Hadoop Triggered Opt/Electrical Data-Center Orchestration Architecture for Reducing Power Consumption

Akira Yamashita, Wataru Muro, Masayuki Hirono, Takehiro Sato, Satoru Okamoto, Naoaki Yamanaka, and Malathi Veeraraghavan*

Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

** University of Virginia, 351 McCormick Road, Charlottesville,VA 22904-4743, USA*

Tel: +81-45-566-1766, e-mail: akira.yamashita@yamanaka.ics.keio.ac.jp

ABSTRACT

In this paper, a data-center network (DCN) system that distinguishes Hadoop job types and allocates optical/electrical circuits to data flows depending on the types automatically is proposed. The proposed system calculates the predicted shuffle value (PSV) of the Hadoop job and determines which type of flow is allocated by comparing the PSV and the threshold value. The PSV can be calculated from a part of the input data based on the shuffle ratio, which is known to have a relationship to the heaviness of the shuffle phase of the Hadoop job. The threshold of PSV can be dynamically changed according to the network load condition to exploit the optical circuit efficiently. By orchestrating the DCN and the Hadoop system, the proposed system achieves the reduction of power consumption. In this study, the orchestration part of the proposed DCN system is implemented and the feasibility of switching data flows between optical and electrical circuits is verified.

Keywords: orchestration, data-center network, Hadoop, optical switch, shuffle-heavy.

1. INTRODUCTION

Many of the services provided on the Internet today are dependent on data-centers (DCs). In recent DCs, tens of blade servers constitute a rack and are connected by a top of rack switch (ToR), and ToRs are connected by layer-2 or -3 switches. Because a large amount of data is exchanged between servers in different racks, power consumption in network equipment has become a significant problem. In order to mitigate this problem, DC network (DCN) architectures which introduces optical circuit switching between ToR switches has been proposed [1]-[3]. Furthermore, large scale distributed parallel processing software such as Hadoop is often used in DCs to construct server clusters. Hadoop jobs can be classified into two types: shuffle-heavy (SH) jobs and shuffle-light (SL) jobs, based on the amount of data exchanged, or shuffled, between the servers. SH jobs, ones with a large amount of shuffle data, have a large impact on power consumption. The ratio of the size of shuffle data to that of input data, called “shuffle ratio (SR)”, is highly dependent on which job is executed in a Hadoop cluster.

In this paper, we propose a new automatic flow switching system which takes into account how heavily the optical network is loaded and dynamically sets a shuffle value threshold to achieve even greater energy reduction. The shuffle value threshold can be calculated by SR and input data size. In addition, threshold value is dynamically change according to network load condition. This orchestration between network and Hadoop can be applicable future green DCN.

2. HADOOP JOB AND SHUFFLE-RATIO

Hadoop is an open source software project to enable distributed parallel processing of large data sets. It uses a programming model called MapReduce to process a large amount of data efficiently in parallel and a file system called Hadoop Distributed File System (HDFS) to ensure data redundancy.

2.1 MapReduce

Figure 1 shows a MapReduce procedure. The procedure can be divided into two main phases: the Map phase and the Reduce phase.

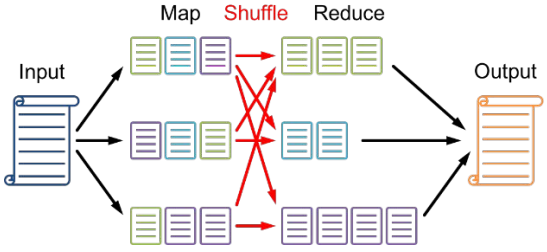

Figure 1. MapReduce model.

In the Map phase, input data is divided into smaller pieces of data. For example, if the input data is a text, it may be divided into lines or words. Then, these pieces are converted into key-value pairs. The key may be

a word, and the value the number of times it appears in the given pieces. Each server in a Hadoop cluster executes this Map procedure in parallel. Before the Reduce phase, the servers exchange these pairs with each other so that all key-value pairs with a common key will be located in one server. This phase is often referred to as the Shuffle phase. In the Reduce phase, each server simultaneously combines the values corresponding to each key it is responsible for and produces the final output data. Because each server can execute both the Map procedure and the Reduce procedure on its own in parallel, a large amount of data can be processed efficiently.

2.2 Shuffle Heaviness

In the Shuffle phase, temporary output data produced in the Map phase is exchanged between the servers. In certain jobs, a large amount is shuffled. For example, a job called TeraSort, which sorts lines of a text file, is known to produce a large amount of shuffle data. On the other hand, WordCount, which counts the number of times each word appears in a file, produces much less. The former types of jobs are called the SH jobs, and the latter the SL jobs. SH jobs contribute more to the power consumption in DCNs.

This difference can be attributed to how well the temporary output data can be compressed in the Map phase. For instance, in WordCount, whose input data is usually a collection of sentences, certain words such as “a” or “the” are likely to appear many times in the input. Because the Map procedure produces key-value pairs and the results of a repeated word can be aggregated into one key-value pair, the Map output of WordCount usually becomes small, making this job as SL. However, in TeraSort, the input data is randomly generated character strings. Therefore, it is less likely that a same key is repeated many times, and the Map output cannot be compressed very well, making this job as SH. In this way, whether a large amount of data is transferred in the Shuffle phase is largely dependent on which job is executed in the cluster.

3. PROPOSED HADOOP AUTOMATIC FLOW SWITCHING SYSTEM

In order to judge whether the job is SH or not, we found that a newly-defined parameter named “shuffle ratio (SR)” could be effective. SR is defined as the index to consider the input data size and the size of shuffled data formed by equation (1).

$$\text{Shuffle Ratio (SR)} = \frac{Shuffled\ Data\ Size}{Input\ Data\ Size} \tag{1}$$

A Hadoop job with a high load shuffle phase will have high SR. Regardless of the input data size, TeraSort (SH) shows high SR and WordCount (SL) shows small SR. According to our evaluation, SRs of TeraSort and WordCount is about 0.8 and 0.05, respectively. Similar results are likely to be obtained for other Hadoop jobs [4]. We found that the shuffle data size of a Hadoop job can be expected by examining a first small part of the input data. For example, let us assume that the input data is divided into *N* blocks (Block 1, 2, …, *N*) as shown in Fig. 2.

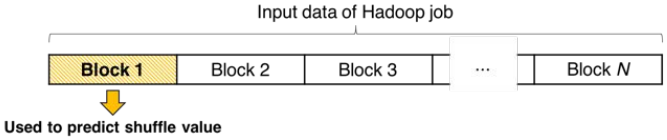

Figure 2. Example of the blocked input data.

The predicted shuffle value (PSV) of the job can be calculated as follows:

$$\text{Shuffle Ratio (Block 1)} = \frac{\text{Shuffled Data Size (Block 1)}}{\text{Input Data Size (Block 1)}} \tag{2}$$

$$\text{Predicted Shuffle Value (PSV)} = \text{Shuffle Ration (Block 1)} \times \text{Input Data Size (Whole)} \tag{3}$$

Therefore, the job category can be determined on-line by obtaining PSV from the first small input data block and setting the threshold properly.

Figure 3 shows the proposed job adaptive DCN system. There are multiple Hadoop clusters in most DCs, and jobs are executed simultaneously in each cluster. In the proposed system, multiple paths are configured by assigning different VLAN IDs to each cluster in layer-2 switching case. VLAN paths are assigned to determine the flow. Each path is set on electrical switching network (ESN) or optical switching network (OSN). The Hadoop job execution works in the following procedure. When the Cluster Manager detects the start of job, initially it sets flow path on ESN to connect servers in the cluster via the Network Manager. The Cluster Manager gives the network connection status in the cluster to the Traffic Monitor. The Traffic Monitor checks the flow in the cluster immediately and notifies the Cluster Manager periodically. The Cluster Manager connects to the cluster periodically and grasps the progress of the job. If the job has progressed from Map phase to Shuffle and Reduce phases, SH job judgment is executed by on-line job analysis. Here, PSV is determined by using the job progress obtained from the cluster and the size of shuffle data acquired from the Traffic Monitor. A Job is judged as SH when the PSV is larger than the threshold. After the judgment, a proper flow path configuration is applied through the Network Manager. Detecting the job completion, the Cluster Manager immediately stops traffic monitoring for the corresponding cluster and sets the flow path configuration to the initial state through the Network Manager. The Cluster Manager continues to monitor the cluster even after the job is over.

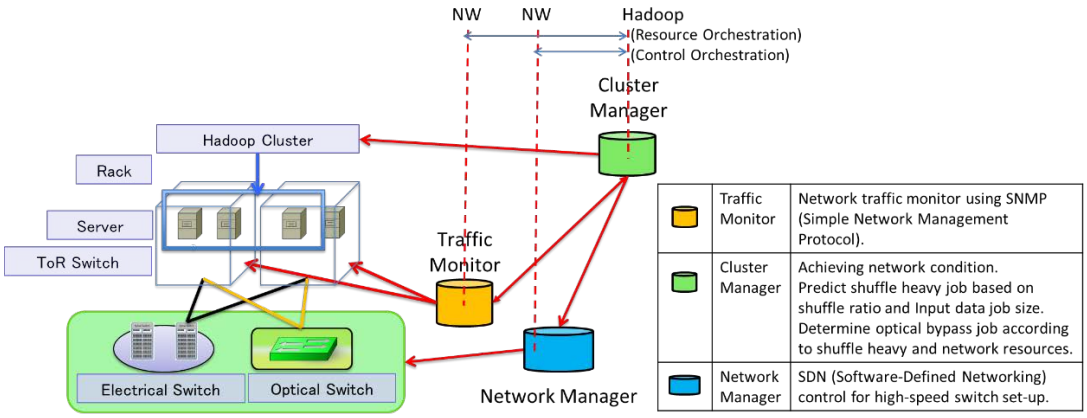

Figure 3. Job adaptive data-center network system.

4. ORCHESTRATION BETWEEN NETWORK AND HADOOP

Optical switch needs smaller power desperation than electrical switch. So, if network load is low, not only SH jobs but also shuffle-middle (SM) or SL jobs must try to use OSN instead of ESN to reduce energy. Therefore, we employ dynamic PSV threshold based on network and Hadoop orchestrator.

Figure 4 shows flowchart of dynamic PSV threshold method. Network load is monitored by the Traffic Monitor. The Cluster Manager changes threshold according to the network load. The Network Manager controls OSN to bypass traffic from ESN to OSN. In Fig. 4, an orchestration curve which is relation between threshold (Th) and network load is also described. If network is high-load, only SH job will use OSN. On the other hands, if the networks load is low, SH, SM, and also SL jobs can use OSN. Figure 5 shows number of jobs and SR distributions under heavy and low networks load. PSV threshold (Th) is determined which job will try to use OSN. As shown in Fig. 4, threshold is dynamically changed. This is realized by orchestration between network and Hadoop.

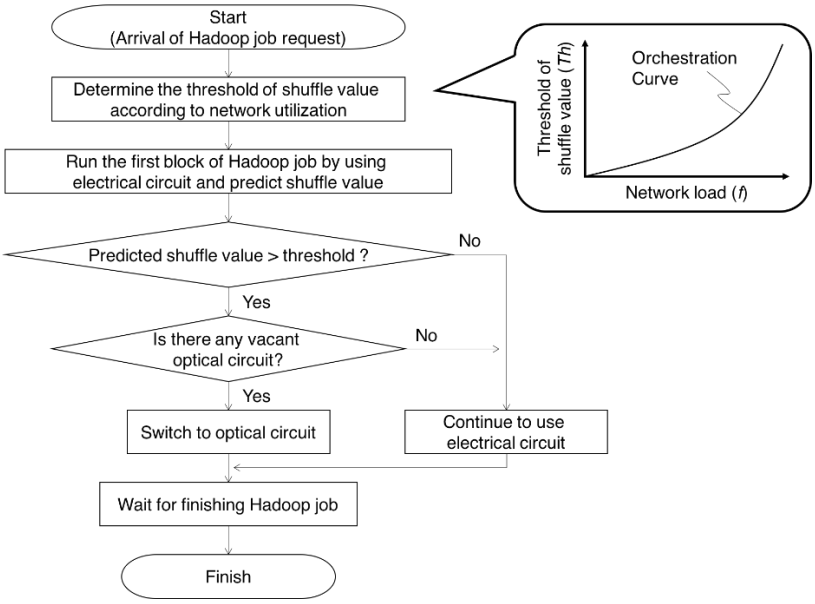

Figure 4. Flowchart of the dynamic shuffle threshold method.

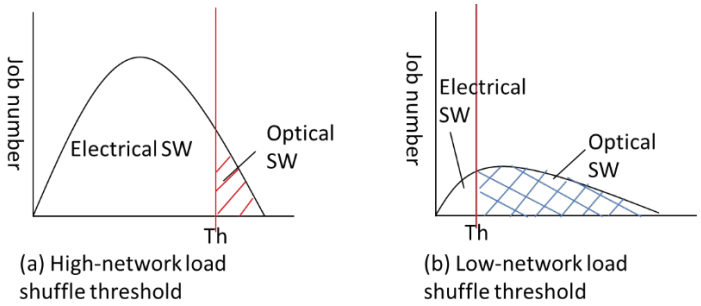

Figure 5. Job classification under high-network load condition and low-network load condition.

5. PROTOTYPE IMPLEMENTATION

A small Hadoop cluster is implemented on PC servers. ToR switches and an ESN are constructed by 10 Gbps Ethernet Switches. MEMS (Micro-Electronics Mechanical Switch) optical switch is used for OSN. TeraSort and WordCount are operated. Traffic flow is set on ESN or OSN. Input file size of both jobs is 30 GB. Figure 7 shows the shuffle data amount in the optical/electric flow paths when the Hadoop cluster executes each job of TeraSort (Fig. 7a) and WordCount (Fig. 7b). The horizontal axis shows the elapsed time from the job start time. The vertical axis on the left side shows the shuffle data size on the optical/electrical flow paths and on the right side shows the progress of the job phases. In TeraSort, in the first 330 s, little data is transferred because only the Map task is executed. As the number of completed Map tasks increases, intermediate data that can be transferred is created and data transfer is started. 450 MB of the first shuffle data is transferred on ESN. After 20 s, the second shuffle data is transferred on OSN. Finally, the amount of data transferred on OSN increases and almost data is not transferred with on ESN. This result means that the application triggered automatic flow switching between ESN/OSN has been properly applied through the on-line job analysis. In WordCount case, 50 MB of the first shuffle data is transferred on ESN and all shuffle data is transferred on ESN.

According to our approximated evaluation, TeraSort which is SH, the proposed system sets the optical flow path during the job execution and nearly 72 J (about 86%) of power consumption is reduced compared with executing the job only with the electrical flow path.

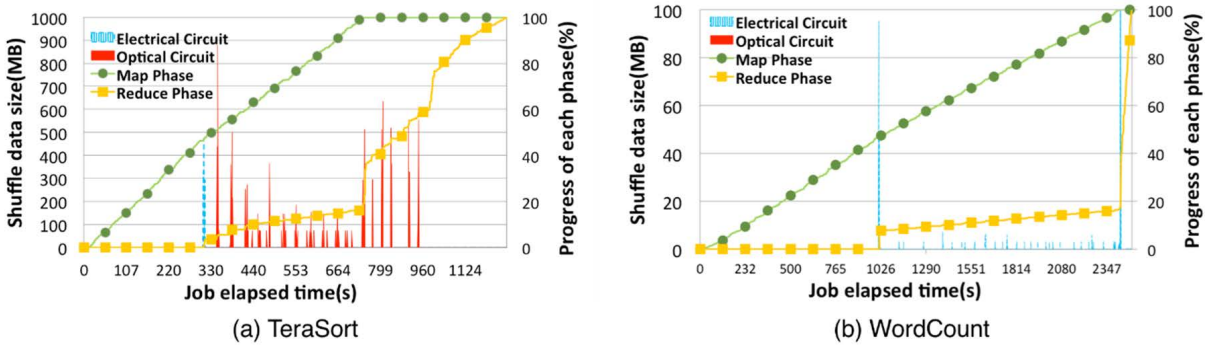

Figure 7. Shuffle data size of each circuit being executed: (a) TeraSort, (b) WordCount.

6. CONCUSION

To reduce power consumption of the data-center networks, introduction of optical circuit switching network is very attractive. In this paper, data-center application triggered automatic flow switching in optical/electrical hybrid data-center network has been proposed. The investigated parameter “shuffle-ratio (SR)” was introduced to the on-line Hadoop job analysis system. The proposed automatic flow switching system can dynamically sets an optical/electrical flow path within a cluster according to Hadoop job traffic condition and network load. The prototype system has been implemented and evaluated.

ACKNOWLEDGEMENTS

This work is supported by ”HOLST (High-speed Optical Layer 1 Switch system for Time slot switching based optical data center networks) Project” funded by New Energy and Industrial Technology Development Organization (NEDO) of Japan.

REFERENCES

[1] N. Farrington *et al.*: Helios: A hybrid electrical/optical switch architecture for modular data-centers, in *Proc. ACM SIGCOMM 2010*, New York, NY, USA, pp. 339-350.

[2] N. Farrington *et al.*: A multiport microsecond optical circuit switch for data center networking, *IEEE Photonics Technology Letters*, vol. 25, no. 16, pp. 1589-1592, Aug.15, 2013.

[3] M. Hirono *et al*: HOLST: Architecture design of energy-efficient data center network based on ultra high-speed optical switch, to appear in *Proc. IEEE LANMAN 2017*, Osaka, Japan, Jun. 2017.

[4] X. Wang *et al*: Dynamic optical circuits in datacenter networks for shuffle-heavy Hadoop applications, in *Proc. iPOP 2016*, Yokohama, Japan, Jun. 2016, paper T4-1.