# PieBridge: A Cross-DR scale Large Data Transmission Scheduling System *

Yuchao Zhang[1,3]†, Ke Xu[1,3], Guang Yao[1,2], Miao Zhang[2], Xiaohui Nie[1]
[1]Tsinghua University [2]Baidu
[3]Tsinghua National Laboratory for Information Science and Technology
zhangyc14@mails.tsinghua.edu.cn, xuke@mail.tsinghua.edu.cn
{yaoguang,zhangmiao02}@baidu.com, nxh15@mails.tsinghua.edu.cn

## ABSTRACT

Cross-DR WAN (Datacenter Region Wide Area Network) with various services are deployed to provide timely data information and analytics for users in a wide range of geographical locations. For its reliability and performance, data duplication synchronization is essential among different IDCs (Internet datacenters). However, this problem poses a challenge. First, data duplication requires huge amount of bandwidth whereas the bandwidth of cross-DR links and the upload/download rates of server interfaces are limited. Second, data transmissions are time sensitive, but the current network cannot complete such tasks in a timely manner. In this work, we present PieBridge, a cross-RD data duplicate transmission platform that accommodates hundreds of TBs of data generated from user applications online data analytics. We deployed PieBridge on the IDCs of Baidu and obtained promising performance results in comparison with the prevalent approaches.

## CCS Concepts

•**Networks** → *Network algorithms;* **Network services;**

## Keywords

Cross-DR WAN; Large-scale Data Transmission

*This work has been supported by NSFC (61472212), 863 Project of China (2015AA010203) and EU MARIE CURIE ACTIONS EVANS (PIRSES-GA-2013-610524).

†Work partly done when author is interned in Baidu.



*SIGCOMM '16, August 22–26, 2016, Florianopolis, Brazil*
 ISBN 978-1-4503-4193-6/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2934872.2959046

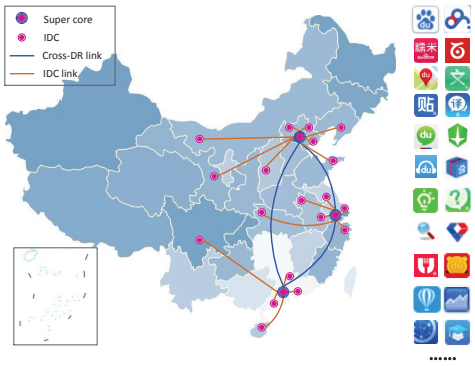



Figure 1: There are three geographically distributed *DRs*, each has a *super core* and handles numbers of *IDCs*. In each IDC, there are a series of *clusters* that consists of tens of thousands of *servers*.

## 1. INTRODUCTION

Large information platform providers, such as Microsoft [2, 4], Google [3, 5] and Baidu, provide timely data information services for end users in a wide range of geographical locations, and multiple IDCs are built for the services. Fig.1 contains IDCs distribution of Baidu that is the largest Chinese search engine in the world. However, timely duplication of large amount of data across these geographically distributed IDCs is known to be a challenge: 1) A service may have hundreds of millions of users and generate several TBs of data on daily basis. The data information is supposed to be synchronized among IDCs through links, which have limited bandwidth and cross traffic from other applications. On the other hand, upload/download server interfaces have limited data rates. 2) Transmission completion time has to be short; users can access the data only after the data synchronization transmission is completed.

In this work we present PieBridge, a centralized data transmission platform in WAN-scale. It schedules data transmission among IDCs, enhances system upload, maximizes the total data traffic, and reduces the data transmission completion time.

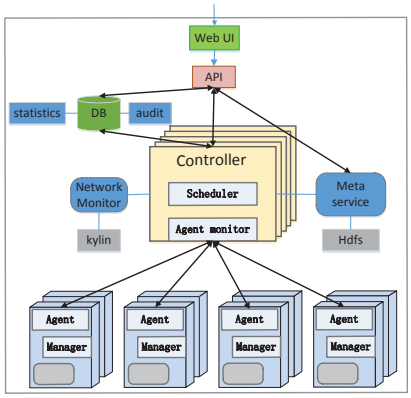



Figure 2: The architecture of PieBridge

# 2. PIEBRIDGE

PieBridge has centralized control with an efficient scheduler that selects the data transmission source for reducing the completion time of data synchronization.

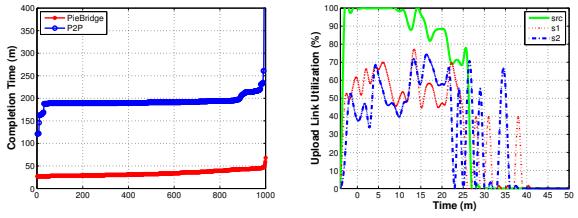
## 2.1 A Scheduling Algorithm

PieBridge scheduling algorithm contains three procedures in one data transmission period: subtask selection, max-traffic scheduling, and subtask merging. First, when a transmission task arrives at the scheduler, we first split it into subtasks to be queued. Second, we maximize the total weighted bandwidth allocation by working on the residual network, a network that keeps track of the residual capacity. We then apply the path augmentation algorithm [1] and add the amount of data of the selected subtask to the chosen path. We repeat the process on the residual network until there is no more augmenting paths. Third, at the end of a scheduling period, we merge the subtasks with the same source/destination into one subtask to cut down the calculation cost in the next scheduling period.

## 2.2 System Design

The architecture of PieBridge is shown in Fig.2 with two main components: 1) A logically centralized *controller* that accepts tasks from users and makes scheduling decision. It consists of two parts: a *scheduler* and an *agent-monitor*. The scheduler is a computation module that executes our scheduling algorithm, and the agent-monitor supports communications with agents. 2) *Agents* implement tasks at each node, control the data transmission, and report the processing status to the agent-monitor. It performs the functions of setting the upload/download rate limits, maintaining the local status information, and managing tasks.

When a user request arrives at PieBridge the controller maintains admission control, and the scheduler makes scheduling decision and informs the involved agents through the agent monitor. Upon receiving an assignment, an agent executes the scheduled data transmission.




(a) Completion time. (b) Upload link utilization.

Figure 3: The evaluation results.

## 2.3 Evaluation

We implement and evaluate PieBridge on the real topology and data traffic matrices of Baidu's WAN networks in *go language*. For a 30 Tbs data duplication, which are stored in *src* IDCs in a distributed way, there are 12 clusters and each downloads one data copy where each cluster is typically equipped with 1,000 servers. We measure PieBridge's completion time versus the most popular approach - P2P. For a particular cluster of 1,000 servers, we show the completion time in Fig.3a. Obviously, PieBridge completes the transmission 3 times faster than P2P, and eliminates the long tail phenomenon. Furthermore, Fig.3b displays the utilization of upload links of the origin source server (*src*) and two destination servers ($s1$ and $s2$), which are in different DRs. PiBridge substantially out performs P2P.

# 3. CONCLUSION

WAN-scale large data transmission is indispensable for the service reliability and cost control. We design, implement, deploy and experiment PieBridge at Baidu network with promising results. It maximizes the communication link bandwidth utilization and significantly reduces the data synchronization completion time.

# 4. REFERENCES

[1] J. Edmonds and R. M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.

[2] C.-Y. Hong, S. Kandula, R. Mahajan, et al. Achieving high utilization with software-driven wan. In *ACM SIGCOMM Computer Communication Review*.

[3] S. Jain, A. Kumar, S. Mandal, et al. B4: experience with a globally-deployed software defined wan. *Acm Sigcomm Computer Communication Review*, 43(4):3–14, 2013.

[4] S. Kandula, I. Menache, R. Schwartz, and S. R. Babbula. Calendaring for wide area networks. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 515–526. ACM, 2014.

[5] A. Verma, L. Pedrosa, M. Korupolu, and others. Large-scale cluster management at google with borg. In *Proceedings of the Tenth European Conference on Computer Systems*, page 18, 2015.