Ordonnancement de coflows dans les réseaux de data centers

Olivier Brun

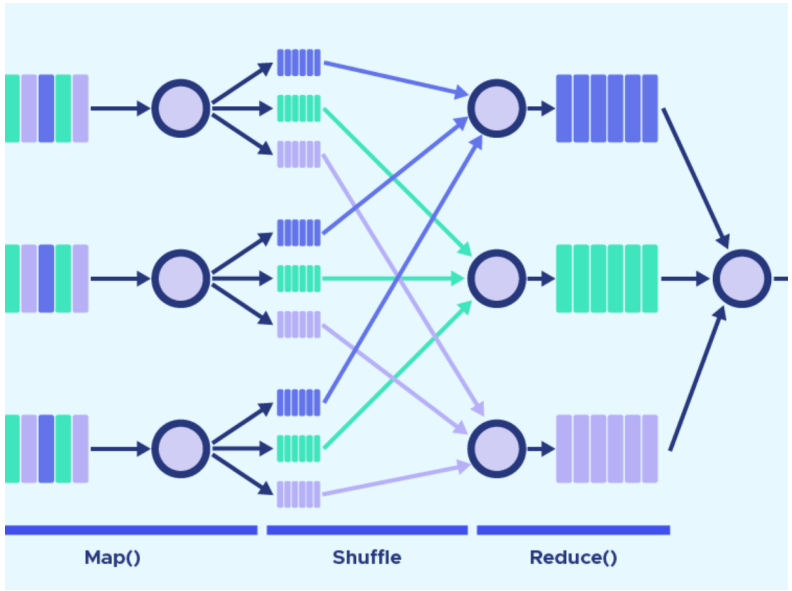
INSA, 2024

Outline

INTRODUCTION

Contexte

- Applications itératives massivement parallèles de big data : Hadoop MapReduce, Apache Spark
- Transferts de données massifs dans le réseau du datacenter (shuffle phase)
  - ✔ Peuvent représenter jusqu'à 50% du temps d'exécution total



- Le réseaux traite chacun des flots concurrents de manière indépendante !

Ordonnancement de Coflow

- Ordonnancement des flots concurrents des différentes applications
  - ✔ Coflow : ensemble des flots concurrents d'une même application
  - ✔ Coflow Completion Time (CCT) : temps de terminaison du dernier flot
  - ✔ Objectif : minimisation du CCT moyen
- Ordonnancement clairvoyant
  - ✔ Ports src/dst et volume des flots connus à l'arrivée d'un coflow
  - ✔ NP-hard, inapproximable en dessous d'un facteur 2
  - ✔ Algorithmes d'approximation efficaces : Varys, Sincronia [1]
- Ordonnancement non clairvoyant
  - ✔ Les volumes des flots sont inconnus
  - ✔ Généralisation des politiques d'ordonnancement *Least Attained Service* (e.g., Aalo) ou *Round Robin* (e.g., BlindFlow)

1.

- ☞ M. Shafiee et *al.*, An improved bound for minimizing the total weighted completion time of coflows in datacenters, IEEE/ACM Trans. Netw., vol. 26, no. 4, 2018.
- ☞ S. Agarwal et *al.*, Sincronia : Near-optimal network design for coflows. in Proc. ACM SIGCOMM, 2018.
- ☞ M. Chowdhury et *al.*,. Near optimal coflow scheduling in networks, in Proc. ACM SPAA, 2019.

FORMULATION DU PROBLEME

Modèle du système et notations

- Modèle du Big-Switch : capacité b_ℓ pour le port ℓ
- Ordonnancement offline d'un ensemble $\mathcal{C} = \{1, 2, \ldots, n\}$ de coflows
  - ✔ Le coflow k est une collection F_k de flots, le flot j ayant un volume $v^{k,j}$
  - ✔ $F_{k,\ell}$ est l'ensemble des flots du coflow k qui utilisent le port ℓ
  - ✔ $r^{k,j}(t)$ est le débit alloué au flot $j \in F_k$ à l'instant t
  - ✔ C_k représente le CCT du coflow k
- Formulation du problème

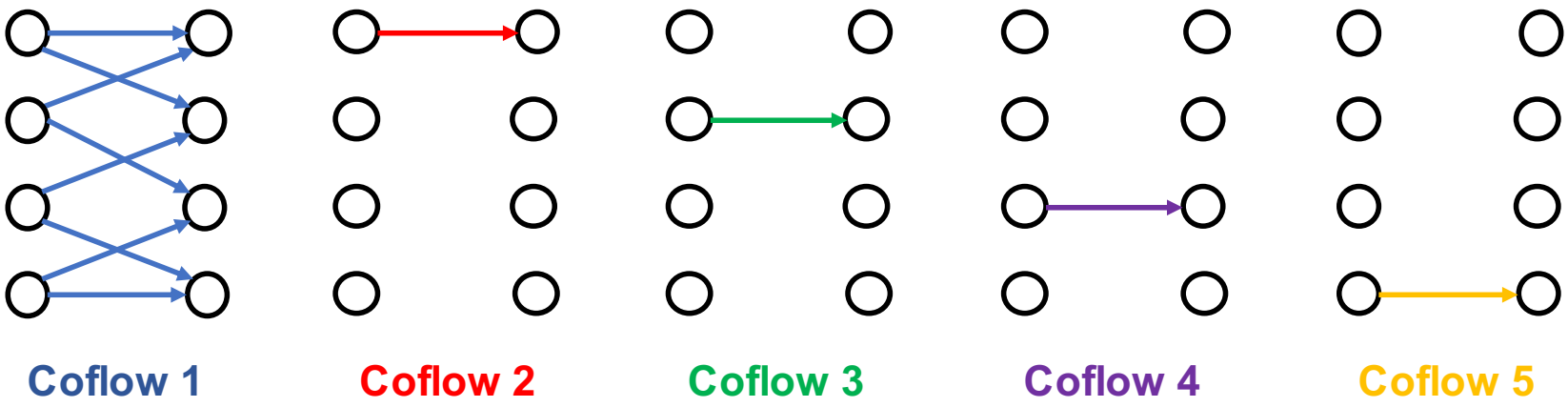
$$\min_{r} \sum_{k \in \mathcal{C}} C_k \tag{P1}$$

$$\text{s.t.} \sum_{k \in \mathcal{C}} \sum_{j \in F_{k,\ell}} r^{k,j}(t) \leq b_\ell, \quad \forall \ell \in \mathcal{L}, \forall t \in \mathcal{T}, \tag{1}$$

$$\int_0^{C_k} r^{k,j}(t)\, dt \geq v^{k,j}, \quad \forall j \in F_k, \forall k \in \mathcal{C}, \tag{2}$$

Exemple

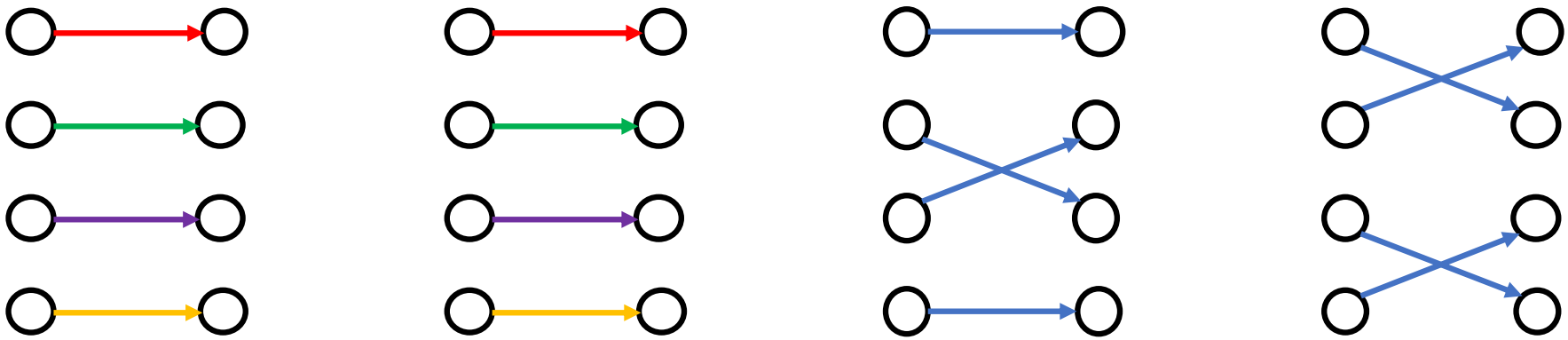
- Tous les ports ont la même bande passante normalisée de 1
- Tous les flots du coflow 1 ont un volume de 1
- Tous les autres flots ont pour volume $2 + \epsilon$



- But : allouer des débits aux flots pour minimiser $(C_1 + \ldots + C_5)/5$

Exemple – Optimum offline clairvoyant

- Formulation MILP indexée par le temps pour le cas clairvoyant [2]



- Le CCT moyen est $OPT = (4 + 4 \times 2)/5 = 2.4$

2.

☞ Y. Magnouche et *al.*, Branch-and-benders-cut algorithm for the weighted coflow completion time minimization problem, INOC 2022.

ORDONNANCEMENT NON-CLAIRVOYANT DE COFLOWS

BLINDFLOW

Ordonnancement non-clairvoyant de coflows – BlindFlow

- Allocation round robin (RR) au port ℓ : $r_\ell(t) = b_\ell / n_\ell(t)$
- Allocation RR généralisée :

$$r^{k,j}(t) = \min\{r_i(t), r_o(t)\} = \frac{1}{\max\{1/r_i(t), 1/r_o(t)\}}$$

  au flot en cours $j \in F_k$ utilisant les ports ingress/egress i et o.

- Allocation de débit de BlindFlow [3] : $r^{k,j}(t) = \frac{1}{1/r_i(t)+1/r_o(t)}$

Théorème

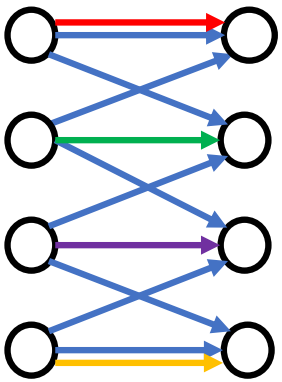
L'allocation de débit de BlindFlow est faisable et fournit une $8 \times p$*-approximation, où* $p = \max_{k \in \mathcal{C}} |F_k|$ *est le nombre maximum de flots qu'un coflow peut avoir.*

3.

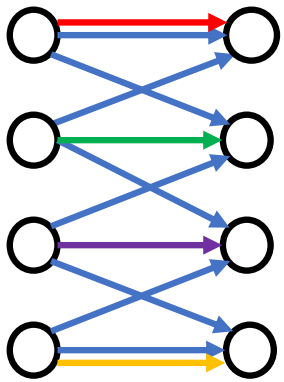
☞ A. Bhimaraju, D. Nayak and R. Vaze, Non-clairvoyant scheduling of coflows, WiOpt 2020, 2020.

Exemple – Allocation RR généralisée

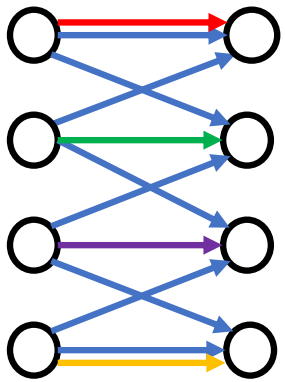
- Tous les ports ont la même bande passante normalisée de 1
- Les flots du coflow 1 ont pour volume 1, les autres ont pour volume 2



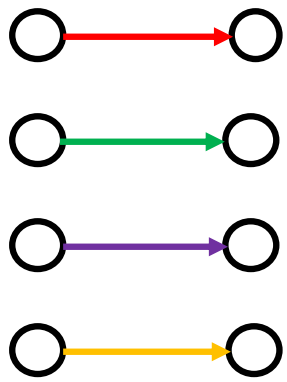
$t = 1$



$t = 2$



$t = 3$



$t = 4$

- Le CCT moyen est $(3 + 4 \times 4)/5 = 3.8 \approx 1.6 \times OPT$ $(8 \times p = 64)$

ORDONNANCEMENT CLAIRVOYANT DE COFLOWS

SINCRONIA

Ordonnancement clairvoyant de Coflows – Sincronia

- La couche transport ne peut pas réaliser une allocation de débit arbitraire aux flots
- Sincronia ordonne les coflows dans un certain ordre, puis s'appuie sur les mécanismes de priorité du réseau
  1. σ-order : le coflow $\sigma(n)$ a priorité sur le coflow $\sigma(n+1)$
  2. Allocation Greedy : un flot est bloqué ssi son port ingress/egress est occupé à servir un flot plus prioritaire
- Dans la suite, on note $p_{\ell,k}$ le CCT du coflow k au port ℓ en isolation

$$p_{\ell,k} = \sum_{j \in F_{k,\ell}} v_{k,j}/b_\ell$$

Inégalités Parallèles

- Ordre $1, 2, \ldots, n$ sur port ℓ : $C_{\ell,1} = p_{\ell,1}, \ldots, C_{\ell,k} = p_{\ell,k} + \sum_{i<k} p_{\ell,i}$
- Pour tout $S \subseteq \mathcal{C}$ et $k \in S$, on a $C_{\ell,k} \geq p_{\ell,k} + \sum_{i \in S, i<k} p_{\ell,i}$ et donc $C_{\ell,k} p_{\ell,k} \geq p_{\ell,k}^2 + p_{\ell,k} \sum_{i \in S, i<k} p_{\ell,i}$, d'où

$$\sum_{k \in S} C_{\ell,k} p_{\ell,k} \geq \sum_{k \in S} p_{\ell,k}^2 + \sum_{k \in S} \sum_{i \in S, i<k} p_{\ell,k} p_{\ell,i},$$

$$= \frac{1}{2} \sum_{k \in S} p_{\ell,k}^2 + \frac{1}{2} \left\{ \sum_{k \in S} p_{\ell,k}^2 + 2 \sum_{k \in S} \sum_{i \in S, i<k} p_{\ell,k} p_{\ell,i} \right\},$$

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

- La borne inférieure $f_\ell(S)$ ne dépend pas de l'ordre et l'inégalité est valide pour tout $S \subseteq \mathcal{C}$

σ-order de Sincronia

- Méthode pour le calcul du σ-order :

$$\text{Min} \sum_{k \in \mathcal{C}} C_k \qquad \text{(P3-Primal)}$$

$$\text{s.t} \quad \sum_{k \in S} p_{\ell,k} C_k \geq f_\ell(S), \ \ell \in \mathcal{L}, S \subseteq \mathcal{C},$$

$$C_k \geq 0, \ k \in \mathcal{C},$$

$$\text{Max} \sum_{\ell \in \mathcal{L}} \sum_{S \subseteq \mathcal{C}} f_\ell(S)\, y_{\ell,S} \qquad \text{(P3-Dual)}$$

$$\text{s.t} \quad \sum_{S: k \in S} \sum_{\ell \in \mathcal{L}} p_{\ell,k} y_{\ell,S} \leq 1, \ k \in \mathcal{C},$$

$$y_{\ell,S} \geq 0, \ \ell \in \mathcal{L}, S \subseteq \mathcal{C}.$$

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

- P3-Primal est une relaxation du problème d'ordonnancement initial

Algorithme primal-dual Sincronia

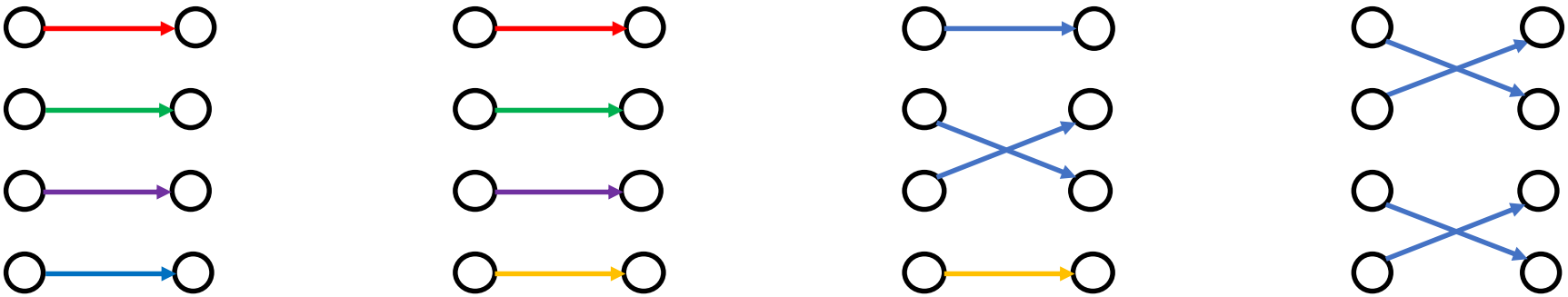
1: Initialize all dual variables $y_{\ell,S}$ to 0 and set $w_k = 1$ for all $k \in \mathcal{C}$
2: $S \leftarrow \mathcal{C}$
3: **for** $t = n \dots 1$ **do**
4: $\quad b \leftarrow \text{argmax}_{\ell \in \mathcal{L}} \sum_{k \in S} p_{\ell,k}$ ▷ Bottleneck port
5: $\quad k^* \leftarrow \text{argmin}_{k \in S} \left(\frac{w_k}{p_{b,k}} \right)$ ▷ Coflow with largest weighted proc. time
6: $\quad C_{k^*} \leftarrow \sum_{k \in S} p_{b,k}$ and $y_{b,S} \leftarrow \frac{w_{k^*}}{p_{b,k^*}}$ ▷ Set primal and dual variables
7: $\quad w_k \leftarrow w_k - w_{k^*} \frac{p_{b,k}}{p_{b,k^*}}$ for all $k \in S$ ▷ Update coflow weights
8: $\quad \sigma(t) \leftarrow k^*$ ▷ Set priority of coflow k^*
9: $\quad S \leftarrow S \setminus \{k^*\}$ ▷ Remove k^* from the set of unscheduled coflows
10: **end for**

Exemple – Sincronia

- Calcul du σ-order

t	b	$\sigma(t)$	$\{w_1, w_2, w_3, w_4, w_5\}$	S
–	–	–	$\{1, 1, 1, 1, 1\}$	$\{1, 2, 3, 4, 5\}$
5	4	5	$\{\epsilon/(2+\epsilon), 1, 1, 1, 0\}$	$\{1, 2, 3, 4\}$
4	3	1	$\{0, 1, 1, 1-\epsilon/2, 0\}$	$\{2, 3, 4\}$
3	3	4	$\{0, 1, 1, 0, 0\}$	$\{2, 3\}$
2	2	3	$\{0, 1, 0, 0, 0\}$	$\{2\}$
1	1	2	$\{0, 0, 0, 0, 0\}$	$\emptyset$

- Allocation de débit Greedy avec $\sigma = \{2, 3, 4, 1, 5\}$



$t = 1$ $t = 2$ $t = 3$ $t = 4$

- Le CCT moyen est $(4 + 3 \times 2 + 3)/5 = 2.6 \approx 1.08 \times OPT$

Admissibilité des solutions primale et duale de Sincronia

Lemme

Les solutions primale et duale produites par Sincronia sont admissibles.

Preuve : Les CCT C_k sont clairement une solution admissible de P3-Primal. Montrons que la solution duale est admissible. Soit $k \in \mathcal{C}$ et $t \in \{1, 2, \ldots, n\}$ tel que $k = \sigma(t)$. On a alors $k \in S(\tau)$ si $\tau = t, \ldots, n$ et $k \notin S(\tau)$ si $\tau \leq t - 1$.

$$\sum_{\ell \in \mathcal{L}} p_{\ell,k} \sum_{S \subseteq \mathcal{C}, k \in S} y_{\ell,S} = \sum_{\ell \in \mathcal{L}} p_{\ell,k} \sum_{\tau=t}^{n} y_{\ell,S(\tau)} = \sum_{\tau=t}^{n} p_{b(\tau),k}\, y_{b(\tau),S(\tau)},$$

$$= \sum_{\tau=t}^{n} p_{b(\tau),k} \frac{w_{\sigma(\tau)}(\tau)}{p_{b(\tau),\sigma(\tau)}}.$$

Or $w_k(\tau - 1) = w_k(\tau) - p_{b(\tau),k} \frac{w_{\sigma(\tau)}(\tau)}{p_{b(\tau),\sigma(\tau)}}$ pour $\tau = t, \ldots, n$

Admissibilité des solutions primale et duale de Sincronia

Preuve (suite) : On a ainsi

$$\begin{aligned}\sum_{\ell \in \mathcal{L}} p_{\ell,k} \sum_{S \subseteq \mathcal{C}, k \in S} y_{\ell,S} &= \sum_{\tau=t}^{n} \left(w_k(\tau) - w_k(\tau-1)\right), \\ &= w_k(n) - w_k(t-1), \\ &= 1,\end{aligned}$$

où la dernière égalité suit de $w_k(t-1) = 0$ et $w_k(n) = 1$. Comme on a $y_{\ell,S} \geq 0$ par construction, on conclut que la solution duale de Sincronia est admissible. □

Ordre sur les CCT de Sincronia

Lemme

Les CCT de Sincronia sont tels que $C_{\sigma(1)} \leq C_{\sigma(2)} \leq \ldots \leq C_{\sigma(n)}$.

Preuve : On a pour tout $1 < t \leq n$

$$
\begin{aligned}
S(t-1) \subset S(t) \implies & \sum_{k \in S(t-1)} p_{\ell,k} \leq \sum_{k \in S(t)} p_{\ell,k} \text{ pour tout } \ell \in \mathcal{L}, \\
\implies & \max_{\ell \in \mathcal{L}} \left(\sum_{k \in S(t-1)} p_{\ell,k} \right) \leq \max_{\ell \in \mathcal{L}} \left(\sum_{k \in S(t)} p_{\ell,k} \right), \\
\implies & C_{\sigma(t-1)} \leq C_{\sigma(t)}.
\end{aligned}
$$

□

Ratio d'approximation de Sincronia

Théorème

Sincronia fournit une solution primale dont le coût est au plus $2\times$ le coût optimal. Comme l'allocation Greedy est aussi 2-optimale, Sincronia garantit un CCT moyen inférieur à $4\times$ l'optimal.

Preuve :

$$\sum_{k\in\mathcal{C}} C_k = \sum_{k\in\mathcal{C}} \left(\sum_{\ell\in\mathcal{L}} p_{\ell,k} \sum_{S\subseteq\mathcal{C}, k\in S} y_{\ell,S} \right) C_k = \sum_{\ell\in\mathcal{L}} \sum_{S\subseteq\mathcal{C}} y_{\ell,S} \sum_{k\in S} p_{\ell,k} C_k,$$

$$= \sum_{t=1}^{n} y_{b(t),S(t)} \sum_{k\in S(t)} p_{b(t),k} C_k.$$

Comme $C_k \leq C_{\sigma(t)}$ pour tout $k \in S(t)$, on obtient

$$\sum_{k\in\mathcal{C}} C_k \leq \sum_{t=1}^{n} y_{b(t),S(t)} C_{\sigma(t)} \left(\sum_{k\in S(t)} p_{b(t),k} \right) = \sum_{t=1}^{n} y_{b(t),S(t)} C^2_{\sigma(t)}.$$

Ratio d'approximation de Sincronia

Preuve (suite) : Avec

$$C_{\sigma(t)}^2 = \left(\sum_{k \in S(t)} p_{b(t),k} \right)^2 \leq \left(\sum_{k \in S(t)} p_{b(t),k} \right)^2 + \sum_{k \in S(t)} \left(p_{b(t),k} \right)^2 = 2\, f_{b(t)}\left(S(t) \right),$$

on obtient

$$\sum_{k \in \mathcal{C}} C_k \leq 2 \sum_{t=1}^{n} f_{b(t)}\left(S(t) \right) y_{b(t),S(t)} \leq 2 \sum_{\ell \in \mathcal{L}} \sum_{S \subseteq \mathcal{C}} f_\ell\left(S \right) y_{\ell,S}^* \leq 2 \sum_{k \in \mathcal{C}} C_k^{OPT},$$

où la dernière inégalité suit du théorème de dualité faible. □

CONCLUSION

Conclusion

- Ordonnancement de coflows
  - ✔ Cas non clairvoyant : BlindFlow et l'allocation RR généralisée garantissent un ratio d'approximation de $8 \times p$
  - ✔ Cas clairvoyant : Sincronia garantit un ratio d'approximation de 4
- Extensions
  - ✔ Ordonnancement de coflows avec deadlines [4]
  - ✔ Placement de tâches et ordonnancement conjoints de coflows [5]

4.

☞ S.-H. Tseng and A. Tang, Coflow deadline scheduling via network- aware optimization. In Proc. Annu. Allert. Conf. Commun. Control Comput., 2018.

5.

☞ Y. Zhao et *al.*, Joint reducer placement and coflow bandwidth scheduling for computing clusters. IEEE/ACM Trans. on Networking, 29(1), 2021.