Joint Optimization of Task Placement and Routing in Minimizing Inter-DC Coflow Completion Time

Yingya Guo*‡, Zhiliang Wang†‡, Xia Yin*‡, Xingang Shi†‡,and Jianping Wu*‡
*Department of Computer Science and Technology, Tsinghua University
†Institute for Network Sciences and Cyberspace, Tsinghua University
‡Tsinghua National Laboratory for Information Science and Technology (TNLIST)
Beijing, P.R. China
Email:{guoyingya, wzl, yxia}@csnet1.cs.tsinghua.edu.cn, {shixg, jianping}@cernet.edu.cn

***Abstract*—With the rapidly growing of geo-distributed applications in the Internet, there is a huge amount of data generated in geo-distributed datacenters everyday. However, because of region privacy concerns and limitation of inter-DC WAN bandwidth, moving all the geo-distributed data to a single datacenter for centralized processing is not practical. Therefore, we intend to process the data where it generates by the big data applications and optimize the coflow routing in inter-DC WAN to improve the performance of the applications. Previous studies consider only routing or task placement optimization, which is inefficient.**

In this paper, we propose an algorithm PRO with an approximation ratio $(1 + \epsilon)$ that jointly optimize the placement of tasks and the routing of a coflow. Our proposed algorithm can efficiently reduce the coflow completion time.

I. Introduction

As with the rapid development of Internet applications, geo-distributed applications generate a huge amount of data across different regions in the inter-Datacenter Wide Area Network (Inter-DC WAN). Big data computing frameworks are deployed to analyze the large scale data in the geo-distributed inter-DC WAN. The traditional method to process the big data across multiple datacenters is to move all the data to a single datacenter and process them in a centralized manner. However, there are some limitations with this method. First, moving huge amount of data consumes a lot of inter-DC WAN bandwidth, which is an expensive and limited resource for inter-DC WAN network. Second, because of privacy and security concerns [1], it is undesirable to move the data across regions, which may expose user information and privacy. Therefore, it is more efficient and secure to process the data in a geo-distributed manner. The intermediate data which is processed locally is smaller in size and is efficient to be transferred, which can greatly reduce the response time of applications and improve user experience.

Map-reduce programming model is the core of the Hadoop, which consists of map, shuffle and reduce phases. The parallel flows between map tasks and a reduce task in the shuffle phase is called a coflow and the completion of the last flow determines the entire coflow completion time (CCT). Therefore, we intend to minimize the completion time of the last flow to reduce the completion time of the entire coflow. The scenario we study is shown in Fig.1. In our system, the map tasks are moved to each datacenter. The data after the processing of the map tasks should be transferred to the reduce tasks through inter-DC WAN. Therefore, the placement of reduce tasks and the routing of the coflows in inter-DC WAN effect the coflow completion time significantly. We need to minimize the coflow completion time to improve the performance of the big data process applications. In the previous work of CCT optimization in inter-DC WAN, they either optimize the reduce tasks placement [1] or the routing of the coflows [2], which is inefficient.

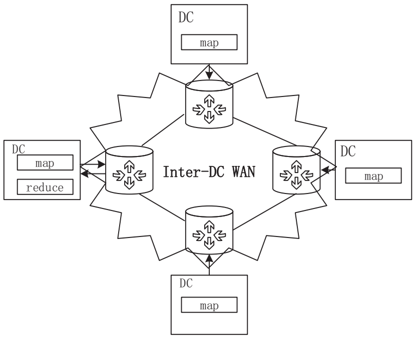

Fig. 1. Map-reduce framework in inter-DC WAN.

In this paper, we propose a framework that jointly optimizes the reduce task placement and routing of the coflows. We formulate the optimization problem as a math programming and propose a heuristic algorithm PRO (Placement and Routing Optimization) to solve it. The CCT can be greatly reduced in our proposed algorithm compared with the CCT that only optimizes the reduce task placement or the routing.

II. Network Model

The Inter-DC WAN can be modeled as a graph $G = (V, E)$, V denotes the router set in inter-DC WAN and E denotes the link set among the routers. $R_l, (l \in E)$ represents the capacity of the link l. Each data center is connected to a router in the inter-DC WAN. We denote the datacenters as $D = \{1, 2, ..d\}$. Each datacenter is assigned map tasks and some datacenters are assigned reduce tasks. $\pi = \{1, 2, ..\gamma\}$ is the set of reduce task. S_i is the set of input datacenters that transfer flows for the

978-1-5090-2991-4/17/$31.00 ©2017 IEEE

reduce task i. We should determine the placement of reduce tasks and the routing of a single coflow among the Inter-DC WAN so that coflow completion time can be minimized. Our optimization problem can be formulated as follows:

$$\textbf{minimize} \ \{y_{ij} * C_{ij}\} \tag{1}$$

$$\sum_{j=1}^{d} y_{ij} = 1, \quad \forall i \in \pi \tag{1a}$$

$$\sum_{i=1}^{\gamma} y_{ij} \leq a_j, \quad \forall j \in D \tag{1b}$$

$$C_{ij} = \textbf{max}_{k \in S_i}(m_{kj}/b_{kj}), \quad \forall i \in \pi, \forall j \in D \tag{1c}$$

$$\sum_{k \in S_i} \sum_{j \in D} \sum_{t: l \in P_{kj}^{t}} b_{kj} x_{kj}^{t} \leq R_l, \quad \forall l \in E \tag{1d}$$

$$\sum_{t} x_{kj}^{t} = 1, \quad \forall k \in S_i, j \in D \tag{1e}$$

$$x_{kj}^{t}, y_{ij} \in \{0, 1\} \tag{1f}$$

C_{ij} is the transfer time to receive all the intermediate data for reduce task i when placed at j datacenter. y_{ij} is a binary variable. $y_{ij} = 1$ denotes that the reduce task i is placed at datacenter j. Equation (1a) requires reduce task i must be placed at one and only one datacenter. Equation (1b) denotes that the number of reduce tasks that are placed in datacenter i should not exceed the maximum number of reduce tasks that can be scheduled on datacenter i. a_j is the maximum number of reduce tasks that can be scheduled on datacenter j and is a pre-set constant. m_{kj} is the volume of the intermediate data, which is transferred between datacenters k and j. b_{kj} is the assigned bandwidth of the flow between datacenters k and j. Equation (1c) computes the coflow completion time for the reduce task i placed at datacenter j. $x_{kj}^{t} = 1$ denotes the flow between datacenters k and j is routed through path t. Equation (1d) constrains that the consumed bandwidth by the coflow should not exceed the link capacity. P_{kj}^{t} is the t-th path between datacenter k and j. Equation (1e) represents that each flow chooses one and only one path between the source and destination. (1f) constrains that x_{kj}^{t}, y_{ij} are binary variables.

There are integer variables and the product of two variables in the objective function, which makes the problem a Mixed Integer NonLinear Programming (MINLP) problem. We can relax the original problem and transform the original problem into a Linear Programming (LP) problem, which can be solved by LP solvers in polynomial time.

III. Algorithm Description

We propose an approximation algorithm PRO (Placement and Routing Optimization) with an approximation ratio of $(1+\epsilon)$ to solve the optimization problem (1). In each iteration, we determine the placement of the reduce tasks that provide the minimum coflow completion time. Then we solve the transformed LP problem and round the solution to minimize the coflow completion time.

IV. Experiments Evaluation

In the simulation, we conduct our experiments on Microsoft inter-DC WAN with 5 DCs and 14 inter-DC links [3]. The intermediate data sizes in each datacenter are randomly generated. We compare our algorithm with RP (random placement and routing optimization) and SPR (placement optimization and k-shortest path routing).

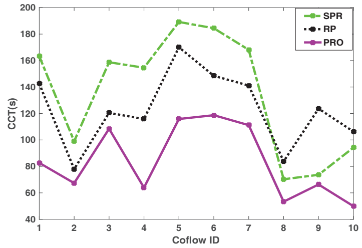

Fig. 2. Completion time for different coflows.

We plot the curves that CCT varies with different intermediate data size under different algorithms. As shown in Fig.2, we can find that under different intermediate data size, our proposed algorithm PRO obtains a lower CCT by jointly optimize the routing and placement of reduce tasks compared with RP and SRP.

V. Conclusion

In this paper, we propose an algorithm PRO to reduce the coflow completion time by jointly optimizing the task placement and coflow routing. Our proposed algorithm PRO can outperform the algorithms that only optimize the task placement or routing.

Acknowledgment

This work is supported by the National High Technology Research and Development Program of China (863 Program) No. 2015AA016105.

References

[1] Z. Hu, B. Li, and J. Luo, “Flutter: Scheduling tasks closer to data across geo-distributed datacenters,” in *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[2] H. Zhang, K. Chen, W. Bai, D. Han, C. Tian, H. Wang, H. Guan, and M. Zhang, “Guaranteeing deadlines for inter-data center transfers,” *IEEE/ACM Transactions on Networking*, 2016.

[3] C.-Y. Hong, S. Kandula, R. Mahajan, M. Zhang, V. Gill, M. Nanduri, and R. Wattenhofer, “Achieving high utilization with software-driven wan,” in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4. ACM, 2013, pp. 15–26.