robert. A polynomial-time algorithm for allocating independent tasks on heterogeneous fork-graphs. In *ISCIS XVII, Seventeenth International Symposium On Computer and Information Sciences*, pages 115–119. CRC Press, 2002.

[2] F. Berman. High-performance schedulers. In I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a New Computing Infrastructure*, pages 279–309. Morgan-Kaufmann, 1999.

[3] H. Casanova, A. Legrand, D. Zagorodnov, and F. Berman. Using Simulation to Evaluate Scheduling Heuristics for a Class of Applications in Grid Environments. Research Report 99-46, Laboratoire de l'Informatique du Parallélisme, ENS Lyon, Sept. 1999.

[4] H. Casanova, A. Legrand, D. Zagorodnov, and F. Berman. Heuristics for Scheduling Parameter Sweep Applications in Grid Environments. In *Ninth Heterogeneous Computing Workshop*, pages 349–363. IEEE Computer Society Press, 2000.

[5] P. Chrétienne, E. G. C. Jr., J. K. Lenstra, and Z. Liu, editors. *Scheduling Theory and its Applications*. John Wiley and Sons, 1995.

[6] M. R. Garey and D. S. Johnson. *Computers and Intractability, a Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1991.

[7] S. M. Johnson. Optimal two- and three-stage production schedules with setup times included. *Naval Research Logistics Quarterly*, 1:61–68, 1954.

[8] M. Maheswaran, S. Ali, H. Siegel, D. Hensgen, and R. Freund. Dynamic matching and scheduling of a class of independent tasks onto heterogeneous computing systems. In *Eight Heterogeneous Computing Workshop*, pages 30–44. IEEE Computer Society Press, 1999.

[9] B. A. Shirazi, A. R. Hurson, and K. M. Kavi. *Scheduling and load balancing in parallel and distributed systems*. IEEE Computer Science Press, 1995.

Appendix

In this section, we provide the raw data generated by the experiments for the 24 representative variants that we have selected. In Figures 10 to 13, we plot the actual execution times (makespans) for each graph type, with the three communication-to-computation cost ratios from top to bottom in each figure. Figures 14 to 17 are the counterpart for relative execution times.

Next, we show in Figure 18 the total file volume transferred from the master to the slaves for graphs of types Forks. As expected, the best heuristics are those which minimize the transfer of the same file to several slaves. Finally, in Table 2 we report the relative cost of each heuristic when solving the $12,000$ tests. For each of the tests, the relative cost of a heuristic is the ratio of the CPU time required by the heuristic and the minimum of the CPU times required by the heuristics.

	two-one	random	partitioned	forks	average
Duration	1.048	1.026	1.06	1.942	1.269
Duration+shared	1.057	1.842	1.076	2.092	1.517
Duration+readiness	1.152	1.395	1.244	1.991	1.446
Duration+locality	1.085	1.292	1.129	2.204	1.427
Payoff	1.032	1.182	1.057	1.01	1.07
Payoff+shared	1.04	2.151	1.073	1.018	1.321
Payoff+readiness	1.134	1.545	1.243	1.062	1.246
Payoff+locality	1.085	1.428	1.138	1.346	1.25
Advance	1.056	1.163	1.121	1.957	1.324
Advance+shared	1.064	1.937	1.116	2.11	1.557
Advance+readiness	1.159	1.405	1.281	2.015	1.465
Advance+locality	1.092	1.432	1.186	2.219	1.482
Johnson	1.138	1.357	1.187	2.382	1.516
Johnson+shared	1.148	2.277	1.191	2.16	1.694
Johnson+readiness	1.207	1.861	1.395	2.379	1.71
Johnson+locality	1.157	1.654	1.283	2.474	1.642
Communication	1.01	1.186	1.046	1.639	1.22
Communication+shared	1.021	2.152	1.062	1.647	1.471
Communication+readiness	1.111	1.466	1.232	1.791	1.4
Communication+locality	1.065	1.43	1.129	1.97	1.398
Computation	1.019	1.099	1.035	2.16	1.328
Computation+shared	1.019	1.099	1.035	2.162	1.329
Computation+readiness	1.123	1.763	1.269	2.119	1.569
Computation+locality	1.072	1.389	1.128	2.304	1.473
Min-min	220.3	728.8	388.4	339.2	419.2
Max-min	214	697.7	330.4	329.3	392.8
Sufferage	227.1	596.5	317	366.2	376.7
Sufferage II	476.2	736.9	532.4	974.2	679.9
Sufferage X	483.9	603.8	511.4	990.1	647.3

Table 2: Relative costs of the heuristics.

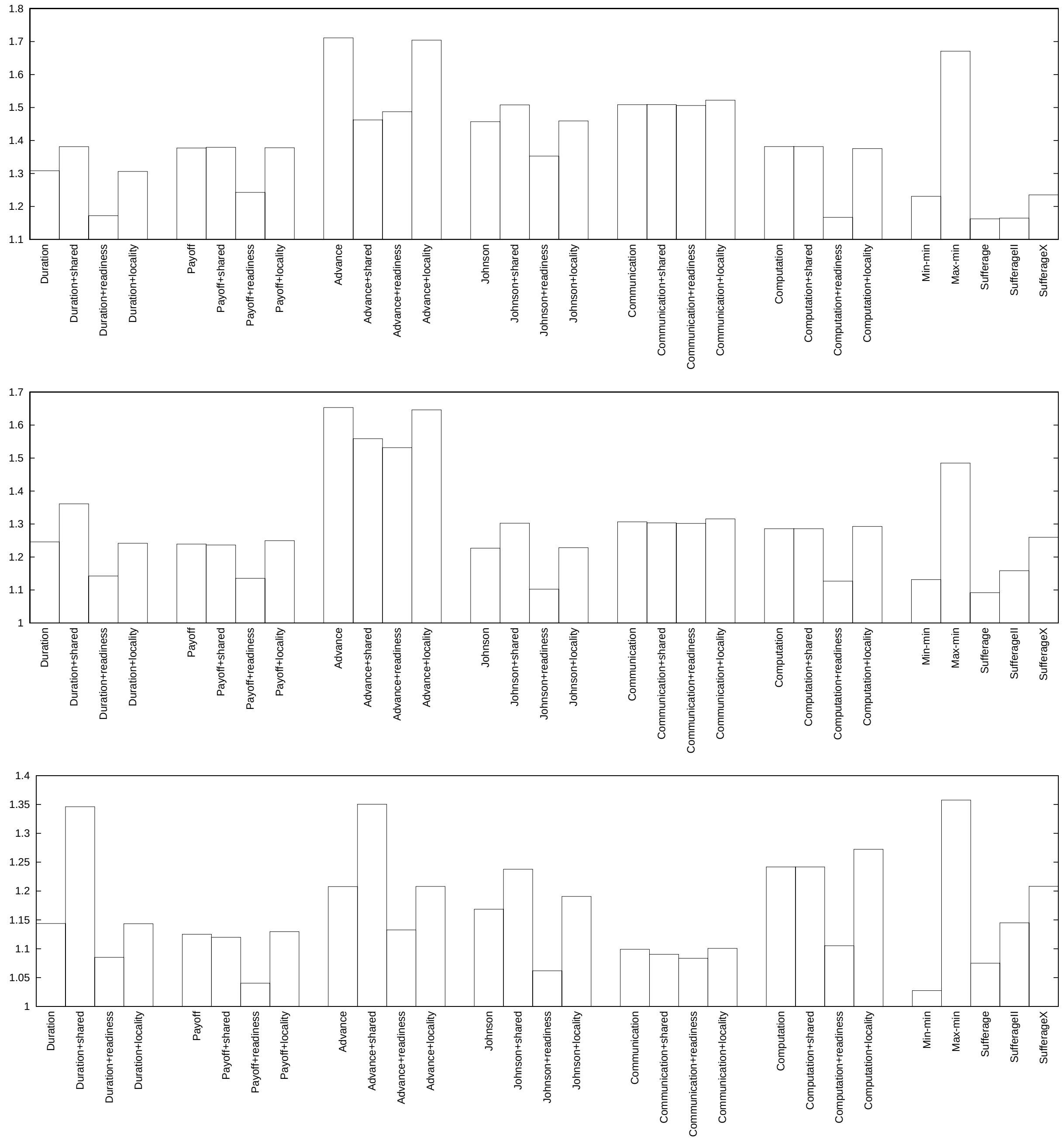


Figure 8: Relative performances of the schedules produced by the different heuristics average on four types of graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

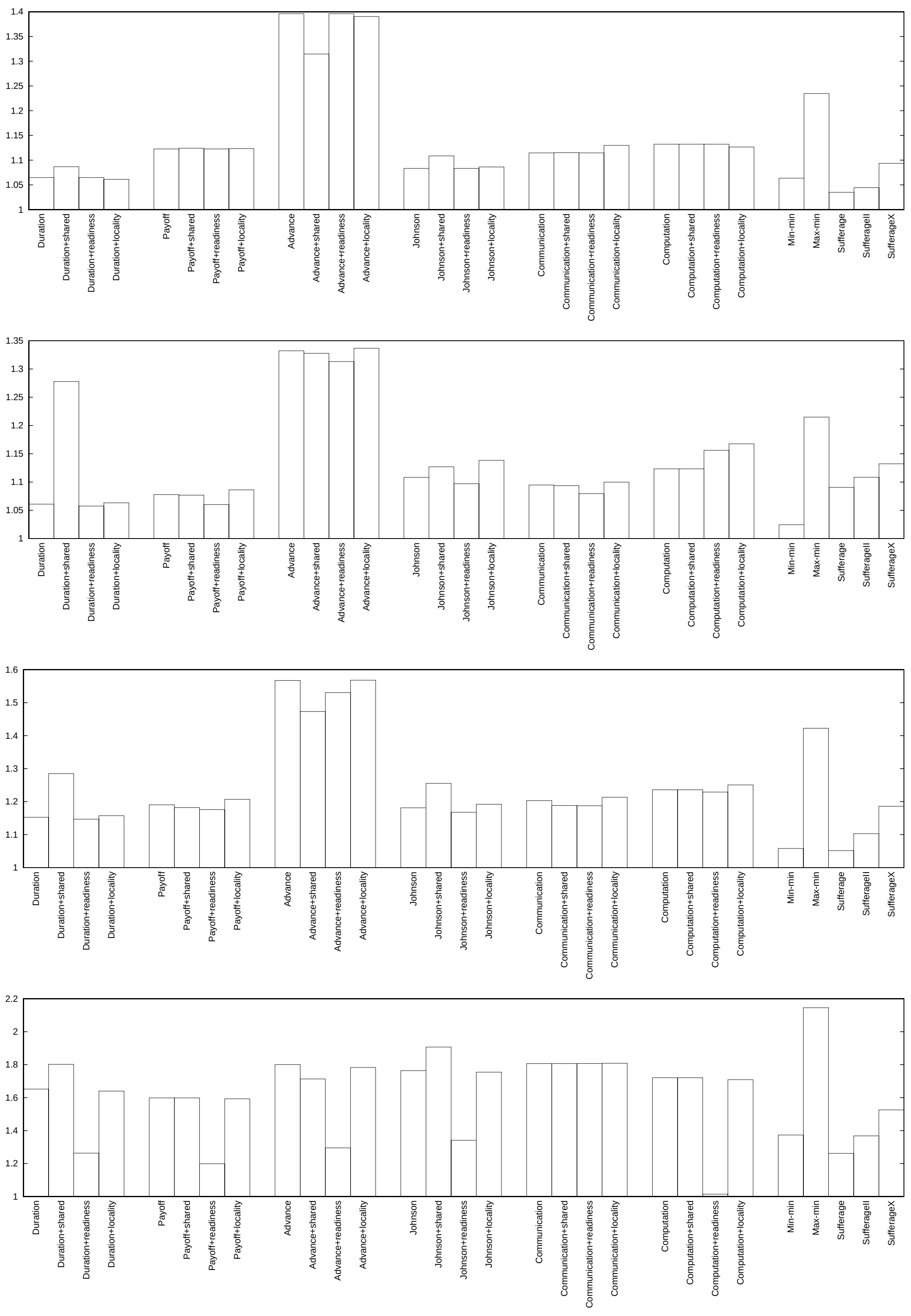


Figure 9: Relative performances of the schedules produced by the different heuristics average on three communication to computation ratios, for the four types of graphs (from top to bottom: *Two-one*, *Random*, *Partitioned*, and *Forks*).

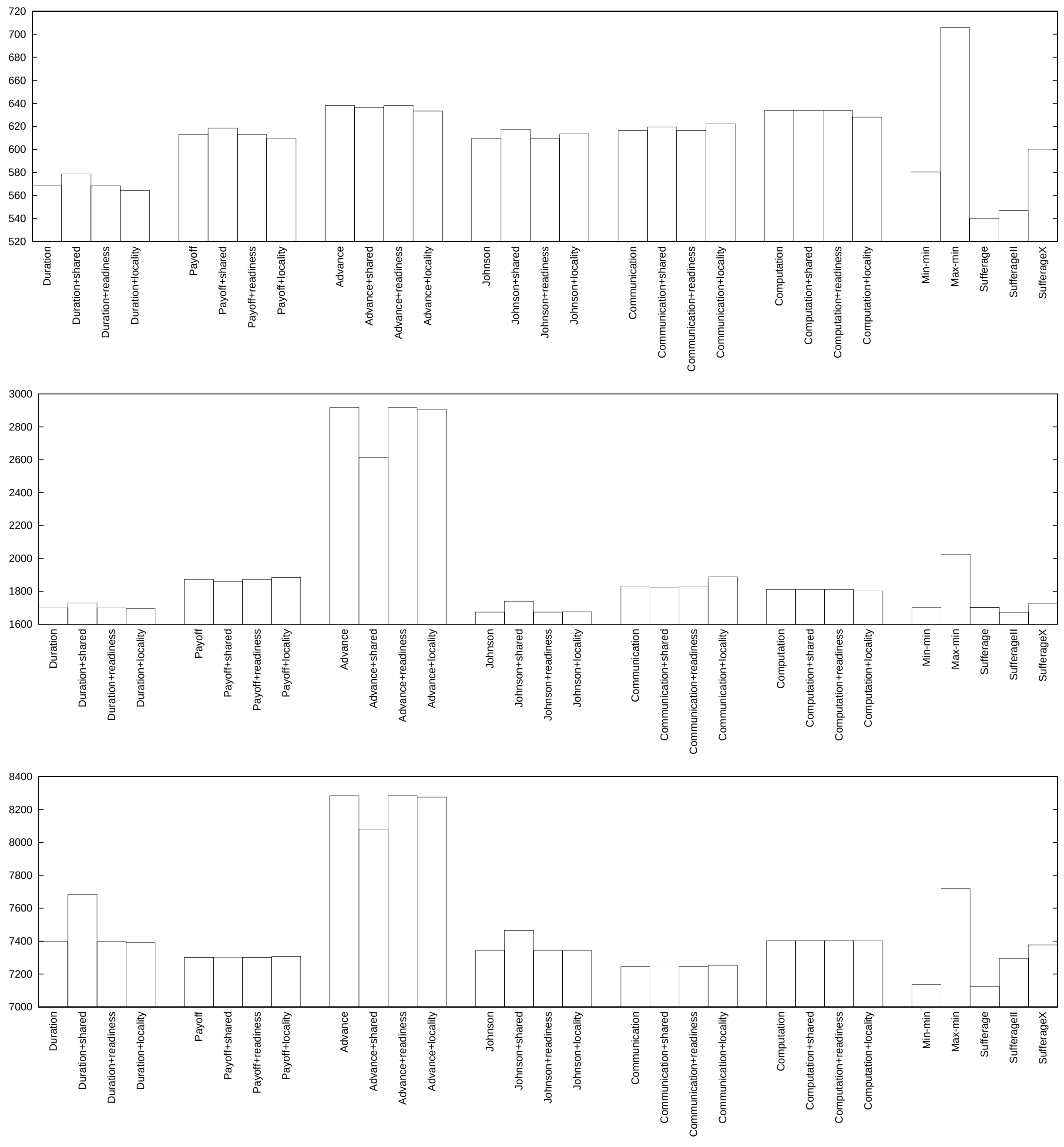


Figure 10: Execution times of the schedules produced by the different heuristics on the *Two-one* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

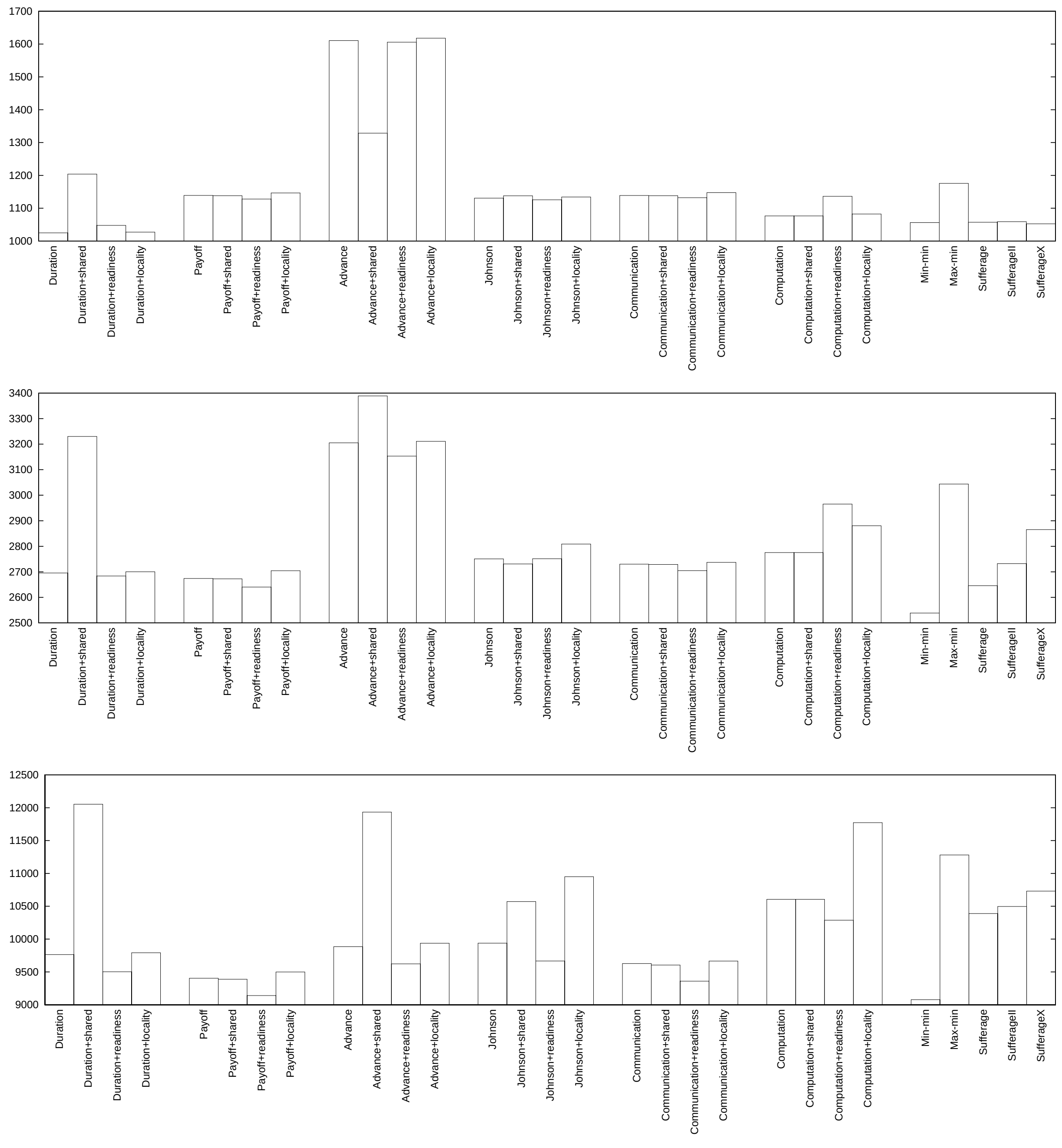


Figure 11: Execution times of the schedules produced by the different heuristics on the *Random* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

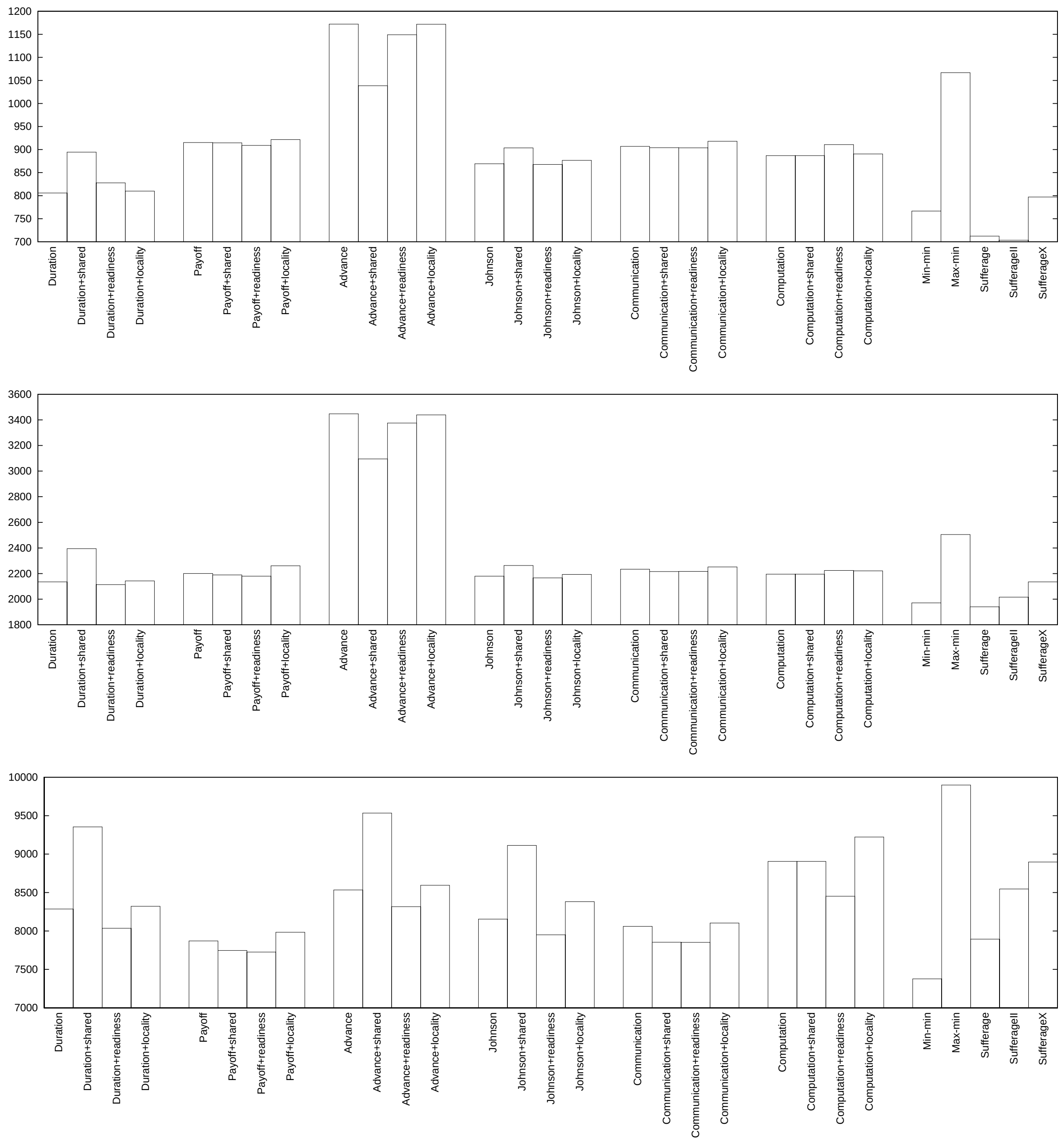


Figure 12: Execution times of the schedules produced by the different heuristics on the *Partitioned* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

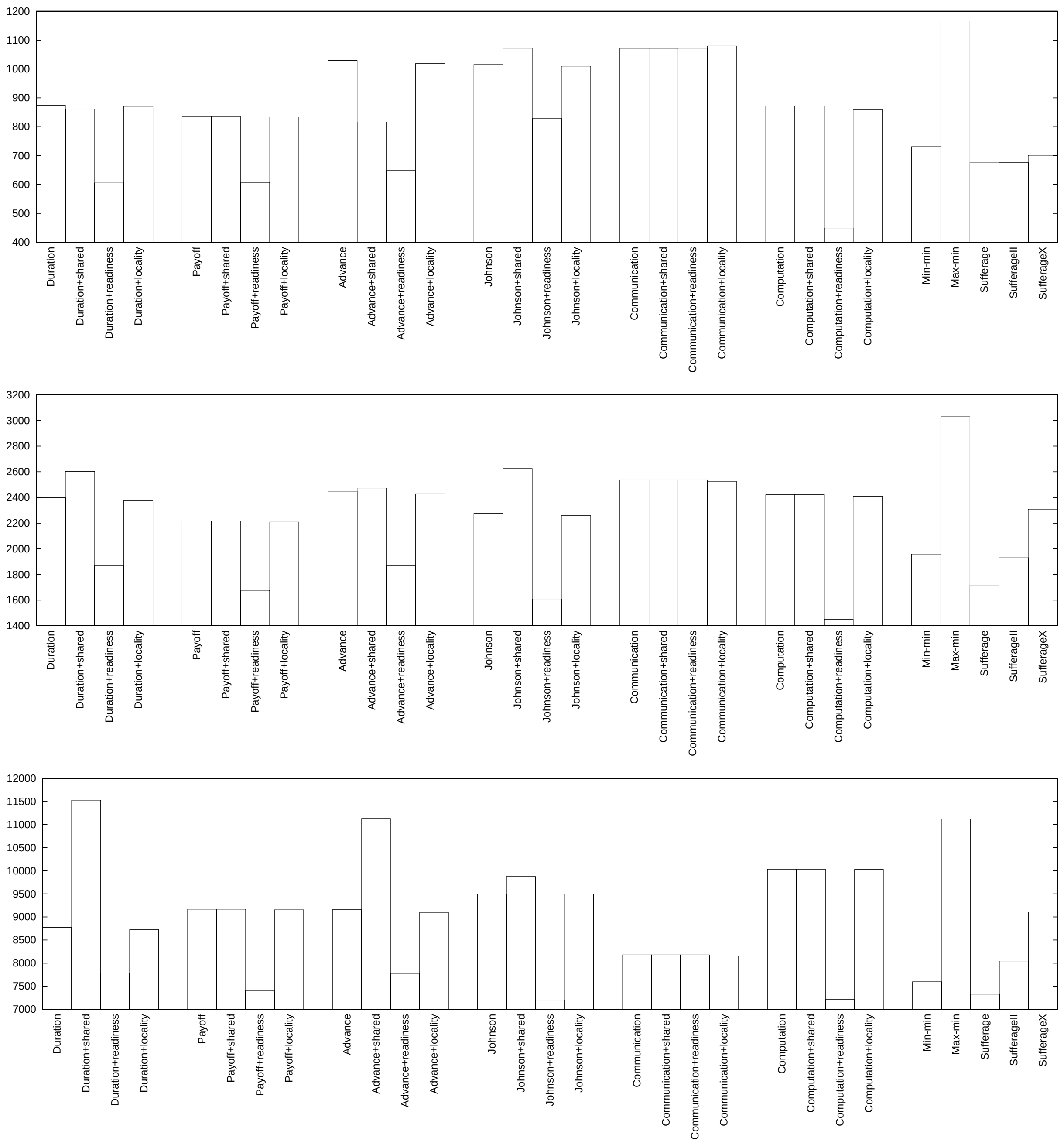


Figure 13: Execution times of the schedules produced by the different heuristics on the *Forks* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

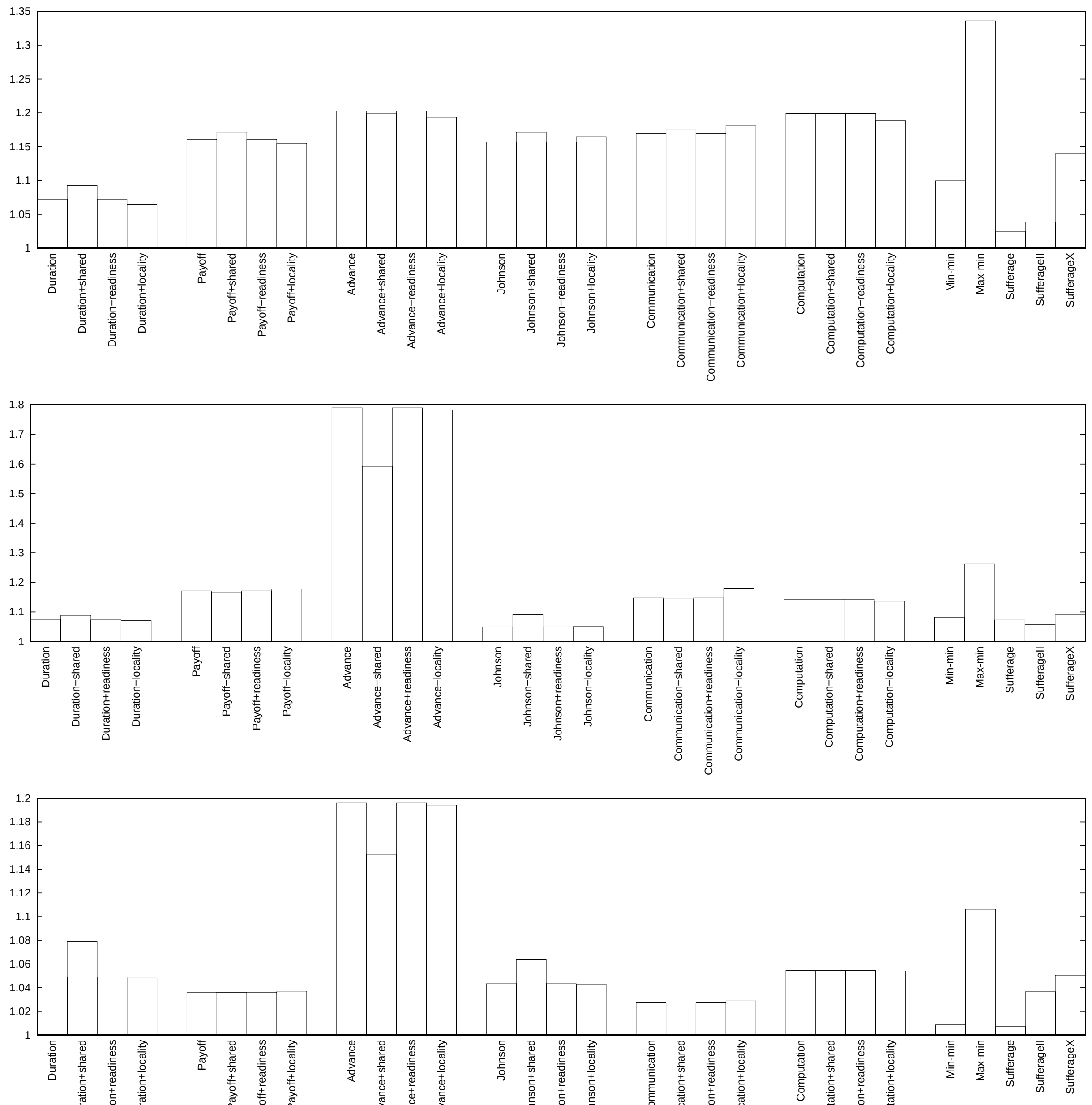


Figure 14: Relative performances of the schedules produced by the different heuristics on the *Two-one* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

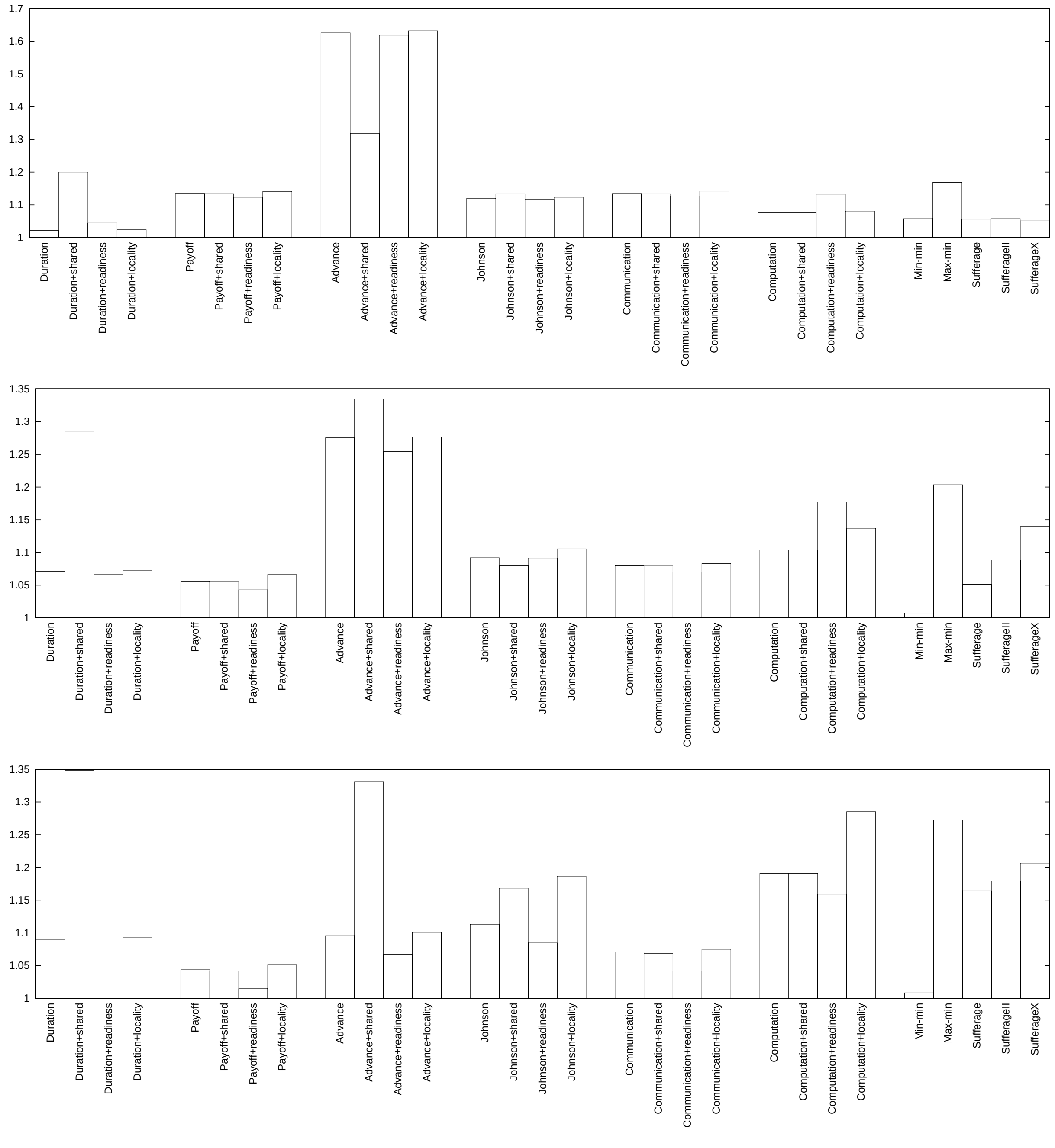


Figure 15: Relative performances of the schedules produced by the different heuristics on the *Random* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

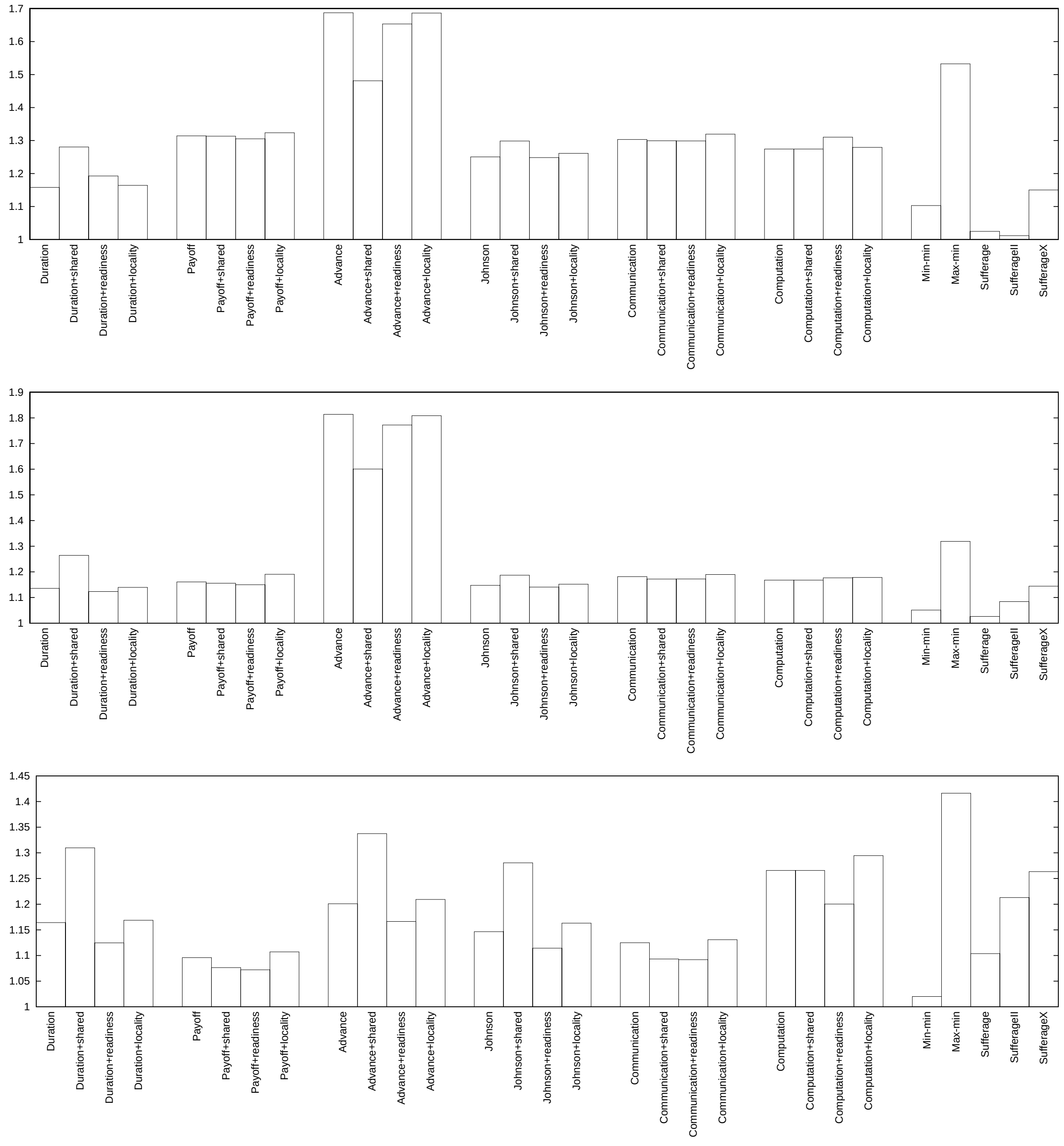


Figure 16: Relative performances of the schedules produced by the different heuristics on the *Partitioned* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

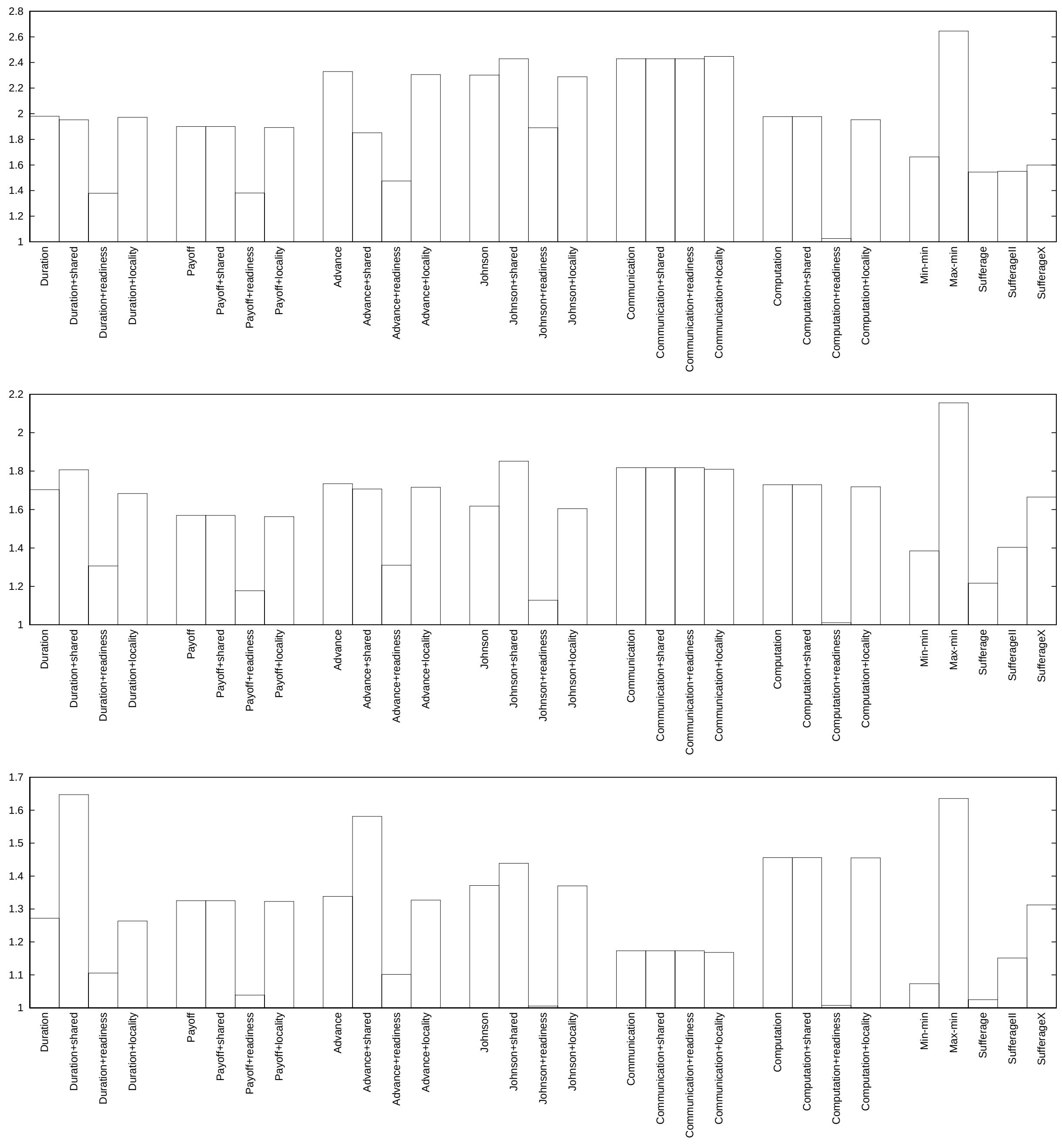


Figure 17: Relative performances of the schedules produced by the different heuristics on the *Forks* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.

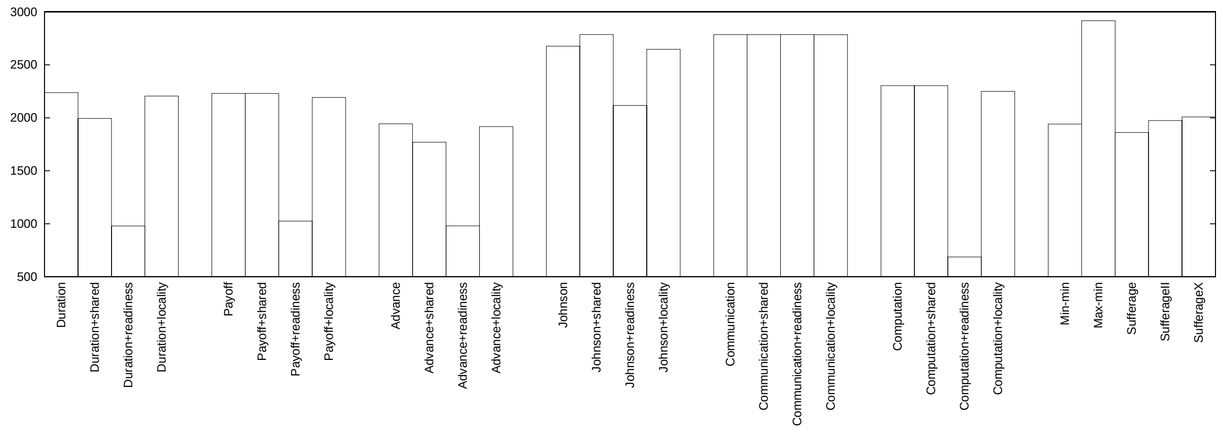
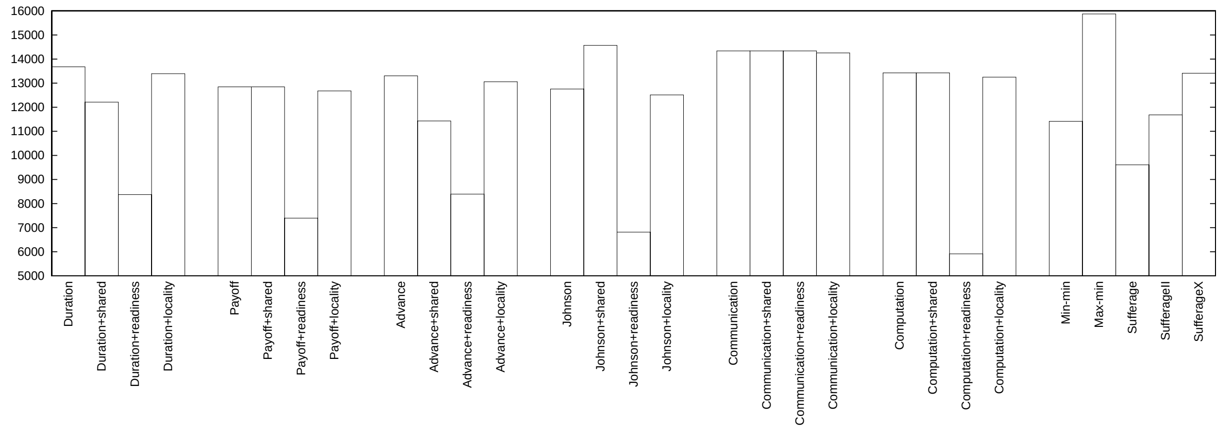
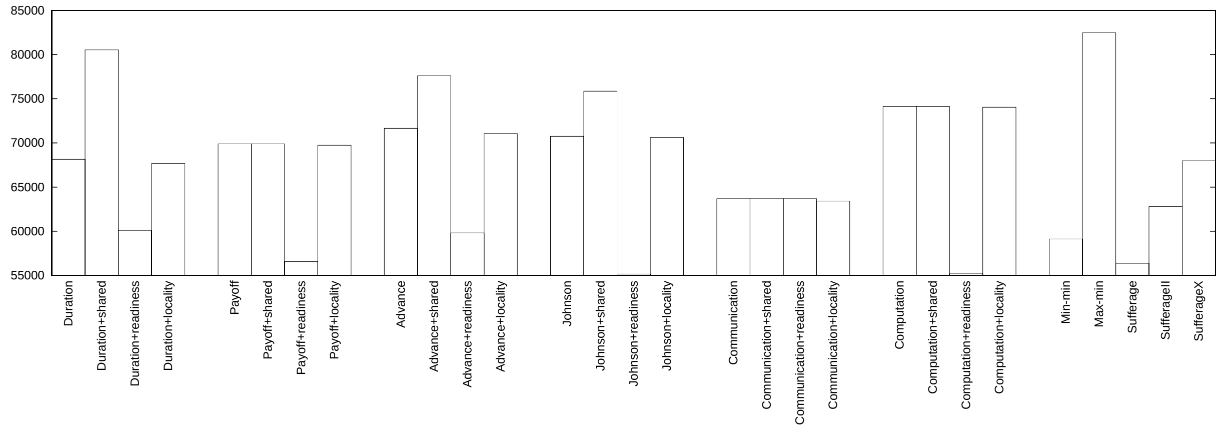




Figure 18: Total amount of files (sum of sizes) transfered by the schedules produced by the different heuristics on the *Forks* graphs with a communication to computation ratio equal, from top to bottom, to: 0.1, 1.0, and 10.