Prediction-based Coflow Scheduling

Olivier Brun, Balakrishna J. Prabhu and Oumayma Haddaji

***Abstract*—We explore the problem of minimizing the weighted average coflow completion time of coflows when flow sizes are unknown but unreliable predictions on them are available. We propose to use Sincronia, a 4-approximation algorithm when actual flow sizes are known, directly on the predictions. For this variant of Sincronia, we obtain a finite upper bound on the approximation ratio that combines unreliability on predictions and the weights of the coflows. On several numerical experiments, it is shown that this bound is too conservative, and that in practice Sincronia with predictions has a much better performance on average as well as in the worst-case.**

***Index Terms*—Coflow scheduling; weighted completion time; predictions; worst-case analysis**

I. Introduction

Cloud computing infrastructures have popularized the execution of massively parallel algorithms on vast amounts of data. Nowadays, most cloud providers allow their customers to launch parallel computations on their own data using Big Data applications such as MapReduce, Hadoop or Spark [10], [23]. These parallel applications typically alternate between computation stages and communication stages, during which the application's tasks exchange intermediate results using the datacenter network.

While parallel task scheduling is a well-explored topic, the scheduling of data flows in datacenter networks has only been studied since 2012, with the foundational work of Chowdhury and Stoica [7]. They introduced the concept of coflow to study the scheduling of data flows from concurrent parallel applications sharing the resources of a datacenter network. A coflow is defined as a collection of semantically-related data flows generated by a parallel application between two iterations. The coflow completion time (CCT) is the time at which all flows of a coflow have completed. In MapReduce, for instance, a coflow is composed of all flows sent from mapper to reducer nodes. These flows are launched as soon as the mapper nodes finish their computation tasks.

Chowdhury and Stoica showed that by scheduling flows from different concurrent coflows, it was possible to significantly improve application performance. Since then, efficient coflow scheduling methods aiming to minimize the average CCT have been studied in the literature. Some of them have been devised for the offline problem, in which all coflows to be scheduled are initially present in the system (or with known release dates), whereas others were devised for the online problem in which coflows arrive over time.

The work of O. Haddaji was done as intern at LAAS-CNRS while she was a student at Ecole Polytechnique de Tunisie.

All the authors are with LAAS-CNRS, University of Toulouse, CNRS, INSA, 31400 Toulouse, France.
E-mails: {brun, balakrishna.prabhu, oumayma.haddaji}@laas.fr.

Furthermore, coflow scheduling was mainly considered in the clairvoyant setting, where upon arrival of a coflow, the source and destination ports as well as the precise volume of each and every of its constituent flow are also revealed. Efficient algorithms (see Sincronia [1], e.g.) can be devised for this setting, but they require *a priori* information about coflows which is hard to obtain in practice. Therefore, some authors have also investigated the non-clairvoyant setting, where on a coflow arrival, only the number of flows and their input/output ports are revealed, while their volumes remain unknown.

The present work goes beyond the conventional clairvoyant and non-clairvoyant settings by exploring an alternative one in which predictions on the flow sizes are revealed to the coflow scheduler. These predictions could be obtained for instance from historical data on previous executions using machine learning (ML) techniques. The issue is that the actual flow size information remains unavailable and the ML predictions are unreliable. In that context, the main question we investigate is how to exploit these predictions for coflow scheduling, and whether it is even advisable to do so.

A. Contributions

We consider the problem of scheduling coflows with the objective of minimizing the average weighted CCT when only predictions on the actual volumes of coflows are given. We propose using Sincronia directly on the predictions and provide worst-case upper bound on its approximation ratio compared to the optimal weighted CCT computed using the actual volumes. This bound is obtained by combining two different upper bounds: one obtained from using the predictions and the other which is independent of the predictions.

Assuming that the predictions lie in the interval $[\mu_{min}x, \mu_{max}x]$, with x being the actual value, we first show that the CCT of Sincronia is at most $4(\mu_{min}^{-1}\mu_{max})^2$ of the optimal value. This upper bound coincides with that of Sincronia when there is no error in the predictions, that is, when $\mu_{min} = \mu_{max} = 1$. However, as the interval of predictions becomes large, this bound can become arbitrarily big, going to infinity as $\mu_{min} \to 0$ or $\mu_{max} \to \infty$. In this case, the bound becomes too large to be useful.

To overcome this drawback, we obtain a second upper bound that is independent of the quality of the predictions and that depends uniquely on the weights of the coflows: $(\sum_k w_k)/(\min_k w_k)$, where w_k is weight of coflow w_k.

The worst-case performance of Sincronia with predictions is then just the minimum of these two upper bounds. It can be seen that performance of Sincronia with predictions remains bounded even when the interval of predictions becomes large.

ISBN 978-3-903176-65-2 © 2024 IFIP

Finally, on several numerical examples we observe that in practice, these bounds are too conservative and the average performance of Sincronia with predictions remains close to that of optimal value computed with the actual volumes.

B. Organization

The paper is organized as follows. Section II is devoted to related work. Section III introduces the main notations used throughout the paper and presents the BlindFlow and Sincronia algorithms proposed for non-clairvoyant and clairvoyant coflow scheduling, respectively. In Section IV, we present our main theoretical results for the setting in which only unreliable ML predictions on flow sizes are available for coflow scheduling. Section V is devoted to numerical results. Finally, some conclusions are drawn and future research directions are suggested in Section VI.

II. Related Work

Since the introduction of the coflow concept by Chowdhury and Stoica [7], several coflow scheduling algorithms have been proposed [21]. The problem can be viewed as a concurrent open-shop scheduling problem, but with coupled resources as transmitting a flow requires capacity on both the input and output ports. Most studies have focused on minimizing the average CCT in the clairvoyant case, that is, assuming complete information on the coflows. Even under this simplifying assumption, the problem is known to be NP-hard [9] and inapproximable below a factor of 2 [18]. The offline minimisation of the average CCT can be formulated as an integer linear problem (ILP) [14], but this approach does not scale well. Research has therefore focused on approximation algorithms and heuristics [1], [2], [20], [22].

One of the first clairvoyant schedulers for minimizing the average CCT was Varys [9], which is still a basic reference today as it introduced many key concepts. Varys schedules coflows iteratively by considering at each stage the bottleneck port (given the remaining coflows to be scheduled), and choosing the coflow with the least volume on that port. The rate allocation to the flows of the chosen coflow is calculated so that they all terminate simultaneously. Another popular scheduler is Sincronia which was proposed by Agarwal et al. [1] and provides the best known approximation ratio. Sincronia is presented in details in Section III-B.

Although algorithms such as Varys or Sincronia provide fairly efficient solutions to the coflow scheduling problem, they require *a priori* information about the coflows, which is difficult to obtain in practice. Existing non-clairvoyant solutions such as Aalo [8] generalize the Least Attained Service (LAS) scheduling discipline to solve this problem (see also [11], [24]). However, they fail to identify the flows transmitted on bottleneck ports which result in a poor rate allocation. The Fai algorithm [13] attempts to solve this problem. Another very relevant work is [4] in which Bhimaraju et al. propose the BlindFlow algorithm for non-clairvoyant coflow scheduling. This algorithm, which generalizes the round robin (RR) scheduling discipline, is shown to be $8\,p$-approximate, where p is the maximum number of flows that any coflow can have.

To the extent of our knowledge, the present paper is the first one that considers a setting in which the actual flow sizes are unknown, but unreliable predictions on them are available. The improvement of online algorithms via ML predictions is a recent line of research which has attracted a lot of studies in the last few years, in particular for online scheduling, see e.g. [3], [6], [15], [16]. Though coflow scheduling significantly differs from the more traditional scheduling problems addressed in these references, our work is inspired by them, and in particular [12], [16].

III. Background Material

As customary for coflow scheduling, we represent the datacenter network using the *big switch model* [9], which abstracts out the datacenter network fabric as one big switch interconnecting servers. The underlying assumption is that the fabric core can sustain 100% throughput and only the ingress and egress ports are potential congestion points. We let $\mathcal{L}$ be the set of ports and denote by b_ℓ he capacity of port $\ell \in \mathcal{L}$.

We consider the offline setting in which all coflows are initially present in the system. We let $\mathcal{C} = \{1, 2, \ldots, n\}$ be the set of coflows. Each coflow k is a collection $F_k = \{1, 2, \ldots, n_k\}$ of flows, where flow j of coflow k is characterized by its source port $s^{k,j}$, its destination port $t^{k,j}$ and its volume $v^{k,j}$. Each coflow k may be assigned a weight w_k (default weight is 1). We also define $F_{k,\ell}$ as the set of flows in F_k that use port ℓ, either as ingress or egress port. In the following, we shall denote by C_k the completion time of coflow $k \in \mathcal{C}$. The objective is to schedule coflows so as to minimize the weighted CCT $\sum_{k\in\mathcal{C}} w_k C_k$.

A. Non-clairvoyant Scheduling of Coflows

In [4], the authors present theoretical guarantees on approximating the sum of coflow completion times in the non-clairvoyant setting, where on a coflow arrival, only the number of flows and their input-output ports are revealed, while the flow volumes are unknown.

They propose an algorithm, called BlindFlow, that divides the capacity of each port among all the flows that require that particular port in proportion to flow weights, so that the maximum rate that can be allocated to an unfinished flow $j \in F_k$ on port ℓ at time t is $r_\ell^{k,j}(t) = w_k b_\ell / W_\ell(t)$, where $W_\ell(t)$ is the sum of the weights of the unfinished flows using port ℓ at time t.

As each flow uses two ports, a natural rate allocation, which is referred to as the round robin (RR) allocation in the following, is to allocate rate $r_{RR}^{k,j}(t) = \min\{r_i^{k,j}(t), r_o^{k,j}(t)\}$ to unfinished flow $j \in F_k$ at time t, where i and o are its input and output ports, respectively.

As argued in [4], any performance guarantee obtained for BlindFlow also holds for better rate allocations, such as the RR allocation. The main result in [4] is stated in Theorem 1.

Theorem 1 ([4]): The rate allocation of BlindFlow is feasible and $8 \times p$ approximate, where $p = \max_{k\in\mathcal{C}} |F_k|$ is the maximum number of flows that any coflow can have.

B. Clairvoyant Scheduling of Coflows

For the clairvoyant setting, we present Sincronia, as the details of this algorithm are required for understanding our own contribution in Section IV. Sincronia is one of the most popular clairvoyant coflow scheduling algorithms [1]. It guarantees a $4-$approximation, the best known approximation ratio. We shall describe below the algorithm used by Sincronia to schedule a set of coflows. In addition to the notations introduced previously, we also let $p_{\ell,k} = \sum_{j \in F_{k,\ell}} \frac{v^{k,j}}{b_\ell}$ denotes the total transmission time of coflow $k \in \mathcal{C}$ at port $\ell \in \mathcal{L}$.

Sincronia first computes a σ-order, that is, a priority order over coflows. If $\sigma(k) < \sigma(k')$, then the flows of coflow k are transmitted in the datacenter network with strict priority over the flows of coflow k'. Sincronia then uses a greedy rate allocation which preserves the σ-order, that is, it is guaranteed that coflow k will be completed before coflow k' (that is $C_k \leq C_{k'}$) if $\sigma(k) < \sigma(k')$. The method for computing the σ-order is based on the following linear-programming problem:

$$\text{Minimize} \sum_{k \in \mathcal{C}} w_k C_k \qquad \text{(LP-Primal)}$$

s.t

$$\sum_{k \in S} p_{\ell,k} C_k \geq f_\ell(S), \quad \ell \in \mathcal{L} \text{ and } S \subseteq \mathcal{C}, \qquad (1)$$

$$C_k \geq 0, \quad k \in \mathcal{C}, \qquad (2)$$

where $f_\ell(S)$ is defined as

$$f_\ell(S) = \frac{1}{2} \sum_{k \in S} (p_{\ell,k})^2 + \frac{1}{2} \left(\sum_{k \in S} p_{\ell,k} \right)^2, \qquad (3)$$

for each port $\ell \in \mathcal{L}$ and each subset $S \subseteq \mathcal{C}$ of coflows. Inequalities (1) correspond to the so-called parallel inequalities [17], [19]. In non-preemptive single-machine scheduling problems, these inequalities define the convex hull of feasible completion time vectors in the absence of precedence constraints. Note that the formulation of problem LP-Primal does not account for the coupling between the resources of the fabric, that is, it ignores the fact that a flow can be transmitted only if its input and output ports are available. In that respect, problem LP-Primal is a relaxation of the original coflow scheduling problem. It is proven in [1] (see Lemma 1) that the optimal value of LP-Primal provides a lower bound for the coflow scheduling problem, that is, $\sum_{k \in \mathcal{C}} w_k C_k^* \leq \sum_{k \in \mathcal{C}} w_k C_k^{OPT}$, where the C_k^* are the optimal values of the completion times in problem LP-Primal whereas the C_k^{OPT} are the completion times in an optimal schedule.

Sincronia solves problem LP-Primal with a primal-dual algorithm. The dual problem is as follows

$$\text{Maximize} \sum_{\ell \in \mathcal{L}} \sum_{S \subseteq \mathcal{C}} f_\ell(S)\, y_{\ell,S} \qquad \text{(LP-Dual)}$$

s.t

$$\sum_{S: k \in S} \sum_{\ell \in \mathcal{L}} p_{\ell,k} y_{\ell,S} \leq w_k, \quad \text{for all } k \in \mathcal{C}, \qquad (4)$$

$$y_{\ell,S} \geq 0, \quad \text{for all } \ell \in \mathcal{L} \text{ and } S \subseteq \mathcal{C}. \qquad (5)$$

The Sincronia primal-dual algorithm is shown in Alg. 1.

Algorithm 1 Sincronia primal-dual algorithm

$y_{\ell,A} \leftarrow 0$ for all $\ell \in \mathcal{L}$ and $A \subseteq \mathcal{C}$.
$S \leftarrow \mathcal{C}$
for $t = n \dots 1$ **do**
  $b \leftarrow \text{argmax}_{\ell \in \mathcal{L}} \sum_{k \in S} p_{\ell,k}$ ▷ Bottleneck port
  $k^* \leftarrow \text{argmin}_{k \in S} \left(\frac{w_k}{p_{b,k}} \right)$ ▷ Largest weighted proc. time
  $C_{k^*} \leftarrow \sum_{k \in S} p_{b,k}$ ▷ Set primal variable
  $y_{b,S} \leftarrow \frac{w_{k^*}}{p_{b,k^*}}$ ▷ Set dual variable
  $w_k \leftarrow w_k - w_{k^*} \frac{p_{b,k}}{p_{b,k^*}}$ for all $k \in S$ ▷ Update weights
  $\sigma(t) \leftarrow k^*$ ▷ Set priority of coflow k^*
  $S \leftarrow S \setminus \{k^*\}$ ▷ Remove k^* from unscheduled coflows
end for

It is easy to prove that Algorithm 1 produces a primal feasible solution and a dual feasible solution. It can also be shown that the completion times of coflows produced by Algorithm 1 are such that $C_{\sigma(1)} \leq C_{\sigma(2)} \leq \ldots \leq C_{\sigma(n)}$. These properties are used in [1] to establish the approximation ratio of Sincronia, which is formally stated Theorem 2 below.

Theorem 2: The cost of the schedule obtained with Algorithm 1 is at most two times the optimal cost. As the Greedy rate allocation is also 2-optimal, this implies that Sincronia is a 4-approximation algorithm.

IV. Coflow Scheduling with Predictions

In this section, we consider the setting in which the actual flow sizes are unavailable and only unreliable ML predictions on them can be used for scheduling coflows. We first establish in Sections IV-A and IV-B the approximation ratio of Sincronia when ran on predictions as a function of the prediction error. We then propose in Section IV-C a consistent and robust prediction-based algorithm for coflow scheduling. Some proofs are omitted due to the lack of space but can be found in [5].

A. Sincronia with Predictions

In this section, we analyze the approximation ratio of Sincronia when, instead of the actual flow sizes $v^{k,j}$ of a coflow $k \in \mathcal{C}$, the algorithm only knows some predictions $\hat{v}^{k,j}$ on the flow sizes. These predictions can be written as $\hat{v}^{k,j} = v^{k,j} + \Delta v^{k,j}$, where $\Delta v^{k,j}$ represents the prediction error on the size of flow $j \in F_k$ and lies in the interval $[-v^{k,j}, \infty)$. The predicted total transmission time of coflow $k \in \mathcal{C}$ on port $\ell \in \mathcal{L}$ is then

$$\hat{p}_{\ell,k} = \sum_{j \in F_k} \frac{\hat{v}^{k,j} x_\ell^{k,j}}{b_\ell} = p_{\ell,k} + \eta_{\ell,k},$$

where $\eta_{\ell,k} = \sum_{j \in F_k} \frac{\Delta v^{k,j} x_\ell^{k,j}}{b_\ell}$ represents the prediction error on the transmission time on port ℓ. Define

$$\mu_{min} = 1 + \min_{\ell,k} \left(\frac{\eta_{\ell,k}}{p_{\ell,k}} \right) \text{ and } \mu_{max} = 1 + \max_{\ell,k} \left(\frac{\eta_{\ell,k}}{p_{\ell,k}} \right),$$

We then have the following inequalities:

$$\mu_{min}\, p_{\ell,k} \leq \hat{p}_{\ell,k} \leq \mu_{max}\, p_{\ell,k}, \tag{6}$$

for all ports $\ell \in \mathcal{L}$ and all coflows $k \in \mathcal{C}$.

When ran on predictions, Sincronia produces a σ-order and compute coflow completion times $\hat{C}_k$ which are a feasible primal solution, that is, $\hat{\mathbf{C}}$ is a feasible solution to problem LP-Primal in which the actual processing times $p_{\ell,k}$ have been replaced by the predicted processing times $\hat{p}_{\ell,k}$. Likewise, the constants $f_\ell(S)$ have been replaced by the constants $\hat{f}_\ell(S)$ obtained by using the predicted values $\hat{p}_{\ell,k}$ instead of the actual values $p_{\ell,k}$. As a primal-dual algorithm, Sincronia also computes the dual variables $\hat{y}_{\ell,S}$ which provide a feasible solution to dual problem in which predicted processing times are used instead of the actual ones.

We shall assume for simplicity that Sincronia schedule the coflows in the order $1, 2, \ldots, n$, which implies that $\hat{C}_1 \leq \hat{C}_2 \leq \ldots \leq \hat{C}_n$. Our goal is to obtain an upper bound on the approximation ratio obtained by scheduling coflows in that order, as a function of μ_{min} and μ_{max}. Note that $\hat{C}_k$ is only the *estimated completion time* of coflow k. In the following we denote by $\tilde{C}_k$ the actual completion time of coflow $k \in \mathcal{C}$ obtained by scheduling coflows in the order $1, 2, \ldots, n$. We first establish a bound on the predicted weighted sum of completion times.

Lemma 1: Let $\hat{C}_k$ be the predicted completion time of coflow k when Algorithm 1 is ran over the predictions, and C_k^{OPT} be its completion time in an optimal schedule. Then

$$\sum_{k \in \mathcal{C}} w_k \hat{C}_k \leq \frac{2\,\mu_{max}^2}{\mu_{min}} \sum_{k \in \mathcal{C}} w_k C_k^{OPT}, \tag{7}$$

Proof: Please see [5]. ∎

We now use Lemma 1 to establish the approximation ratio of the prediction-based Sincronia algorithm.

Theorem 3: Scheduling coflows in the order determined by Sincronia with predictions as inputs yields a schedule whose weighted sum of completion times is at most $4 \times \left(\frac{\mu_{max}}{\mu_{min}}\right)^2$ the optimal one.

Proof: Please see [5]. ∎

Note that the approximation ratio in Theorem 3 is identical to that of Sincronia for perfect predictions, that is, when $\mu_{max} = \mu_{min} = 1$. More generally, the prediction-based Sincronia algorithm computes the same σ-order as the clairvoyant one when $\mu_{max} = \mu_{min}$. However, the bound in Theorem 3 suggests that Sincronia is not robust to prediction errors, as its approximation ratio grows unboundedly as the ratio μ_{max}/μ_{min} increases. We shall see in Section IV-B that this is not the case.

To illustrate the above result, consider the scenario where, although unreliable, the relative prediction error $\frac{\Delta v^{k,j}}{v^{k,j}}$ on the flow sizes is guaranteed to lie in the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$ uniformly for all flows j and all coflow $k \in \mathcal{C}$. In that case, it is easy to see $\mu_{min} \geq \frac{1}{2}$ and $\mu_{max} \leq \frac{3}{2}$, so that Sincronia with predictions as inputs guarantees a $36-$approximation, that is, a worst-case approximation ratio which is $9\times$ that of Sincronia when ran over the actual flow sizes. As we shall see in Section V, the experimental performance of the prediction-based Sincronia algorithm is far better than this theoretical guarantee.

B. Approximation ratio independent of predictions

The bound in Theorem 3 depends upon the quality of predictions. If μ_{min} is close to 0, then the bound goes towards infinity. This happens because the bound makes use of the completion times, $\hat{C}_k$, computed from the predicted processing times, $\hat{p}_{\ell,k}$s, which can highly different from the actual processing times. For example, the predictions can be of the order of ϵ a small quantity while the actual ones could be of the order of 1. The predicted completion times will be of the order of ϵ which is much smaller than the actual completion times which will be of the order of 1.

A bound that does suffer from this drawback can be obtained from the observation that Sincronia is an order-based scheduling algorithm. Its total weighted completion time will be upper bounded by that of the worst-case order, that is:

$$\sum_k w_k \tilde{C}_k \leq \max_\sigma \sum_k w_{\sigma(k)} C_{\sigma(k)}. \tag{8}$$

Beside, the optimal total weighted completion time is at least that of the last coflow in an optimal order:

$$\sum_k w_k C_k^{OPT} \geq w_{\sigma^*(n)} C_{\sigma^*(n)} \geq \left(\min_k w_k \right) \max_\ell \sum_k p_{\ell,k}$$

where $\sigma^*(n)$ is the index of the coflow finishing last. The second inequality is a consequence of the fact that the latest completion time cannot be earlier than the time required to empty the most loaded link failing which there will be a coflow that would not have finished.

Combining the above two observations, we get the following bound on the approximation ratio of Sincronia with predictions:

$$\frac{\sum_k w_k \tilde{C}_k}{\sum_k w_k C_k^{OPT}} \leq \frac{\max_\sigma \sum_k w_{\sigma(k)} C_{\sigma(k)}}{(\min_k w_k) \max_\ell \sum_k p_{\ell,k}} \tag{9}$$

Using this inequality, we arrive at the following approximation ratio.

Theorem 4: Sincronia with predictions is a $2\frac{\bar{w}}{w_{min}}$ approximation algorithm, where $\bar{w} = \sum_k w_k$ and $w_{min} = \min_k w_k$.

Proof: For a given order σ, we have the following inequality bounding the completion times of coflows:

$$C_{\sigma(k)} \leq 2 \max_{\ell} \sum_{j=1}^{k} p_{\ell,\sigma(j)} \tag{10}$$

from which we can bound the numerator of the RHS of (9) as

$$\begin{aligned} \max_{\sigma} \sum_{k} w_{\sigma(k)} C_{\sigma(k)} &\leq 2 \max_{\sigma} \sum_{k} w_{\sigma(k)} \max_{\ell} \sum_{j=1}^{k} p_{\ell,\sigma(j)} \\ &\leq 2 \max_{\sigma} \left(\sum_{k} w_{\sigma(k)} \right) \max_{\ell} \sum_{j=1}^{n} p_{\ell,\sigma(j)} \\ &= 2 \left(\sum_{k} w_k \right) \max_{\ell} \sum_{j=1}^{n} p_{\ell,j} \end{aligned} \tag{11}$$

For the second inequality, we have replaced k by n in $\sum_{j=1}^{k} p_{\ell,\sigma(j)}$ which makes this sum independent of k. For the equality in the third line, observe that $\sum_{j=1}^{k} p_{\ell,\sigma(j)}$ as well as the first sum are independent of the order since we are now including all the n coflows and not just the first k. The claimed ratio now follows by substituting (11) into (9). ∎

Note that Theorem 4 implies that Sincronia is robust to prediction errors. In particular, when all coflow weights are equal, as considered in Section V, we have $\bar{w} = n w_{min}$ and the average CCT of Sincronia with predictions as inputs cannot be worse than $2n$ the optimal one, independently of the prediction error. Moreover, combining the upper bounds in Theorem 3 and Theorem 4, we obtain our final upper bound on the approximation ratio of the prediction-based algorithm.

Corollary 1: The approximation ratio of Sincronia with predictions as inputs is upper bounded by $\min \left\{ 4 \times \left(\frac{\mu_{max}}{\mu_{min}} \right)^2, 2 \frac{\bar{w}}{w_{min}} \right\}$.

Proof: The proof follows from theorems 3 and 4. ∎

C. A consistent and robust prediction-based algorithm

In [16], the authors propose a prediction-based approach for scheduling jobs of unknown sizes on a single machine so as to minimize the average completion time of the jobs. In this section, we extend this approach to coflow scheduling. We combine the Sincronia clairvoyant coflow scheduling algorithm with the RR non-clairvoyant one. Sincronia uses predictions to schedule coflows in the fabric over a fraction λ of time, while RR schedules the coflows the rest of the time. In other words, the rate allocation to flow $j \in F_k$ at time t is $r^{k,j}(t) = \lambda \, r^{k,j}_{SP}(t) + (1-\lambda) \, r^{k,j}_{RR}(t)$, where $r^{k,j}_{SP}(t)$ is the rate allocated to this flow by Sincronia when ran over predictions. We can obtain the following guarantee on the resulting algorithm.

Theorem 5: Running in parallel Sincronia with predictions and RR yields an algorithm with competitive ratio

$$\min \left(\frac{4}{\lambda} \left(\frac{\mu_{max}}{\mu_{min}} \right)^2, \; \frac{2}{\lambda} \frac{\bar{w}}{w_{min}}, \; \frac{8\,p}{1-\lambda} \right) \tag{12}$$

In particular, this algorithm is $4/\lambda$-consistent and $\min \left\{ \frac{2}{\lambda} \frac{\bar{w}}{w_{min}}, \frac{8\,p}{1-\lambda} \right\}$-robust. That is, its approximation ratio is $4/\lambda$ for perfect predictions and by $\min \left\{ \frac{2}{\lambda} \frac{\bar{w}}{w_{min}}, \frac{8\,p}{1-\lambda} \right\}$ independently of the prediction error.

Proof: The stated result is a direct application of Lemma 3.1 in [16] to coflow scheduling. The proof then directly follows from Theorem 1 in [16] and Theorem 1. ∎

The algorithm combining Sincronia with predictions and RR gives an option to trade-off consistency and robustness. In particular, greater trust in the predictions suggests setting λ close to 1, as this leads to a competitive ratio which is approximately 4 when $\mu_{max} \approx \mu_{min} \approx 1$. In that case, the algorithm performs as well as the clairvoyant Sincronia in the worst case. On the other hand, setting λ close to 0 is conservative and guarantees a competitive ratio of $8 \times p$ (as the non-clairvoyant RR algorithm) even for terrible predictions. As the prediction-based Sincronia is robust to prediction errors, this is however relevant only if $\bar{w} > 4 p w_{min}$.

V. Numerical results

In this section, we compare the average CCTs obtained with the prediction-based Sincronia algorithm against that of other clairvoyant and non-clairvoyant algorithms on random problem instances. We first describe below the procedure for generating these instances. In Section V-A, we present the numerical results obtained on small instances for which the clairvoyant offline optimum can be computed. Section V-B provides the results obtained on larger problem instances, whereas Section V-C investigates the performance obtained by running the prediction-based Sincronia and RR in parallel.

We generate random problem instances with an algorithm which works as follows. The algorithm takes as input the number of instances to be generated, the number n of coflows and the number L of ports in each instance. For each instance, it is assumed the the first $L/2$ ports are input ports, while the other ports are output ports. The capacity of each port is 1. For each coflow k, the weight w_k is 1 and there is a flow between input port ℓ and output port ℓ' with probability q, where q is another input parameter of the algorithm. The number of flows per coflow therefore follows a binomial distribution[1] and the mean number of flows per coflow is $qL^2/4$. The size of each flow is randomly generated from a (truncated) Gaussian distribution with mean m and standard deviation σ, these parameters being also given as inputs to the algorithm.

Given a problem instance, we generate a given number of random predictions as follows. We compute the predicted volumes as $\hat{v}^{k,j} = u^{k,j} \times v^{k,j}$, where $u^{k,j}$ is uniformly distributed in the interval $[1-\delta, 1+\delta]$. We vary the coefficient δ so as to evaluate the impact of the prediction quality on the performance of the scheduling algorithm.

A. Comparison against the clairvoyant offline optimum

We first consider small instances for which the mixed-integer linear program proposed in [14] can be solved. We

[1]However, it is assumed that there is at least one flow per coflow.

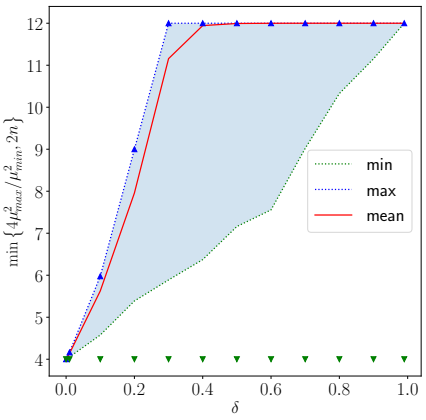

Fig. 1. Minimum, maximum and mean values of $\min\left\{4\frac{\mu_{max}^2}{\mu_{min}^2}, 2n\right\}$ as a function of δ for one of the type-1 instances with $100,000$ predictions.

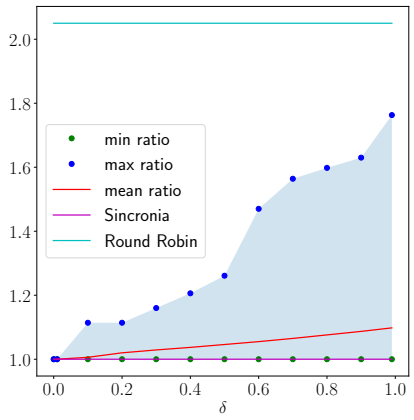

Fig. 2. Minimum, maximum and average values of the ratio $\frac{SIN_{pred}}{OPT}$ as a function of δ for one of the type-1 instances.

use Gurobi as MILP solver, with a time limit of 15 minutes. Two types of instances are considered:

- **Type-1 instances**: We randomly generate $1,000$ instances with $n = 6$ coflows and $L = 6$ ports. We use the parameters $m = 5$, $\sigma = 2$ and $q = \frac{1}{3}$, so that the average number of flows per coflow is 3. The parameter δ takes values in the set $\{0, 0.01, 0.1, \ldots, 0.9, 0.99\}$. For each instance and each value of δ, we generate randomly either $10,000$ or $100,000$ predictions of flow sizes.
- **Type-2 instances**: We randomly generate 100 instances with $n = 8$ coflows, $L = 12$ ports and $q = \frac{1}{4}$, so that there are on average 9 flows per coflow. The other parameters are as above, except that we use only $10,000$ predictions for each value of δ and each instance.

We first evaluate how the upper bound $\min\left\{4\frac{\mu_{max}^2}{\mu_{min}^2}, 2n\right\}$ on the approximation ratio of Sincronia evolves as we vary the prediction error δ. Figure 1 shows the evolution of the empirical minimum, maximum and mean values obtained for this quantity as the prediction error δ varies for one of the type-1 instances (for which, $n = 6$). The results obtained with the other type-1 or type-2 instances are similar. The figure also shows with dots the theoretical minimum value 4 and maximum value $\min\left\{4(1+\delta)^2/(1-\delta)^2, 12\right\}$ obtained for each value of δ. We note from Figure 1 that the empirical maximum matches the theoretical value and that the average value is quite close to the maximum value. For $\delta \geq 0.4$, the performance guarantee for the prediction-based Sincronia corresponds to that of the worst σ-order.

We now consider the experimental performance of Sincronia when ran over predictions. Figure 2 shows the minimum, maximum and mean values obtained for the ratio $\frac{SIN_{pred}}{OPT}$ as δ varies for one of the type-1 instances, where SIN_{pred} is the average CCT obtained by Sincronia using predictions whereas OPT is the clairvoyant offline optimum computed with Gurobi. We generated $100,000$ predictions for each value of δ. We also denote by SIN and RR the average CCT obtained with the clairvoyant Sincronia and RR, respectively, and show the ratios $\frac{SIN}{OPT}$ and $\frac{RR}{OPT}$, which are independent of δ, on the figure. Note that for perfect predictions, that is for $\delta = 0$, the experimental performance of Sincronia with predictions matches that of the clairvoyant Sincronia. The minimum value of $\frac{SIN_{pred}}{OPT}$ is never below 1.0, as expected, and its maximum value is at most 1.76, that is, at most a 76% degradation with respect to the clairvoyant offline optimum. However, on average, the performance degradation is fairly low since it is in the order of 5% for this instance.

Figure 3 summarizes the results obtained over $1,000$ type-1 instances, using $10,000$ predictions. Here, the mean value of $\frac{SIN_{pred}}{OPT}$ is obtained by averaging over the $1,000$ instances and the $10,000$ predictions. Similarly, its maximum (resp. minimum) value corresponds to the greatest (least) ratio obtained over the $1,000$ instances and the $10,000$ predictions. The figure also shows the average values of the ratios $\frac{SIN}{OPT}$ and $\frac{RR}{OPT}$. Note that for $\delta = 0$, even though there is no prediction error, the minimum, maximum, and mean values of $\frac{SIN_{pred}}{OPT}$ do not coincide as they are obtained for different instances. We observe that the minimum value of the ratio remains at 1, which shows that wrong predictions do not necessarily harm Sincronia and can even improve its performance in some cases. We also observe that on average the performance of the prediction-based Sincronia remains remarkably close to that of the clairvoyant Sincronia, even for terrible predictions. We note that Sincronia with predictions performs remarkably well in comparison to RR even with poor predictions.

Figure 4 is similar to Figure 3 and presents the ratios $\frac{SIN_{pred}}{OPT}$ obtained over 100 type-2 instances, using $10,000$ predictions for each one and each value of δ. We note that the minimum value occasionally falls below 1 because for some instances Gurobi was unable to reach the optimal solution within the time limit of 15 minutes. The maximum value does not exceed 1.6. Moreover the average performance of the prediction-based Sincronia always remains quite close to

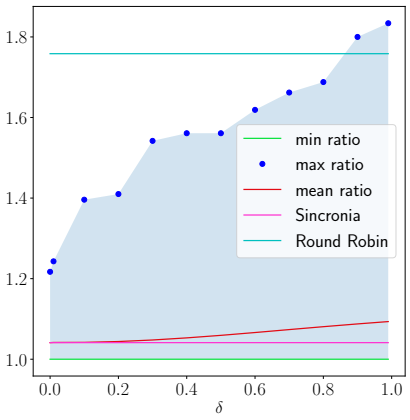

Fig. 3. Minimum, maximum and average values of the ratio $\frac{SIN_{pred}}{OPT}$ over $1,000$ type-1 instances as a function of δ.

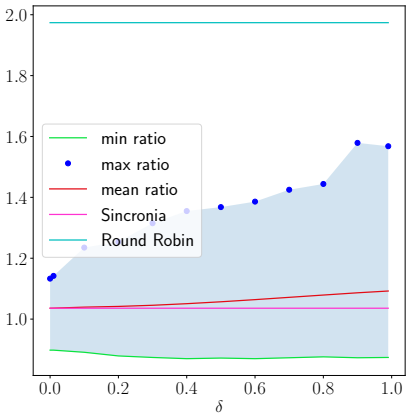

Fig. 4. Minimum, maximum and average values of the ratio $\frac{SIN_{pred}}{OPT}$ over 100 type-2 instances as a function of δ.

that of the clairvoyant Sincronia. Another observation is that the average approximation ratio of RR remains close to 2, a significantly higher value than what is achieved by Sincronia, even with extremely poor predictions.

B. Comparison against the clairvoyant Sincronia

We now consider larger problem instances for which the clairvoyant offline optimum cannot be computed. Specifically, we consider four datasets with $n = 10$, 15, 20 and 30 coflows, respectively. Each dataset contains 100 randomly generated instances with $L = 10$ ports, $m = 5$, $\sigma = 2$ and $q = 0.2$ (that is, 5 flows per coflow on average). For each dataset and each instance in this dataset, we vary the parameter δ from 0 to 0.99 and generate $10,000$ random predictions of flow sizes. The number of coflows and the number of ports are kept to relatively low values as $4 \times 100 \times 10,000 \times 12 = 48,000,000$ network simulations are required, which is of course a time-consuming affair.

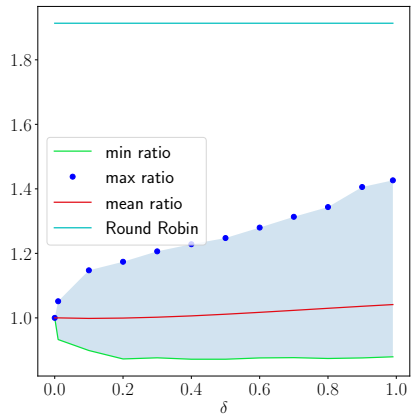

Fig. 5. Minimum, maximum and average values of the ratio SIN_{pred}/SIN over 100 instances with 15 coflows and 10 ports as a function of δ.

As the clairvoyant offline optimum is no more available, we choose as reference the clairvoyant Sincronia. Figure 5 plots the minimum, maximum and average values of the ratio $\frac{SIN_{pred}}{SIN}$ as a function of δ. We observe that the average performance of Sincronia with predictions is quite close to that it obtains in the clairvoyant setting, even with extremely poor predictions (at most a 5% degradation). The worst-case performance increases with δ, but remains surprisingly close to that obtained by the clairvoyant Sincronia (at most $+42\%$ with respect to the clairvoyant Sincronia). We also observe that on average RR performs poorly as compared to the clairvoyant Sincronia.

Figure 6 plots the results obtained for 30 coflows. The same observations can be made as for Figure 5 . However, a comparison with the latter figure shows that increasing the number of coflows while maintaining the same number of links leads to enhanced worst-case performance for the prediction-based Sincronia (at most $+35\%$ for the maximum value). The results obtained with 10 and 20 coflows are not presented here due to the lack of space but are similar.

C. Running Sincronia with predictions and RR in parallel

The empirical results obtained above show that Sincronia performs very well and much better than RR even when feed with terrible predictions. In this section, we therefore investigate the performance obtained by running the two algorithms in parallel, focusing on a large value of λ. We choose $\lambda = 0.95$.

Figure 7 shows the maximum and average values of the ratio $\frac{\lambda SIN_{pred}+(1-\lambda)RR}{SIN}$ over 100 instances with 6 coflows and 6 ports as a function of δ for $\lambda = 0.95$ and for $\lambda = 1.0$. As expected, a slight degradation of the average performance is observed when the weight of RR is increased from 0% to 5%. In terms of worst-case empirical performance, we do not observe any improvement, except for $\delta = 0.99$. Results obtained on larger problem instances with more coflows are

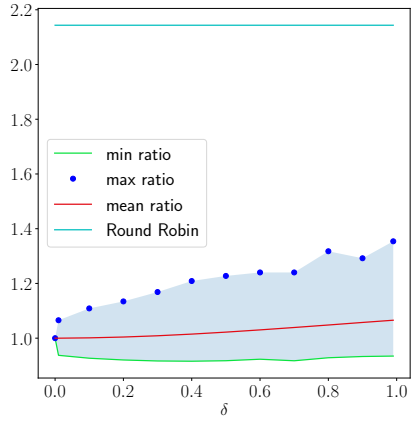

Fig. 6. Minimum, maximum and average values of the ratio SIN_{pred}/SIN over 100 instances with 30 coflows and 10 ports as a function of δ.

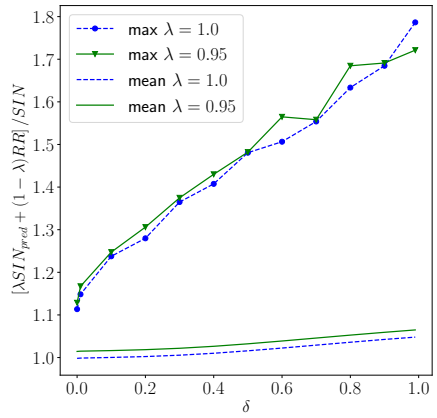

Fig. 7. Maximum and average values of $\frac{\lambda SIN_{pred}+(1-\lambda)RR}{SIN}$ for $\lambda = 0.95$ and $\lambda = 1.0$ over 100 instances with $n = L = 6$ as a function of δ.

similar. We therefore conclude that there is no clear benefit of running both algorithms in parallel.

VI. Conclusion

We have established an upper bound on the approximation ratio of Sincronia with predictions as inputs and proposed to combine it with a RR rate allocation. Our numerical results show that Sincronia performs well even when feed with terrible predictions, with average CCT which are quite close on average to those obtained under the clairvoyant setting. The worst performance degradation is limited and tends to decrease as the number of coflows is increased. These results suggest that there is little or no benefit in combining Sincronia with predictions and a RR rate allocation. Moreover, they suggest that operating Sincronia with ML predictions could be an efficient solution in practical scenarios with hundreds of coflows and ports. Confirmation of these conclusions for other models of the prediction errors will require further study and is left for future work.

References

[1] Saksham Agarwal, Rachit Agarwal, Shijin Rajakrishnan, David Shmoys, Akshay Narayan, and Amin Vahdat. Sincronia: Near-Optimal Network Design for Coflows. In *Proc. ACM SIGCOMM*, pages 16–29, 2018.

[2] Saba Ahmadi, Samir Khuller, Manish Purohit, and Sheng Yang. On scheduling coflows. *Algorithmica*, 82:3604 – 3629, 2020.

[3] Eric Balkanski, Vasilis Gkatzelis, and Xizhi Tan. Strategyproof scheduling with predictions. In Yael Tauman Kalai, editor, *ITCS 2023*, volume 251 of *LIPIcs*, pages 11:1–11:22, 2023.

[4] Akhil Bhimaraju, Debanuj Nayak, and Rahul Vaze. Non-clairvoyant scheduling of coflows. In *WiOPT 2020*, pages 81–88, Volos, Greece, June 2020. IEEE.

[5] Olivier Brun, Balakrishna Prabhu, and Oumayma Haddaji. Prediction-based Coflow Scheduling. https://laas.hal.science/hal-04231370, October 2023.

[6] Woo-Hyung Cho, Shane G. Henderson, and David B. Shmoys. Scheduling with predictions. *CoRR*, abs/2212.10433, 2022.

[7] Mosharaf Chowdhury and Ion Stoica. Coflow: A networking abstraction for cluster applications. In *Proc. ACM HotNets*, pages 31–36, Redmond, Washington, 2012.

[8] Mosharaf Chowdhury and Ion Stoica. Efficient coflow scheduling without prior knowledge. *SIGCOMM Comput. Commun. Rev.*, 45(4):393–406, August 2015.

[9] Mosharaf Chowdhury, Yuan Zhong, and Ion Stoica. Efficient Coflow Scheduling with Varys. In *Proc. ACM SIGCOMM*, pages 443–454, 2014.

[10] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[11] Yuanxiang Gao, Hongfang Yu, Shouxi Luo, and Shui Yu. Information-agnostic coflow scheduling with optimal demotion thresholds. In *IEEE ICC*, 2016.

[12] Sungjin Im. Non-clairvoyant scheduling with predictions. In *Proc. International Symposium on Artificial Intelligence and Mathematics 2022 (ISAIM 2022), Fort Lauderdale, Florida, USA, Jan 3-5*, 2022.

[13] Libin Liu, Hong Xu, Chengxi Gao, and Peng Wang. Bottleneck-aware coflow scheduling without prior knowledge. In *IEEE INFOCOM Workshops 2020*, pages 50–55, 2020.

[14] Youcef Magnouche, Sébastien Martin, Jérémie Leguay, Francesco De Pellegrini, Rachid Elazouzi, and Cédric Richier. Branch-and-benders-cut algorithm for the weighted coflow completion time minimization problem. In *Proc.INOC 2022, Aachen, Germany*, 2022.

[15] Michael Mitzenmacher. Scheduling with predictions and the price of misprediction. In Thomas Vidick, editor, *ITCS 2020*, volume 151 of *LIPIcs*, pages 14:1–14:18, USA, January 2020.

[16] Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving online algorithms via ml predictions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[17] M. Queyranne. Structure of a simple scheduling polyhedron. *Mathematical Programming*, 58:263–285, 1993.

[18] Sushant Sachdeva and Rishi Saket. Optimal inapproximability for scheduling problems via structural hardness for hypergraph vertex cover. In *2013 IEEE Conf. on Computational Complexity*, 2013.

[19] Andreas S. Schulz. Scheduling to minimize total weighted completion time: Performance guarantees of lp-based heuristics and lower bounds. In *Conf. on Integer Prog. and Combinatorial Optimization*, 1996.

[20] Li Shi, Yang Liu, Junwei Zhang, and Thomas Robertazzi. Coflow scheduling in data centers: routing and bandwidth allocation. *IEEE Trans. Parallel Distrib. Syst.*, 32(11):2661–2675, 2021.

[21] Shuo Wang, Jiao Zhang, Tao Huang, Jiang Liu, Tian Pan, and Yunjie Liu. A survey of coflow scheduling schemes for data center networks. *IEEE Commun. Mag.*, 56(6):179–185, 2018.

[22] Zhiliang Wang, Han Zhang, Xingang Shi, Xia Yin, Yahui Li, Haijun Geng, Qianhong Wu, and Jianwei Liu. Efficient scheduling of weighted coflows in data centers. *IEEE Trans. on Par. and Dist. Systems*, 30(9), 2019.

[23] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, Ion Stoica, et al. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

[24] Tong Zhang, Fengyuan Ren, Ran Shu, and Bo Wang. Scheduling coflows with incomplete information. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–10, 2018.