This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

An Exploration of Designing a Hybrid Scale-Up/Out Hadoop Architecture Based on Performance Measurements

Zhuozhao Li, Haiying Shen, *Senior Member, IEEE,* Walter Ligon, and Jeffrey Denton

Abstract—Scale-up machines perform better for jobs with small and median (KB, MB) data sizes, while scale-out machines perform better for jobs with large (GB, TB) data size. Since a workload usually consists of jobs with different data size levels, we propose building a hybrid Hadoop architecture that includes both scale-up and scale-out machines, which however is not trivial. The first challenge is workload data storage. Thousands of small data size jobs in a workload may overload the limited local disks of scale-up machines. Jobs from scale-up and scale-out machines may both request the same set of data, which leads to data transmission between the machines. The second challenge is to automatically schedule jobs to either scale-up or scale-out cluster to achieve the best performance. We conduct a thorough performance measurement of different applications on scale-up and scale-out clusters, configured with Hadoop Distributed File System (HDFS) and a remote file system (i.e., OFS), respectively. We find that using OFS rather than HDFS can solve the data storage challenge. Also, we identify the factors that determine the performance differences on the scale-up and scale-out clusters and their cross points to make the choice. Accordingly, we design and implement the hybrid scale-up/out Hadoop architecture. Our trace-driven experimental results show that our hybrid architecture outperforms both the traditional Hadoop architecture with HDFS and with OFS in terms of job completion time, throughput and job failure rate.

Index Terms—Hadoop; scale-up; scale-out; remote file system; hybrid architecture;

1 Introduction

MapReduce [17] is a framework designed to process a large amount of data in the parallel and distributed manner on a cluster of computing nodes. Hadoop, as a popular open source implementation of MapReduce, has been deployed in many large companies such as Facebook, Google and Yahoo!. In the last decade, the amount of computation and data increases exponentially [9]. This trend poses a formidable challenge of high performance on MapReduce and motivates many researchers to explore to improve the performance from different aspects such as job scheduling [3, 5, 21, 22], short jobs performance optimization [19] and intermediate data shuffling [7, 15, 16, 34].

A common sense in the IT community is that a larger Hadoop cluster of machines is always better for processing big data, i.e., a very large volume of data in terabytes, petabytes or exabytes. Recent studies indicate that most jobs in the production workloads (e.g., at Facebook [14] and Microsoft [31] clusters) usually have input/shuffle/output sizes in the MB to GB range. These production clusters with many small jobs suffer from poor performance because the existing Hadoop MapReduce clusters were not originally designed for short and latency-sensitive jobs [19]. Therefore, optimizing the performance of short jobs is important. To process the current production workloads, *Appuswamy et al.* [10] claimed that scale-up machines is a better option than scale-out machines. Scale-up is vertical scaling, which means adding more resources to the nodes of a system, typically stronger processors and RAM. Scale-out is horizontal scaling, which refers to adding more nodes with few processors and RAM to a system [10].

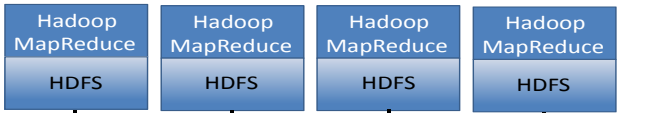


Fig. 1: Typical Hadoop with HDFS local storage.

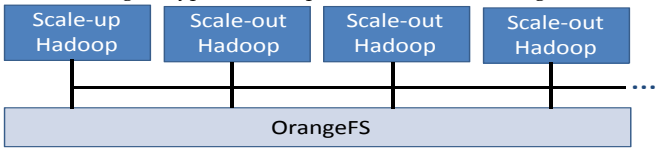


Fig. 2: Hybrid scale-up/out Hadoop with a remote storage.

A real-world workload usually consists many jobs handling diverse data size levels and computations. Also, in this big data era, the data size handled by jobs has been increasingly larger. We calculated the Cumulative Distribution Function (CDF) of the data sizes of more than 6000 jobs in the Facebook synthesized workload trace FB-2009 [13] and show the results in Figure 3. We see that the input data size ranges from KB to TB. Specifically, 40% of the jobs process less than 1MB small datasets, 49% of the jobs process 1MB to 30GB median datasets, and the rest 11% of the jobs process more than 30GB large datasets. Such a workload requires both scale-up and scale-out machines to handle datasets with different size levels. Therefore, it is not

- *Haiying Shen is the corresponding author.*
- *Z. Li, H. Shen, and W. Ligon are with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29631.*
- *J. Denton is with the CITI group of CCIT as the Deputy Director of Data Science and system administrator of Palmetto at Clemson University.*
- *E-mail: {zhuozhl, shenh, walt, denton}@clemson.edu*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

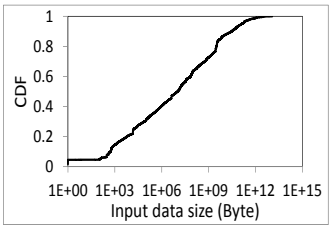


Fig. 3: CDF of input data size in the Facebook synthesized trace.

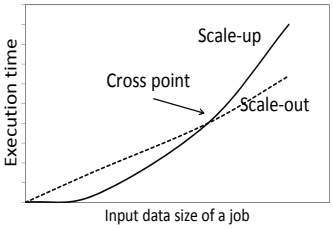


Fig. 4: Cross point.

practical to simply decide to use scale-up machines or scale-out machines for a workload.

Motivated by these observations, we see the great potential to design a cost-efficient Hadoop architecture with coexistence of both scale-up and scale-out machines to improve the performance of real-world workloads. That is, we aim to improve the cluster performance without spending more money. In such a hybrid cluster, we utilize the features of both scale-up machines and scale-out machines. Scale-up machines have more powerful CPU and memory but fewer CPU cores, and hence are better for small jobs but not large jobs. On the other hand, scale-out machines have more CPU cores, and hence it is more beneficial to process a larger amount of data than scale-up machines.

However, designing a hybrid cluster is non-trivial due to two main challenges.

• Proper data storage to enable both scale-up machines and scale-out machines efficiently access data needed. The Facebook trace shows that thousands of jobs (85%) handle small or median data size levels. Since majority jobs are better to run in scale-up machines, these jobs may overload the local disks of scale-up machines. Also, both scale-up and scale-out machines may request the same workload dataset, which leads to data transmission between the machines and may degrade the application performance.

• Adaptively scheduling a job to either scale-up cluster or scale-out cluster that benefits the job the most. Simply referring to job input data size may not be sufficient and there may exist other factors that determine the performance difference between scale-up and scale-out clusters.

In this paper, we aim to investigate the feasibility of designing such a hybrid architecture and start the initial exploration. Hadoop Distributed File System (HDFS) is the file system designed to closely work with Hadoop MapReduce (Figure 1). To handle the first challenge, we could let HDFS consider both scale-out and scale-up machines equally as datanodes for data distribution. However, in this case, scale-up machines must frequently access data in scale-out machines, which not only degrades their job performance but also consumes their bandwidth. We then explore handling this challenge by using a remote dedicated storage system (e.g., OrangeFS [6] (Figure 2). The dedicated storage offloads I/O load from the compute nodes and enables the data sharing between scale-up machines and scale-out machines easily. To handle the second challenge, we need to decide the *cross points* of different jobs. As shown in Figure 4, *cross point* is the size of input data for a job, at which the scale-up and scale-out clusters provide similar performance for the job, and if the actual size is higher or lower than the cross point, one cluster provides better performance. We found that the cross point of a job is determined by multiple factors, such as shuffle/input data size ratio and output data size.

To this end, by taking advantage of the Clemson University Palmetto HPC cluster that successfully configured Hadoop by replacing the local HDFS with the remote OrangeFS (OFS), we configured four architectures as shown in Table 1: scale-up machines with OFS (denoted by up-OFS), scale-up machines with HDFS (denoted by up-HDFS), scale-out machines with OFS (denoted by out-OFS), and scale-out machines with HDFS (denoted by out-HDFS). We then measure the performance of representative Hadoop applications (i.e., shuffle-intensive and map-intensive) on these architectures. Through the measurement, we aim to see if the use of a remote file system can provide efficient data storage as we expected and whether it brings about any side-effect to scale-up cluster or scale-out cluster. More importantly, we study the benefits gained from scale-up cluster and scale-out cluster, respectively, for different jobs, based on which we can decide where to run a given job.

Through our performance measurement, we confirm the benefits of the remote file system, identify the factors (i.e., shuffle/input ratio, and output data size) that determine the performance differences on the scale-up and scale-out clusters and their cross points to make the choice. Accordingly, we design a hybrid scale-up/out Hadoop architecture. In this architecture, different jobs in a workload can be executed on either scale-up or scale-out cluster that benefits them the most, thus achieving higher workload performance. In this paper, we use execution time to evaluate application performance. Our contributions are summarized below:

1. We have identified comparable scale-up cluster and scale-out cluster, built four architectures shown in Table 1 and optimized their configurations to achieve the best performance by trial of experiments.
2. We have conducted thorough experiments of different applications on the four architectures with different input data sizes and provided an insightful analysis on the performance.
3. Based on our measurement analysis, we design a scheduler, which helps decide whether to execute a job on the scale-up or scale-out cluster to gain the most benefit. We then design a hybrid scale-up/out Hadoop architecture that incorporates this scheduler and uses a remote file system.
4. We have conducted experiments driven by the Facebook synthesized workload trace, which show that our hybrid architecture outperforms both the traditional Hadoop architecture with HDFS and with OFS in terms of the execution time, throughput and success rate of jobs.

As far as we know, our work is the first that i) studies the application performance on the four architectures in Table 1, ii) proposes the idea of a hybrid scale-up/out Hadoop architecture to better serve a real-world workload with jobs handling diverse data size levels and computations, and iii) introduces a method to build the hybrid scale-up/out Hadoop architecture. Our new architecture is only an initial design and has many aspects to improve but we expect it can stimulate many researches on this topic.

The remainder of this paper is organized as follows. Section 2 describes the configurations of scale-up and scale-out machines for the HPC-based Hadoop. Section 3 presents the performance measurements for different types of applications on the four architectures. Section 4 presents our pro-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

TABLE 1: Four architectures in our measurement.

	Scale-up	Scale-out
OFS	up-OFS	out-OFS
HDFS	up-HDFS	out-HDFS

posed hybrid scale-up/out Hadoop architecture. Section 5 presents the trace-driven experimental results of our architecture compared with the traditional Hadoop. Section 6 gives an overview of the related work. Section 7 concludes the paper with remarks on our future work.

2 Optimization of the HPC-Based Hadoop MapReduce Configurations

In this section, we introduce the details on how we configured Hadoop MapReduce on the Clemson Palmetto HPC cluster, which is ranked as the top five fastest supercomputers at public universities in United States and the 66^{th} fastest supercomputers globally [4]. The HPC cluster makes it easy to build the hybrid scale-up/out Hadoop architecture due to two reasons. First, a HPC center have different kinds of machines with different number of CPU cores and RAM size, which allows us to build the architecture without any further cost to buy new machines.

Second, the configuration of Hadoop with a remote storage makes the coexistence of scale-up and scale-out machines in Hadoop possible. These machines can share the same datasets and access their required data easily without frequent data transmission between machines.

2.1 Introduction of Hadoop MapReduce

MapReduce is a framework that processes a large dataset in parallel using a large number of nodes. HDFS is the file system designed to closely work with Hadoop MapReduce [2]. HDFS generally consists of a namenode that manages the metadata of the cluster and multiple datanodes used to store the data blocks. The running process of a MapReduce job is composed of three phases: map, shuffle and reduce. Each node has a specific number of map and reduce slots. Given the input data, HDFS divides it to $\frac{input\ data\ size}{block\ size}$ number of data blocks and stores the blocks into datanodes. In the map phase, the job tracker assigns each mapper to process one data block in a datanode. The output of all the mappers is intermediate data (i.e., shuffle data). In the shuffle phase, the shuffle data is then partitioned and shuffled to corresponding reduce nodes. The shuffle data is copied to the reduce nodes' memory first. If the shuffle data size is larger than the size of in-memory buffer (which is determined by the heap size), the shuffle data will be spilled to local disk. In the reduce phase, the reducers aggregate the shuffle data and generate the final output.

2.2 Hadoop MapReduce on HPC Cluster

We use myHadoop [25] to automatically configure Hadoop on the Clemson Palmetto HPC cluster. Recently, a Java Native Interface (JNI) shim layer has been implemented on the Clemson Palmetto HPC cluster that allows Hadoop MapReduce to store input/output on a remote storage file system (i.e., OFS) directly. OFS is a parallel file system that distributes data across multiple servers. The remote storage in general has much faster I/O performance than local disks [1]. Moreover, because of the centralization of remote storage, it is much easier to manage and maintain than the local storage. With OFS, no modifications to the Hadoop source code and MapReduce jobs are required.

2.3 Experiment Environment

In the experiments, we use Hadoop version 1.2.1. We use two machines for scale-up Hadoop MapReduce. Each scale-up machine is equipped with four 6-core 2.66GHZ Intel Xeon 7542 processors, 505GB RAM, 91GB hard disk. On Clemson Palmetto HPC center, the remote file system connects to the computing nodes with 10GB Myrinet. In order to maintain a fair comparison between Hadoop with HDFS and Hadoop with OFS architecutres, we select 10GB Myrinet as the interconnection between the nodes. Additionally, we do not aim to measure the absolution in the performance measurement. We aim to compare the performance of the four architectures in the comparable environments and propose a cross point determination model based on the measurement results. Finally, although Ethernet is the most common used interconnection for Hadoop systems, it is showed that in HPC architecture, the high speed interconnection Myrinet and Infiniband can also provide competitive performance [25]. The scale-out cluster consists of twelve machines, each of which has two 4-core 2.3GHZ AMD Opteron 2356 processors, 16GB RAM, 193GB hard disk, and 10Gbps Myrinet interconnections.

The reason that we select two scale-up machines and twelve scale-out machines is because it makes the scale-up and scale-out clusters have the same price cost (according to the investigation of market), thus makes the performance measurements comparable. Previous research [10] used only one scale-up machine and hence did not consider the network performance between scale-up machines in the performance study. In order not to exclude the network factor in the performance study, we use two scale-up machines and comparably 12 scale-out machines.

We compare the performance of different types of applications on four architectures in Table 1. For the HDFS configuration, if one of the machines acts as both namenode and datanode, it will degrade the performance of Hadoop. Since OFS itself has metadata servers, in order to achieve fair comparison, we use an additional machine to serve as namenode in HDFS.

2.4 Hadoop Configurations

It is important for us to optimize the configurations of both scale-up and scale-out machines, either with OFS or HDFS, to achieve the best application performance of the four architectures.

Heap size In Hadoop, each map and reduce task runs in a JVM. The heap size is the memory allocated to each JVM for buffering data. If the memory is full, the data in memory is spilled to the local disk, which introduces overhead. By default, the heap size is 200MB for each JVM. Current modern servers (regardless of scale-out or scale-up machines) always provide sufficient memory for us to increase the heap size to reduce the overhead and improve the JVM performance. However, if the heap size is too large, the memory used for heap is wasted and the out of memory error may occur. To achieve the best performance and also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

avoid the out of memory error [10] in our experiments, through trial and error, we set the heap size to 8GB per task on scale-up machines, and to 1.5GB and 1GB for shuffle-intensive and map-intensive applications on scale-out machines, respectively.

Map and reduce slots To ensure the best performance, the total number of map and reduce slots is set to the number of cores for both scale-up and scale-out machines. For example, if we use machines with 4-core CPUs, the sum of map and reduce slots is equal to 4. Therefore, in our experiments, each scale-up machine has 24 map and reduce slots, while each scale-out machine has 8 map and reduce slots in total.

Remote file system strip size In HDFS, a file is broken into small blocks and each data block is processed by one map task. It is important to set the block size properly, which cannot be too small or too large. We set the HDFS block size to 128MB to match the setting in the current industry clusters. Similarly, OFS stores data in simple stripes (i.e., similar as blocks in HDFS) across multiple storage servers in order to facilitate parallel access. The stripe size is 4KB in default. In order to compare OFS fairly with HDFS, we also set the stripe size to 128MB.

The number of remote storage servers There are 32 remote storage servers in OFS in the Clemson HPC cluster. These 32 remote storage servers are connected with high-speed interconnection Myrinet, which is a high-speed local area network and has much lower protocol overhead than standard Ethernet. Currently, these 32 remote storage servers are sufficient to provide low latency for around 2000 machines on Palmetto and hence we do not need to worry that this remote file system is not scalable to large Hadoop MapReduce clusters. In our experiments, we use 8 remote servers to store each file in parallel.

RAM drive of scale-up machines The scale-up machines provide a large amount of memory size (i.e., 505GB in the experiments), which is an advantage of the scale-up machines. Even though we set the heap size to 8GB, there is much memory space left. To fully take advantage of the unused memory in the scale-up machines, Palmetto enables to use half of the total memory size as *tmpfs*, which serves the same functions as RAMdisk. On the other hand, since the memory size is limited on the scale-out machines (i.e., 16GB), we do not use memory as RAMdisk.

Shuffle data placement Although Clemson Palmetto HPC cluster allows us to configure Hadoop with remote file system OFS, we only can place input and output data to OFS, but cannot place shuffle data to OFS. We need to utilize the local file system to store the shuffle data. For the scale-up machines in our experiments (with HDFS and OFS), we place the shuffle data on RAMdisks, which improves the shuffle data I/O performance. For scale-out machines, we store the shuffle data in the local disks (i.e., HDD).

3 Performance Measurement

In this section, we compare the performance of shuffle-intensive and map-intensive jobs on the four architectures in Table 1. Shuffle-intensive applications have large shuffle data size, while map-intensive applications generally do not contain large shuffle data size. We expect to provide a thorough analysis on how different applications benefit from scale-up and scale-out clusters, with remote and local storage respectively.

3.1 Types of Applications

The applications we use include *Wordcount*, *Grep*, *Terasort*, the write test and read test of *TestDFSIO*. Among them, *Wordcount*, *Grep* and *Terasort* are typical shuffle-intensive applications. Specifically, *Wordcount* and *Grep* have only relatively large input and shuffle size but small output size, while *Terasort* has relatively large input, shuffle and output size. We generated the input data by BigDataBench [33] based on the Wikipedia datasets for *Wordcount*, *Grep*, and *Terasort*. The write test and read test of *TestDFSIO* are typical map-intensive applications. We measure the following metrics of each job:

- *Execution time*, which is the job running time and calculated by the job ending time minus job starting time.
- *Map phase duration* calculated by the last map task's ending time minus the first map task's starting time.
- *Shuffle phase duration* calculated by the last shuffle task's ending time minus the last map task's ending time.
- *Reduce phase duration*, which is the time elapsed from the ending time of the last shuffle task to the end of the job.

Note that due to the limitation of local disk size, up-HDFS cannot process the jobs with input/output data size greater than 80GB.

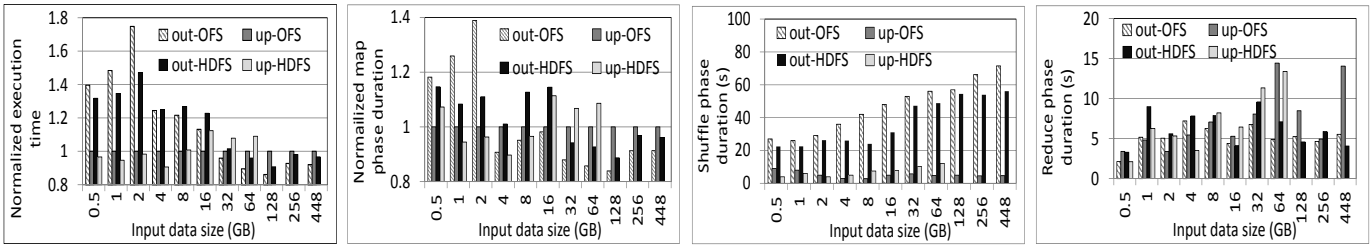
Since the execution time and map phase duration of jobs with different input data sizes differs greatly, it is difficult to see the experimental results of small data sizes in the figures. Therefore, we normalize execution time and map phase duration results by the results of up-OFS, since we only focus on the performance comparison among up-OFS, up-HDFS, out-OFS, and out-HDFS, rather than the exact execution time or map phase duration. For example, if a job running on up-OFS and up-HDFS has an execution time of 10 and 20 seconds, respectively, then up-OFS on the figure is shown as 1, while up-HDFS on the figure is shown as 2. And we only need to know from the figure that up-OFS has better performance than up-HDFS.

3.2 Performance of Shuffle-Intensive Applications

In this section, we show the performance evaluation of shuffle-intensive applications: *Wordcount*, *Grep* and *Terasort*. As we mentioned above, all these three applications include a large amount of shuffle data.

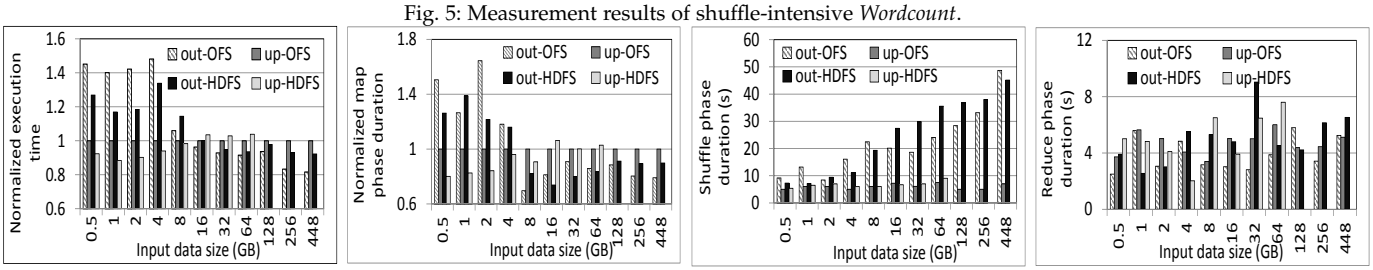
Figures 5(a), 6(a), and 7(a) show the execution time of *Wordcount*, *Grep* and *Terasort* versus different input data sizes, respectively. We see that when the input data size is small (0.5-8GB), the performance of *Wordcount*, *Grep* and *Terasort* all follows: up-HDFS>up-OFS>out-HDFS>out-OFS. Recall that the number of required map slots equals $\lceil \frac{input\ data\ size}{block\ size} \rceil$. The scale-up machines have better performance when the input data size is small because of three reasons. First, the two scale-up machines can provide the majority of required map slots of the small jobs (defined as jobs with small input data size), which means that the jobs can be completed in only a few task waves. The number of *map (reduce) waves* of a job is calculated by the number of distinct start times from all mappers (reducers) of the job. Thus, small jobs can benefit more from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems



(a) Execution time. (b) Map phase duration. (c) Shuffle phase duration. (d) Reduce phase duration.

Fig. 5: Measurement results of shuffle-intensive *Wordcount*.



(a) Execution time. (b) Map phase duration. (c) Shuffle phase duration. (d) Reduce phase duration.

Fig. 6: Measurement results of shuffle-intensive *Grep*.

more powerful CPU resources of the scale-up machines than from scale-out machines. Second, these jobs are all shuffle-intensive and their performance is very related to the memory resource. The map outputs are copied to the reduce nodes' memory, which is limited by the heap size. A larger heap size makes it less likely to spill the map outputs to the local disks. The more memory resource of scale-up machines provides larger heap size and hence enhances the performance of these shuffle-intensive applications. Third, for the shuffle data placement, the RAMdisks in the scale-up machines are much faster than the local disks in the scale-out machines.

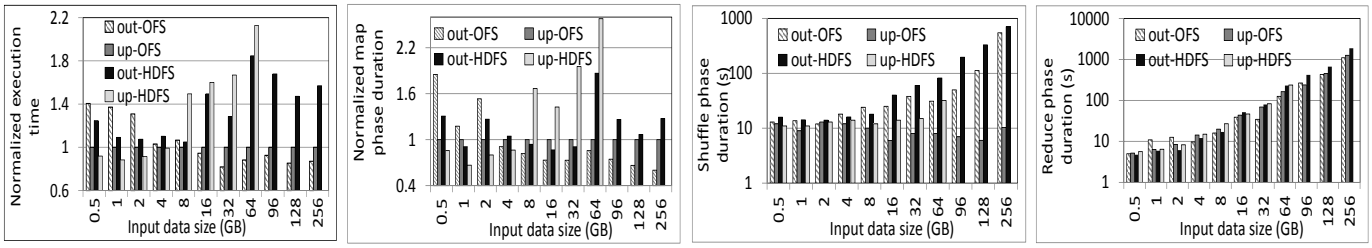
When the input data size is small, the performance of out-HDFS is around 20% (calculated by $\frac{OFS-HDFS}{HDFS}$) better than out-OFS, and up-HDFS is around 10% better than up-OFS. Although a remote file system has better I/O performance than HDFS [1], its advantage cannot be shown for small jobs. This is caused by the network latency in the communication with the remote file storage, which is independent on the data size. When the data size is small, the execution time is also small and the network latency occupies a relatively high portion of the total execution time. Then, the performance of the remote file system becomes slightly worse than the local file system. However, we see that up-OFS performs around 10-25% better than out-HDFS, which means that scale-up Hadoop with remote file system outperforms the traditional scale-out Hadoop with HDFS.

We also see from the figures that when the input data size is large (>16GB), the performance of *Wordcount* and *Grep* follows out-OFS>out-HDFS>up-OFS>up-HDFS, while *Terasort* follows out-OFS>up-OFS>out-HDFS>up-HDFS. It means that scale-out machines are better for the shuffle-intensive jobs with large input data size (i.e., large jobs) than scale-up machines. The reason is that a large input data size usually requires a large number of map slots, which however is the primary bottleneck of scale-up machines though they have more powerful CPU resources. The requirements of more map slots and less task waves of large jobs make them benefit more from the scale-out machines. OFS performs better than HDFS because the more powerful dedicated remote servers in OFS and the high speed HPC interconnections (i.e., 10Gbps Myrinet) can provide a higher I/O performance than HDFS. *Terasort* performs different from *Wordcount* and *Grep* on up-OFS and out-HDFS since the sorting program not only has relatively large shuffle data size but also large output data size, while *Wordcount* and *Grep* have a negligible output data size compared to *Terasort*. It means that with OFS, *Terasort* reads input data for map tasks from OFS and writes the output of reduce tasks to OFS, while *Wordcount* and *Grep* only take advantage of OFS during reading input data for map tasks. Therefore, *Terasort* benefits twice from the higher I/O rate of OFS, which results in its better performance on up-OFS than out-HDFS.

Furthermore, we see from the figures that as the input data size increases, the performance on scale-up machines decreases while the performance on scale-out machines increases. The three jobs have different performance degrading speed on scale-up machines as the input data size increases though they are all shuffle-intensive applications. The cross points of input data size of *Wordcount*, *Grep* and *Terasort* are close to 32GB, 16GB and 16GB, respectively. That is, when the input data size of a *Wordcount* (or *Grep*, *Terasort*) job is smaller than 32GB (or 16GB), then it performs better in scale-up machines, otherwise, it performs better in scale-out machines. This cross point difference between the jobs is caused by the different shuffle/input ratio calculated by $\frac{shuffle\ data\ size}{input\ data\ size}$. Given the same input data size, if a job's shuffle data size is large, it can benefit more from the large memory and fast RAMdisk of the scale-up machines in the shuffle phrase, thus reduces the shuffle phase duration. In our experiments, regardless of the input data size of the jobs, the shuffle/input ratio of *Wordcount*, *Terasort* and *Grep* are always around 1.6, 1 and 0.4, respectively. Therefore, the larger shuffle data size of *Wordcount* leads to slower application performance degradation on the scale-up machines when the input data size increases and a larger cross point than other applications.

Since the execution time is determined by the durations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems



(a) Execution time. (b) Map phase duration. (c) Shuffle phase duration. (d) Reduce phase duration.

Fig. 7: Measurement results of shuffle-intensive *Terasort*.

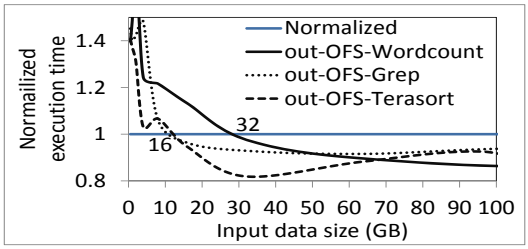


Fig. 8: Cross points of *Wordcount*, *Grep* and *Terasort*.

in the map, shuffle and reduce phases, we then study these broken-down durations. Figures 5(b), 6(b) and 7(b) show the map phase duration of *Wordcount*, *Grep* and *Terasort*, respectively. We see that the map phase duration of these jobs has similar performance trends and order as the job execution time. When the input data size is small (0.5-8GB), the map phase duration is shorter on scale-up than on scale-out; when the input data size is large (>16GB), the map phase duration is shorter on scale-out than on scale-up. This is mainly because the map phase duration consists of more waves of map tasks on scale-up machines than on scale-out machines. Comparing OFS and HDFS in either scale-up or scale-out machines, we see that when the input data size is between 0.5 and 8GB, the map phase duration of these jobs are 10-50% shorter on HDFS. However, up-OFS still outperforms out-HDFS by 10-25% because of the higher benefits from scale-up machines for small jobs. When the input data size is larger than 16GB, the map phase duration is 10-40% shorter on OFS than on HDFS, no matter on the scale-up or scale-out cluster.

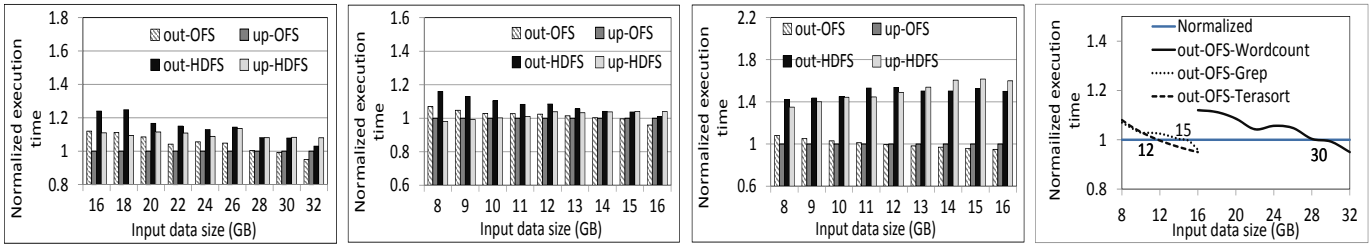
Figures 5(c), 6(c) and 7(c) show the shuffle phase duration of *Wordcount*, *Grep* and *Terasort*, respectively. We see that the shuffle phase duration is always shorter on scale-up machines than on scale-out machines. This is because the shuffle phase benefits from the larger memory resource and the RAMdisk of scale-up machines. We then can conclude from the map phase and shuffle phase figures: the cross point appears when the benefit of shuffle phase from scale-up machines is not able to compensate the drawback of scale-up machines due to fewer map and reduce slots.

Figures 5(d), 6(d) and 7(d) show the reduce phase duration of *Wordcount*, *Grep* and *Terasort*, respectively. The reduce phase of *Wordcount* and *Grep* just aggregates the map outputs which have small size, but the reduce phase of *Terasort* needs to sort the map outputs which has the same size as the input data. Therefore, *Wordcount* and *Grep* use only a few seconds during reduce phase after the shuffle phase, while *Terasort* uses around 5-1800 seconds. The scale-out machines have more reduce slots than scale-up machines and hence the reduce phase duration of *Terasort* acts similarly as the execution time and the map phase duration: when the input data size is small (0.5-8GB), the reduce phase duration is shorter on scale-up than on scale-out; when the input data size is large (>16GB), the reduce phase duration is shorter on scale-out than on scale-up. We see neither OFS nor HDFS affects the reduce phase duration of *Wordcount* and *Grep*. For *Terasort*, the reduce phase duration is 20-65% shorter on HDFS when the input data size is small (0.5-8GB) in either the scale-up or scale-out cluster. However, benefiting from scale-up machines, up-OFS has reduce phase duration 10-25% shorter than out-HDFS for *Terasort*. It is 20-70% shorter on OFS than on HDFS no matter on scale-up or scale-out cluster when the input data size is large (>16GB). Therefore, *Terasort* benefits from OFS during both map and reduce phase when the input data size is large, resulting in its higher execution time difference between OFS and HDFS. When the data is small, up-OFS is a better choice than traditional Hadoop with HDFS.

To illustrate the cross points of *Wordcount*, *Grep* and *Terasort*, we draw Figure 8, which shows the normalized execution time of each job on the scale-out cluster by its execution time on the scale-up cluster (e.g., $\frac{execution\ time\ on\ scale-out(Grep)}{execution\ time\ on\ scale-up(Grep)}$), denoted by out-OFS-Wordcount, out-OFS-Grep and out-OFS-Terasort, respectively. Recall that the shuffle/input ratio of *Wordcount*, *Terasort* and *Grep* is always around 1.6, 1 and 0.4, respectively. We from the figure that *Wordcount* with a larger shuffle/input ratio has a higher cross point (i.e., near 32GB) than the cross point (i.e., near 16GB) of *Grep* and *Terasort* with smaller shuffle/input ratios. A higher shuffle/input ratio leads to a higher cross point, and vice versa. Near the cross point, the benefit from scale-out cluster due to large input data size equals the benefit from scale-up cluster due to large shuffle data size. When the input data size is smaller than the cross point, scale-up cluster is a better choice, otherwise, scale-out cluster is a better choice.

As we mentioned above, a larger shuffle/input ratio results in a larger cross point. Intuitively, *Terasort* has a larger shuffle/input ratio (i.e., 1.0) than *Grep*'s shuffle/input ratio (i.e., 0.4), and hence *Terasort*'s cross point should be larger than the *Grep*'s. However, as shown in Figure 8, *Terasort* and *Grep* have the same cross point at 16GB, which does not appear as we expect. To investigate the cross point more accurately, we further conducted performance measurement of *Wordcount*, *Terasort* and *Grep* with the input data size range (16, 32) GB, (8,16) GB, and (8, 16) GB, respectively. Figures 9(a), 9(b) and 9(c) show the normalized execution time of *Wordcount*, *Terasort* and *Grep*. Figure 9(d) is the cross point figure newly generated from the above three figures.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems



(a) *Wordcount*. (b) *Grep*. (c) *Terasort*. (d) Cross point.

Fig. 9: Analysis of cross points for *Wordcount*, *Grep* and *Terasort*.

We see that the cross points of *Wordcount*, *Terasort* and *Grep* are 30, 12, and 15.

3.3 Performance of Map-Intensive Applications

In this section, we show the performance evaluation of the read and write test of *TestDFSIO*. In the read (or write) test, each map task is responsible for reading (or writing) a file. Therefore, the number of mappers is equal to the number of files. In either read or write test, there is only one reduce task, which collects and aggregates the statistics of the map tasks, such as completion time of read/write and file size.

Figures 10(a) and 11(a) show the normalized execution time of *TestDFSIO* write and read test versus output and input data size, respectively. Again, scale-up machines are the best for small write/read jobs (1-5GB) because small jobs do not require many mappers and scale-up machines can provide the required map slots and better CPU than scale-out machines. However, the execution time difference of *TestDFSIO* between scale-up and scale-out is not as significant as the shuffle-intensive applications. This is because *TestDFSIO* is a map-intensive application, which does not have large shuffle data. Therefore, the large memory benefit of scale-up machines for improving shuffle phase duration is not exhibited in map-intensive applications.

On the other hand, for large write/read data size ($\geq$10GB) (i.e., a large number of mappers), the performance follows out-OFS>up-OFS>out-HDFS. For large jobs, running on scale-out machines with remote storage is better than scale-up machines with remote storage because scale-out machines have more map slots and large jobs can be completed in fewer task waves. Running on scale-out machines with local storage results in the worst performance because the local disks are slower than the remote storage and the remote file system OFS can provide higher I/O performance than HDFS.

Figures 10(b), 10(c) and 10(d) and Figures 11(b), 11(c) and 11(d) show the map, shuffle and reduce phase durations of the write/read test, respectively. Since the map phase of *TestDFSIO* completes the majority work in the jobs, while the shuffle phase only collects the statistics and the reduce phase simply aggregates the results, we see that in both the write and read tests, the map phase duration exhibits a similar performance trends as the execution time. The shuffle and reduce phase durations of both tests are quite small (<8s), and they exhibit no specific relationships and are not affected by either OFS or HDFS. Comparing OFS and HDFS in the scale-up or scale-out cluster, when the write/read data size is small (1-5GB), HDFS leads to 10-20% shorter map phase duration. However, up-OFS still generates 5-15% shorter map phase duration than out-HDFS. When the write/read data size is large ($\geq$10GB), OFS leads to $50-80\%$ shorter map phase duration, a significant improvement.

We conclude that for map-intensive jobs in our experiment environment, if the write/read data size is small (1-5GB), the scale-up machines are the better choice because of better CPU. On the other hand, if the write/read data size is large ($\geq$10GB), scale-out machines can achieve better performance because of more map and reduce slots and better I/O performance of OFS over HDFS.

Figure 12 shows the normalized execution time of the write (denoted by out-OFS-Write) and read (denoted by out-OFS-Read) tests of *TestDFSIO* on the scale-out cluster by its execution time on the scale-up cluster, respectively. We see that the cross point is around 10GB for both tests. Since the shuffle size (in KB) is negligible (which makes the shuffle/input ratio close to 0) in both tests, these map-intensive jobs benefit little from the scale-up machines during the shuffle phase.

Again, to investigate the cross point more accurately, we further conducted performance measurement of write/read test of *TestDFSIO* at the range (6, 14)GB, respectively. Figures 13(a) and 13(b) show the normalized execution time of write/read test of *TestDFSIO*. Figure 13(c) is the cross point figure newly generated from Figures 13(a) and 13(b). Note that the write test actually has only a very small input data size but has a varying output data size in the measurement. We confirm that the cross points of write and read tests of *TestDFSIO* are both close to 10. As map-intensive applications have a small shuffle data size, map-intensive applications achieve less benefit from the large RAMdisks of the scale-up machines during the shuffle phase. We conclude that the cross points for map-intensive applications are smaller than those of shuffle-intensive applications.

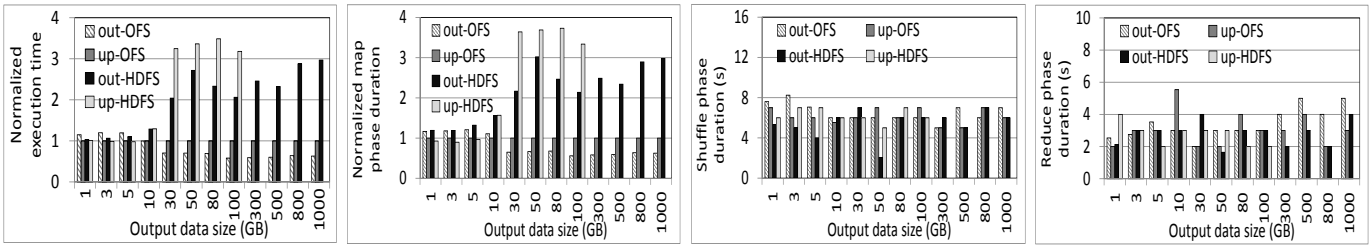
3.4 Analysis of Cross Points of Different Applications

In this section, we will discuss how different factors affect the cross point in detail. According to Figures 9(d) and 13(c), we use a tuple to represent the relationship between the shuffle/input ratio and cross point of each application. For example, the tuples of *Wordcount*, *Terasort*, *Grep* and read test of *TestDFSIO* are (1.6, 30), (1, 12), (0.4, 15), and (0, 10), respectively. If we consider a linear relationship between the shuffle/input ratio (S/I) and cross point (CP), we have:

$$CP = a * S/I + b. \quad (1)$$

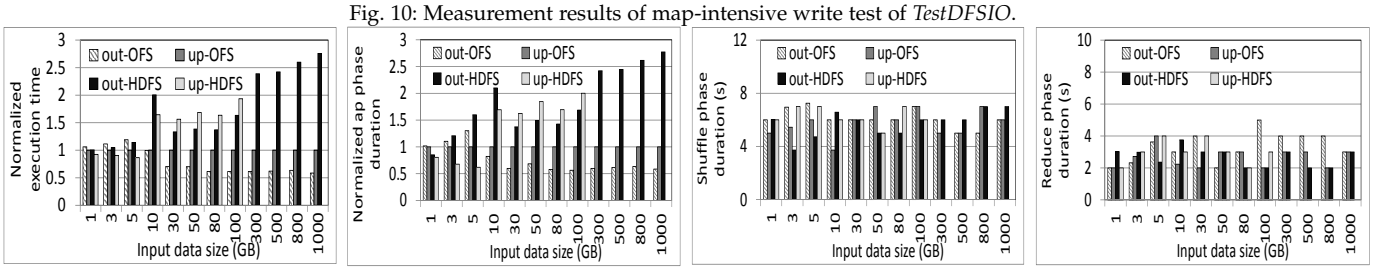
First, we substitute the tuples of cross points in Equation (1). We are able to calculate that one possible answer is $a = 12.5, b = 10$. Further, we find that the tuples of *Wordcount*, *Grep* and read test of *TestDFSIO* satisfy Equation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems



(a) Execution time. (b) Map phase duration. (c) Shuffle phase duration. (d) Reduce phase duration.

Fig. 10: Measurement results of map-intensive write test of *TestDFSIO*.



(a) Execution time. (b) Map phase duration. (c) Shuffle phase duration. (d) Reduce phase duration.

Fig. 11: Measurement results of map-intensive read test of *TestDFSIO*.

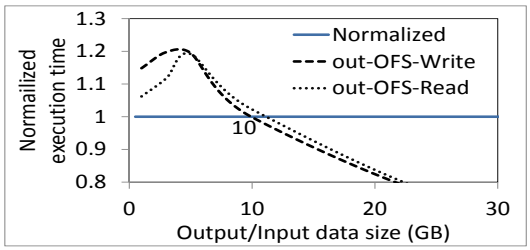


Fig. 12: Cross point of write and read test of *TestDFSIO*.

(1). However, *Terasort*'s tuple does not satisfy this equation, which is caused by the output data size of *Terasort*. In detail, the output data sizes of Wordcount, Grep and read test of TestDFSIO are very close to zero regardless of their input data size, and hence their output data sizes do not have much impact on the cross points. However, the output data size of Terasort increases as its input and shuffle data size increases, and hence has impact on the cross point determination.

We further investigate how the output data size affects the cross point by using the write test of *TestDFSIO*. In Figure 13(a), we see that the cross point of write test of *TestDFSIO* is 10GB output data size. This means that scale-up machines are better for the jobs with output data size smaller than 10GB, while scale-out machines are better for the jobs with output data size larger than 10GB. This is because the scale-out machines benefit from more CPU cores on writing the output data to the file system in parallel.

Similarly, we can conclude that the cross point of *Terasort* is affected by the output data size. In detail, if we substitute *Terasort*'s tuple in Equation (1), the cross point should be 22.5GB. However, the output data size of *Terasort* introduces a bias to the cross point, since the benefit is offset by the worse performance on scale-up machines when the output data size is greater than 10GB. Therefore, the measured cross point is less than 22.5GB. In other words, in order to compensate the poor performance introduced by the output data size (22.5-10=12.5GB greater than 10GB), the calculated cross point (22.5GB) is biased by 22.5-12=10.5GB. As shown in Figure 12, we can consider a linear relationship between the increase of output data size and the cross point.

Based on the above analysis, we conclude the process to calculate cross point for one job in our cluster as follows. First, we calculate the cross point CP based on Equation (1), Next, we check whether the cross point of this job is offset by the worse performance on scale-up machines caused by the output data size.

$$Output_{CP} = Output/Input * CP \geq 10, \tag{2}$$

where $Input$ and $Output$ are the input and output data size of the job, respectively, and $Output_{CP}$ is the output data size of the job corresponding to the cross point. If $Output_{CP}$ is greater than 10GB, the cross point calculated from Equation (1) needs to be offset, otherwise, it is not offset. Therefore, we have,

$$CP = \begin{cases} CP, & Output_{CP} < 10 \\ CP - \frac{10.5}{12.5} * (Output_{CP} - 10) & Output_{CP} \geq 10, \end{cases} \tag{3}$$

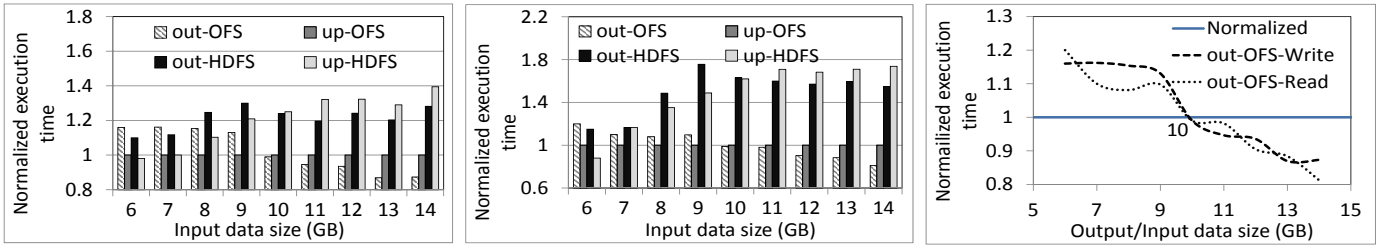
Summary of this section

• Hadoop with a remote file system improves the performance of Hadoop with HDFS, when jobs are scheduled to run on the hybrid cluster.

• We use a linear model for the cross point determination, since a linear model is sufficient to summarize the observations from the measurements and further determine the cross point in a small error. We will evaluate the accuracy of our model of cross point determination in the evaluation in Section 5.3.

Previous studies [30, 32] show that the execution time of a MapReduce job often has some linear relationship with its data size, as shown in Figure 4. The difference between scale-up machines and scale-out machines is that the two curves (execution time versus data size) have different slopes, which are affected by some factors. Since the impacted factors affect the slopes of the curves and the cross point is the intersection of the two curves, we can conclude that the cross point has some linear relationship with the impacted factors. Impacted factors are summaries as follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems



(a) Write test of *TestDFSIO*. (b) Read test of *TestDFSIO*. (c) Cross point.

Fig. 13: Analysis of cross points for write/read test of *TestDFSIO*.

• Input data size, shuffle/input ratio, and output data size all affect the benefits gained from scale-up and scale-out clusters. When the input data size is small, the scale-up cluster outperforms the scale-out cluster, and when the input/output data size is large, the scale-out cluster outperforms scale-up machines. The cross point depends on the shuffle data size; a larger shuffle size leads to more benefits from the scale-up machines and vice versa.

4 A Hybrid Scale-up/out Hadoop Architecture

A traditional Hadoop cluster only has scale-out machines in general. However, we have demonstrated in Section 3 that different jobs perform better on scale-up machines or on scale-out machines. In current real-world workloads, the jobs are increasingly diverse mix of computations and data size levels. Recall that in Section 1 improving the performance of small jobs is important for production clusters. Therefore, we propose a new Hadoop architecture including both scale-up and scale-out machines to achieve better performance for real-world workloads.

As indicated in Section 1, there exist two main challenges for building such a new architecture: data storage and scale-up or scale-out selection. For the data storage challenge, HDFS and storing all data blocks needed by scale-up machines in themselves are not efficient methods. Besides, this latter solution requires the modification of the function of the namenode to decide where to store the data blocks based on the jobs handling them. Since usually data is distributed before jobs are launched, this solution introduces not only extra overhead and but also complexity to the namenode's function. We then use a remote file system (e.g., OFS) for the hybrid Hadoop architecture, the feasibility and advantage of which are shown in Section 3. Since both scale-up and scale-out machines can be mounted with the same remote file system on HPC, jobs can read/write data no matter they are scheduled to the scale-up or scale-out clusters without data transmission between machines. Moreover, using a remote file system allows us to implement the hybrid architecture without modifying the namenode code for data distribution. Intuitively, it seems that this hybrid architecture improves the performance of small jobs at the cost of the performance of large jobs because the traditional Hadoop cluster has more map and reduce slots than the hybrid architecture. However, we demonstrate in Section 5 that even with the hybrid architecture, the performance of large jobs is improved due to better I/O performance of OFS over HDFS and less slot competition for large jobs.

To handle the second challenge, we leverage our observations in Section 3 to make the decisions based on job characteristics. Our experiments show that when a job has shuffle/input ratio between 0.4 and 1, if the input data size is smaller than 16GB, scale-up is a better choice, otherwise, scale-out is a better choice. When a job has shuffle/input ratio equals 1.6, if the input data size is smaller than 32GB, scale-up is a better choice, otherwise, scale-out is a better choice. We generalize 1.6 to ratios greater 1. We consider jobs with shuffle/input ratios less than 0.4 as map-intensive jobs, for which when the input data size is smaller than 10GB, scale-up is a better choice, otherwise, scale-out is a better choice. Based on these observations, we design an algorithm to decide whether scale-up or scale-out is a better choice for a given job based on its shuffle/input ratio and input data size. The pseudo-code of this algorithm is shown in Algorithm 1.

However, we mentioned that the output data size also affects the determination of cross point. Since Algorithm 1 does not consider the output data size factor, we further propose an advanced scheduling based on the analysis of Section 3.4, which jointly considers the shuffle/input ratio, the input data size, and the output data size. The pseudo-code of this advanced algorithm is shown in Algorithm 2.

We assume that job characteristics (i.e., the shuffle/input ratio, the input data size and the output data size) are already known by the users, which means that either the users once ran the jobs before (i.e., recurring jobs) or the jobs are well-known as map-intensive or shuffle-intensive. Previous works [8, 20, 23] show that a large number of jobs in production workloads are recurring jobs, which run periodically and have predictable characteristics such as input data size and shuffle/input ratio. Authors in [23] show that the future job characteristics for such jobs can be predicted with an error as low as 6.5%, which allows the users to take advantage of our scheduling algorithm. If the users do not know the shuffle/input ratio of the jobs, we consider the jobs as map-intensive (i.e., shuffle/input ratio less than 0.4) and hence the cross points of the jobs are smaller (i.e., 10GB). This is because we need to avoid scheduling any large jobs to the scale-up machines. Otherwise, it would result in performance degradation of small jobs. For example, a job actually has a shuffle/input ratio of 0.1, input data size of 15GB, and output data size of close to 0. From Equation (1), the job has a cross point of 12.5GB. However, if the user does not know the shuffle/input ratio and we consider the job as a shuffle-intensive job, it will be scheduled to the scale-up machines. However, it actually should be scheduled to scale-out machines. Due to this wrong scheduling, it increases the number of jobs on scale-up machines and degrades the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

performance of other small jobs, which should be avoided.

Algorithm 1 Selecting scale-up or scale-out for a given job without considering the output data size.

Inputs: $NextJob$: next job in the queue
$Input$: input data size of the job
S/I: shuffle/input ratio
while $NextJob$ exists in the job queue **do**
  if $shuffle/input\ ratio > 1$ **then**
    if $InputDataSize < 32GB$ **then**
      Scale-up $\leftarrow NextJob$
    else
      Scale-out $\leftarrow NextJob$
    end if
  else if $0.4 \leq shuffle/input\ ratio \leq 1$ **then**
    if $InputDataSize < 16GB$ **then**
      Scale-up $\leftarrow NextJob$
    else
      Scale-out $\leftarrow NextJob$
    end if
  else
    if $InputDataSize < 10GB$ **then**
      Scale-up $\leftarrow NextJob$
    else
      Scale-out $\leftarrow NextJob$
    end if
  end if
end while

Algorithm 2 Selecting scale-up or scale-out for a given job considering the output data size.

Inputs: $NextJob$: next job in the queue
$Input$: input data size of the job
S/I: shuffle/input ratio
$Output$: output data size of the job
while $NextJob$ exists in the job queue **do**
  Calculate the cross point (CP) based on Equations (1), (2), and (3)
  if $Input < CP$ **then**
    Scale-up $\leftarrow NextJob$
  else
    Scale-out $\leftarrow NextJob$
  end if
end while

Note that we calculate the cross points according to the measurement results from our cluster configurations. We can conduct more experiments with other different jobs to make the algorithm more accurate. Also, different cluster configurations of scale-up and scale-out machines will lead to different cross point results and the coefficients of Equation (1) may not be fit for other cluster configurations. In this paper, we only attempt to show the factors affecting the scale-up and scale-out selection and how they affect the selection decision, and provide a method to design the selection algorithm for the hybrid architecture. Other designers can follow the same method to measure the cross points in their clusters and develop the hybrid architecture.

5 Performance Evaluation

In this section, we use the Facebook synthesized workload FB-2009 [14] to evaluate the performance of our hybrid scale-up/out Hadoop architecture compared with traditional Hadoop architecture. The CDF of input data size of this workload is shown in Figure 3. We see that more than 80% of jobs have an input data size less than 10GB. Due to the limitation of space, please refer the other characteristics of this workload to [14].

5.1 Experimental Setting

We used 2 scale-up machines and 12 scale-out machines to deploy the hybrid scale-up/out architecture with the OFS remote file system, and the hardware configurations of these machines are the same as explained in Section 3. In order to evaluate the two scheduling algorithms in Section 4, we deployed the hybrid scale-up/out architecture with Algorithm 1 (Hybrid in short) and the advanced Algorithm 2 (Hybrid-adv in short), respectively. As a baseline, we deployed a traditional Hadoop cluster (THadoop in short) with HDFS and a Hadoop cluster with remote file system OFS (RHadoop in short) using 24 scale-out machines (which have comparably the same total cost as the machines in the hybrid architecture) and one additional namenode. Since the trace workload is synthesized from a 600-machine cluster and we did our experiments on 24 machines, we shrank the input/shuffle/output data size of the workload by a factor of 5 to avoid disk insufficiency. We ran the Facebook workload consecutively on these two architectures based on the job arrival time in the traces. In this experiment, we assumed that the characteristics of all the submitted jobs are known based on the FB-2009 trace. We refer to the jobs that are scheduled to scale-up cluster and scale-out cluster by our scheduler as *scale-up jobs* and *scale-out jobs*, respectively.

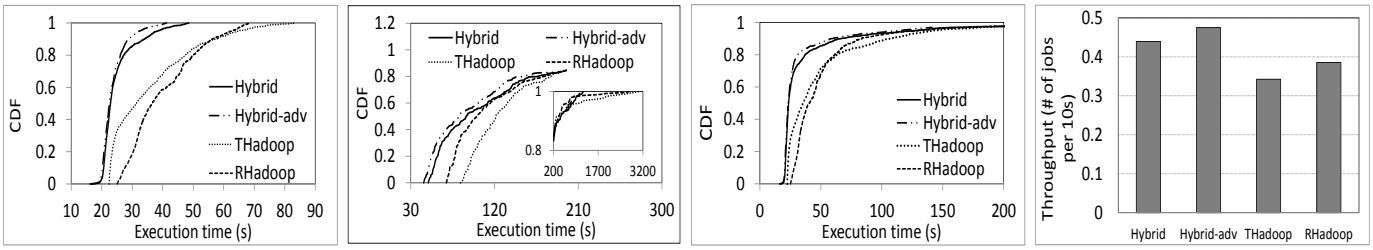
5.2 Performance Analysis

Figure 14(a) shows the CDF of the execution time of the scale-up jobs in the workload. We see that the execution time distribution of Hybrid and Hybrid-adv are much broader than THadoop and RHadoop. Their maximum execution time for scale-up jobs is 48.53s, 41.89s, 83.37s and 68.17s on Hybrid, Hybrid-adv, THadoop and RHadoop, respectively. This result demonstrates the effectiveness of the hybrid architecture in improving the performance of small jobs. In addition, Hybrid-adv has a broader execution time distribution than Hybrid, which means that Hybrid-adv outperforms Hybrid for the scale-up jobs. This is because Hybrid-adv provides a more accurate scale-up or scale-out selection for the jobs, which allows the jobs to accurately select the suitable machines and achieve more performance improvement. RHadoop has the worst performance because OFS performs worse than HDFS for small input data sizes on the scale-out Hadoop.

Figure 14(b) shows the CDF of the execution time of the scale-out jobs in the workload. In order to illustrate the figure clearly, we only show the scale-out jobs with execution time less than 200s in the main figure, and use the embedded small figure to show those with execution time greater than 200s. For scale-out jobs, the maximum execution time is 1207s, 1099s, 3087s and 2734s on Hybrid, Hybrid-adv, THadoop and RHadoop, respectively. The percent of jobs completed after 1099s on THadoop and RHadoop are 4.7% and 1.7%, respectively. Benefitting from the higher I/O performance of OFS, RHadoop outperforms THadoop for large input sizes.

Intuitively, it seems that we improve the performance of scale-up jobs at the cost of the performance of scale-out jobs and the scale-out jobs should have better performance on THadoop because THadoop has more map and reduce slots than Hybrid and Hybrid-adv. However, Figure 14(b) shows

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems



(a) Execution time of scale-up jobs. (b) Execution time of scale-out jobs. (c) Execution time of all jobs. (d) Throughput.

Fig. 14: Measurement results of the Facebook workload experiment.

that Hybrid and Hybrid-adv still outperform THadoop and RHadoop even for scale-out jobs, indicating that Hybrid and Hybrid-adv improve not only the performance of scale-up jobs but also the scale-out jobs. There are two main reasons. The first reason is because Hybrid and Hybrid-adv is configured with the remote file system, which provides better performance for jobs with large input sizes than HDFS, as shown in Section 3. The second reason is that although there are more map and reduce slots in THadoop, a large amount of scale-up jobs in the workload occupy the slots and have poor performance due to less powerful CPU, thus resulting in a long time before releasing the occupied slots to scale-out jobs. On the contrary, in Hybrid and Hybrid-adv, all the scale-up jobs are run on the scale-up cluster, while scale-out jobs run on the scale-out cluster. Therefore, scale-out jobs in Hybrid and Hybrid-adv do not need to compete with scale-up jobs for slots or with other scale-out jobs because only 15% of the jobs in the workload are scale-out jobs. Similarly, this is also the reason that Hybrid and Hybrid-adv outperform RHadoop for scale-out jobs. Therefore, scale-out jobs in Hybrid and Hybrid-adv can always be provided with more map and reduce slots than THadoop or RHadoop, resulting in better performance in Hybrid and Hybrid-adv. In addition, Hybrid-adv outperforms Hybrid, since Hybrid-adv provides a more accurate scale-up or scale-out selection for the jobs, which allows the jobs to achieve the best benefits from the hybrid architecture.

Figure 14(c) show the CDF of the execution time of all the jobs in the workload. To make the figure clear, we eliminated 2.57%, 2.21%, 2.31% and 2.35% of the jobs with execution time greater than 200 seconds on Hybrid, Hybrid-adv, THadoop and RHadoop, respectively. We see that the execution time distribution on Hybrid and Hybrid-adv are broader than THadoop and RHadoop, which implies that Hybrid and Hybrid-adv provide better performance due to the higher performance for both scale-up and scale-out jobs as explained previously. Hybrid-adv outperforms Hybrid due to the aforementioned reason that Hybrid-adv provides a more accurate scale-up or scale-out selection for the jobs.

Moreover, we observe that Hybrid and Hybrid-adv not only improves the performance of jobs, but also generate fewer job running failures. Specifically, 98%, 98%, 82.3% and 89.2% of jobs were successfully executed on Hybrid, Hybrid-adv, THadoop and RHadoop, respectively. The failures on THadoop and RHadoop were observed generally in job groups with similar arrival time and the failed groups appeared multiple times. Specifically, the failures occur when there are many small jobs and several large jobs submitted to THadoop and RHadoop, and the large jobs have reduce tasks that take a long time to complete. A large number of small jobs that have completed their map tasks must wait for the reduce slots in queue. Then, all the shuffle data from these small jobs is spilled to the local disks, which overloads the local disks and generates job failures. On the other hand, the jobs are less likely to fail on Hybrid and Hybrid-adv because the workload is split into a large percent (85%) of scale-up jobs and a small percent (15%) of scale-out jobs. The short execution time of scale-up jobs prevents from forming a long queue on scale-up machines. Scale-out machines process only a small percent of scale-out jobs, which makes it less likely to generate slot competition or a long job queue. However, the utilization of the remote file system for input and output data avoids overloading local disks, resulting in less failures on Hybrid, Hybrid-adv, and RHadoop than on THadoop. Note that when the authors [14] of the Facebook synthesized workload ran this workload (with decreased input/shuffle/output size by a factor of 3) on a 200-machine cluster, 8.4% of the jobs failed. Although their configurations and hardware have some differences from ours, our 98% success rate is still very promising.

Figure 14(d) shows the throughput, defined as the number of jobs finished per ten seconds. We see that the throughput of Hybrid, Hybrid-adv, THadoop and RHadoop is 0.44, 0.48, 0.34 and 0.38, respectively. RHadoop has better throughput than THadoop because of its higher success rate. Hybrid-adv improves the throughput of Hybrid, THadoop and RHadoop by 8%, 39% and 23%, respectively. This result demonstrates the effectiveness of Hybrid and Hybrid-adv in improving the throughput. In summary, our trace-driven experiments show that the proposed hybrid architecture outperforms the traditional Hadoop architecture in terms of job execution time, failure rate and throughput.

5.3 Sensitivity and Accuracy Analysis

The benefits of our Hybrid cluster depend on (a) the accuracy that the job characteristics (e.g., input, shuffle, and output data size) can be predicted, and (b) the model to determine cross point. In this section, we evaluate the Hybrid cluster's performance to the variation of these factors. In the following figures, we only evaluate the results of Hybrid-adv.

Error in predicted job characteristics. As mentioned before, we define the type of an application by its shuffle/input ratio. Besides, the cross point is also related to shuffle/input ratio. Therefore, we varied the error of shuffle/input ratio of jobs. Figure 15(a) shows the reduction of throughput on Hybrid-adv versus the error rate of shuffle/input ratio. We see that, as the shuffle/input ratio decreases, it does not reduce the throughput on Hybrid-adv much. Note that the decrease of shuffle/input ratio leads to a smaller cross

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

point according to Equation (1). A smaller cross point could result in many scale-up jobs running on scale-out machines. However, even when the cross point is decreased, we see that it still provides a better throughput than THadoop and RHadoop because there are still plenty of scale-up jobs running on scale-up machines. However, as the shuffle/input ratio increases, it does reduce the throughput of Hybrid-adv severely. This is because once the shuffle/input ratio increases, the cross point also increases. This allows many scale-out jobs to run on scale-up machines, which leads to a poor performance of the scale-up jobs since the scale-out jobs consume a large amount of resources on scale-up machines for a long time.

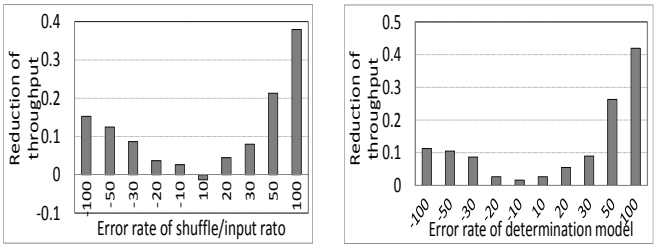
Error in the model to determine cross point. In Hybrid, we roughly determine the cross point as 32GB for shuffle-intensive applications with shuffle/input ratio greater than 1, 16GB for shuffle-intensive applications with shuffle/input ratio greater than 0.4 but smaller than 1, and 10GB for map-intensive applications. In Hybrid-adv, a fine-grained model of cross point determination is defined by Equations (1), (2), and (3). In order to test the robustness and accuracy of Hybrid-adv, we varied the cross point with error, while the shuffle/input ratios of all the jobs are set according the trace.

Figure 15(b) shows the reduction of throughput on Hybrid-adv versus the error rate of cross point. We see that, as the cross point cutoff line decreases, there is not much reduction in throughput because of the aforementioned reason that there are still plenty of scale-up jobs running on scale-up machines. When the cross point cutoff line increases 10%-30%, the reduction of throughput in Hybrid-adv is lower than 10%, which still provides significant better performance than THadoop and RHadoop. Hence, a small error of cross point determination does not affect much on the performance of the Hybrid cluster, which indicates the robustness of Hybrid cluster.

As the cross point increases, the reduction of throughput on Hybrid-adv increases. When the cross point increases by 100%, the reduction of throughput on Hybrid-adv reaches 41%. This is because setting a much higher cross point allows many scale-out jobs to run on scale-up machines, which consumes a large amount of resources on scale-up machines for a long time. It results in queuing the scale-up jobs for a long time, and hence leads to a poor performance on Hybrid-adv.

Further, Figure 15(b) also reflects the accuracy of our current model on cross point determination. First, when the cross point varies slightly, there is not much impact on the throughput. Second, if we set the cross point higher or lower, it results in a poor throughput performance on Hybrid-adv. It indicates that our model determines the cross point in a relatively accurate range. Otherwise, it would lead to a significant throughput reduction on Hybrid-adv.

In summary, we see that the performance of Hybrid-adv is not quite sensitive to the error of job characteristics in 30%. On the other hand, as the error rate increases, the performance of Hybrid-adv can be severely impacted. However, recent studies [18, 23] show that the job characteristics can often be accurately predicted, since there are a large amount of recurring jobs in production cluster. They demonstrate that the job characteristics can be estimated with a small error of 6.5% in average. Additionally, we demonstrate that our model can determine the cross point in a relatively accurate range.

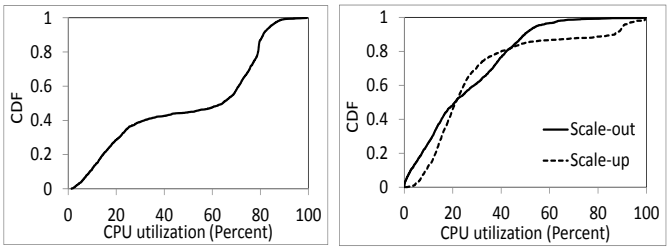


(a) Error in predicted shuffle/input ratio of jobs. (b) Error in the cross point determination model.

Fig. 15: Sensitivity and accuracy analysis.

5.4 Power Consumption Discussion

In modern clusters, power consumption becomes a significant expense. In this section, we expect to estimate the power consumption of the hybrid cluster, compared to the traditional scale-out cluster.



(a) CPU utilization on THadoop. (b) CPU utilization on Hybrid.

Fig. 16: Power consumption analysis.

It has been shown that the main power consumers in a machine are the CPUs and memory [11, 27]. It is also reasonable to assume that the power consumption of other elements like memory and hard disks is constant [11, 27].

We use a well-known CPU utilization centric model [11, 12, 27] to estimate the power consumption. Additionally, previous studies demonstrate that the power consumption has a linear relationship with the CPU utilization. In detail, the power consumption can be estimated by the following equation.

$$P = P_{idle} + \alpha * (P_{max} - P_{idle}), \tag{4}$$

where α is the CPU utilization, P_{idle} is the power consumption at idle and P_{max} is the power consumption at maximum performance. Through empirical measurement of different servers, this model can estimate the power consumption of a machine within an error rate of ± 5 percent [11, 27].

Therefore, in order to estimate the power consumption of THadoop and Hybrid-adv, we measured CPU utilization every second using the *top* command when they ran the Facebook workload. We show the results in a period of 1800 seconds. The metric we use is average CPU utilization each second, which is calculated by

$$\frac{\sum CPU\ utilization\ of\ all\ the\ machines}{number\ of\ machines}. \tag{5}$$

Figure 16(a) shows the CDF of average CPU utilization each second on THadoop. Figure 16(b) shows the CDF of the average CPU utilization each second on Hybrid-adv, breaking down by scale-up and scale-out machines.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

After checking the specifications of scale-up machine and scale-out machines, we find that each scale-up machine consume 1200W at maximum, while each scale-out machine consumes 450W at maximum. The power consumption at idle is observed to be around 70% [11, 27]. Therefore, we consider P_{idle} for each scale-up machine as 840W and P_{idle} for each scale-out machine as 315W.

Now, we use Equation (4) to calculate the power consumption each second. Summing up the power consumption for 1800 seconds, we find that THadoop consumes around $1.61 * 10^7$J in total, while Hybrid-adv consumes $1.43 * 10^7$J in total. Therefore, we can conclude that Hybrid-adv consumes less power than THadoop.

In summary, Hybrid-adv cluster can not only provide better job execution time and throughput for MapReduce workloads, but also consumes less power than THadoop. Some previous works such as GreenMR [28] propose power consumption models to predict the power consumption of scale-up jobs and scale-out jobs, and then utilizes the models to schedule different jobs to reduce the power consumption. Our Hybrid-adv cluster can accommodate the power consumption models to guide our scheduling decision to further reduce the power consumption and improve the performance. We leave this improvement of Hybrid-adv cluster as our future work.

6 Related Work

MapReduce [17] has been a popular framework that performs parallel computations on big data. Cluster provisioning, configuring and managing for the Hadoop clusters is essential, which requires thorough understanding of the workloads. Recently, there are many efforts devoted to characterizing the workloads in real-world cluster. Chen *et al.* [14] analyzed two production MapReduce traces from Yahoo! and Facebook in order to establish a vocabulary for describing MapReduce workloads. Their another work [13] characterized new MapReduce workloads, which are driven in part by interactive analysis and with heavy use of query-like programming frameworks such as Hive on top of MapReduce. Ren *et al.* [29] characterized the workload of Taobao production Hadoop cluster to provide an understanding of the performance and the job characteristics of Hadoop in the production environment. Kavulya *et al.* [24] analyzed MapReduce logs from the $M45$ supercomputing cluster. Appuswamy *et al.* [10] conducted an evaluation of representative Hadoop jobs on scale-up and scale-out machines, respectively. They found that scale-up machines achieve better performance for jobs with data size at the range of MB and GB. Our work is different from the above workload characterization works is that our work is the first that compare the performance of the Hadoop workloads in the four architectures shown in Table 1.

The above works also indicate that the majority jobs in production workloads generally do not process large data size. Since MapReduce [17] is primarily designed to process large data sets, there is a conflict between the goal of MapReduce and current production workloads. Improving the performance of small jobs attracts much attention from the research community. For example, Elmeleegy [19] presented Piranha, a system to optimize the small jobs without affecting the larger jobs. Motivated by this paper, we focused on finding out if the coexistence of scale-up and scale-out machines can improve the performance of workloads with a majority of small jobs.

Many works focus on improving the performance of the MapReduce clusters from different aspects such as job scheduling [3, 5, 21, 22], intermediate data shuffling [7, 15, 16, 34] and improving small job performance [19]. The work in [7] replaces HDFS with the Lustre file system and places shuffle data in Lustre. MapReduce online [15] sends shuffle data directly from map tasks to reduce tasks without spilling the shuffle data to the disks in order to reduce the shuffle phase duration. Camdoop [16] performs in-network aggregation of shuffle data during data forwarding in order to decrease the network traffic. Wang *et al.* [34] proposed JVM-Bypassing shuffling for Hadoop to avoid the overhead and limitations of the JVM. Unlike these previous works that focus on improving the performance of traditional Hadoop, our work focuses on designing a new hybrid scale-up/out Hadoop architecture that fully utilizes both the advantages of scale-up and scale-out machines for different jobs in a workload to improve its performance.

7 Conclusion

Since a real-world workload usually has many jobs with increasingly diverse mix of computations and data size levels, solely using either scale-up or scale-out cluster to run a workload cannot achieve high performance. Thus, in this paper, we explore building a hybrid scale-up/out Hadoop architecture. However, building such an architecture faces two main challenges. First, how to distribute data blocks of a workload dataset to avoid degrading node performance caused by limited local disk or data transmission between nodes. Second, how to decide whether to use scale-up or scale-out cluster for a given job. To handle these challenges, we have conducted performance measurement of different applications on four HPC-based Hadoop platforms: scale-up machines with OFS, scale-out machines with OFS, scale-up machines with HDFS, and scale-out machines with HDFS. Based on our measurement results, we design a hybrid scale-up/out Hadoop architecture, which uses a remote file system rather than HDFS and has a scheduler to determine using scale-up or scale-out for a given job to achieve better performance. We further conducted experiments driven by the Facebook synthesize workload to demonstrate that this new architecture outperforms both the traditional Hadoop with HDFS and with OFS. This is the first work that proposes the hybrid architecture and we propose a simple scheduler to handle the key challenges. We consider our work as a starting point and expect it will stimulate many other works on this topic. In our future work, we will further develop the scheduler to be more comprehensively, such as the load balancing between the scale-up machines and scale-out machines, HDFS block size, and the number of map/reduce tasks of jobs, etc. For example, if many small jobs arrive at the same time without any large jobs, all the jobs will be scheduled to the scale-up machines, resulting in imbalance allocation of resources between the scale-up and scale-out machines.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2016.2573820, IEEE Transactions on Parallel and Distributed Systems

ACKNOWLEDGEMENTS

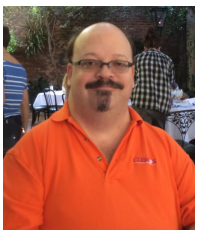
We would like to thank Mr. Matthew Cook for his insightful comments. This research was supported in part by U.S. NSF grants NSF-1404981, NSF 1228312 MRI, IIS-1354123, CNS-1254006, and Microsoft Research Faculty Fellowship 8300751. An early version of this work was presented in the Proceedings of ICPP 2015 [26].

REFERENCES

[1] Accelerate Hadoop MapReduce Performance using Dedicated OrangeFS Servers. http://www.datanami.com/2013/09/09/accelerate_hadoop_mapreduce_performance_using_dedicated_orangefs_servers.html. [Accessed in Oct. 2015].
[2] Apache Hadoop. http://hadoop.apache.org/. [Accessed in Oct. 2015].
[3] Capacity Scheduler. http://hadoop.apache.org/docs/r1.2.1/capacity_scheduler.html. [Accessed in Oct. 2015].
[4] Clemson Palmetto. http://newsstand.clemson.edu/media relations/clemson-supercomputers-ranked-in-top-5-fastest-at-public-universities. [Accessed in Oct. 2015].
[5] Fair Scheduler. http://hadoop.apache.org/docs/r1.2.1/fair_scheduler.html. [Accessed in Oct. 2015].
[6] OrangeFS. http://www.orangefs.org. [Accessed in Oct. 2015].
[7] Using Lustre with Apache Hadoop. http://wiki.lustre.org/images/1/1b/Hadoop_wp_v0.4.2.pdf. [Accessed in Oct. 2015].
[8] S. Agarwal, S. Kandula, N. Bruno, M.-C. Wu, I. Stoica, and J. Zhou. Re-optimizing data-parallel computing. In *Proc. of NSDI*, 2012.
[9] S. Agrawal. The Next Generation of Apache Hadoop MapReduce. Apache Hadoop Summit India, 2011.
[10] R. Appuswamy, C. Gkantsidis, D. Narayanan, O. Hodson, and A. Rowstron. Scale-up vs scale-out for hadoop: Time to rethink? In *Proc. of SOCC*, 2013.
[11] A. Beloglazov and R. Buyya. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*, 24(13):1397–1420, 2012.
[12] L. Chen, H. Shen, and K. Sapra. Distributed autonomous virtual resource management in datacenters using finite-markov decision process. In *Proc. of SOCC*, pages 1–13, 2014.
[13] Y. Chen, S. Alspaugh, and R. Katz. Interactive Analytical Processing in Big Data Systems: A CrossIndustry Study of MapReduce Workloads. In *Proc. of VLDB*, 2012.
[14] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz. The Case for Evaluating MapReduce Performance Using Workload Suites. In *Proc. of MASCOTS*, 2011.
[15] T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, K. Elmeleegy, and R. Sears. Mapreduce online. In *Proc. of NSDI*, 2010.
[16] P. Costa, A. Donnelly, A. Rowstron, and G. O'Shea. Camdoop: Exploiting in-network aggregation for big data applications. In *Proc. of NSDI*, 2012.
[17] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *Proc. of OSDI*, 2004.
[18] P. Delgado, F. Dinu, A.-M. Kermarrec, and W. Zwaenepoel. Hawk: hybrid datacenter scheduling. In *Proc. of ATC*, pages 499–510, 2015.
[19] K. Elmeleegy. Piranha: Optimizing Short Jobs in Hadoop. In *Proc. of VLDB Endow*, 2013.
[20] A. D. Ferguson, P. Bodik, S. Kandula, E. Boutin, and R. Fonseca. Jockey: guaranteed job latency in data parallel clusters. In *Proc of EuroSys*, 2012.
[21] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella. Multi-resource Packing for Cluster Schedulers. In *Proc. of SIGCOMM*, 2014.
[22] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. Katz, S. Shenker, and I. Stoica. Mesos: A platform for fine-grained resource sharing in the data center. In *Proc. of NSDI*, 2011.
[23] V. Jalaparti, P. Bodik, I. Menache, S. Rao, K. Makarychev, and M. Caesar. Network-aware scheduling for data-parallel jobs: Plan when you can. In *Proc. of SIGCOMM*, 2015.
[24] S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan. An analysis of traces from a production MapReduce cluster. In *Proc. of CCGRID*, 2010.
[25] S. Krishnan, M. Tatineni, and C. Baru. myHadoop-Hadoop-on-Demand on Traditional HPC Resources. Technical report, 2011.
[26] Z. Li and H. Shen. Designing a hybrid scale-up/out hadoop architecture based on performance measurements for high application performance. In *Proc. of ICPP*, 2015.
[27] L. Minas and B. Ellison. *Energy Efficiency for Information Technology: How to Reduce Power Consumption in Servers and Data Centers*. Intel Press, 2009.
[28] Z. Niu, B. He, and F. Liu. Not all joules are equal: Towards energy-efficient and green-aware data processing frameworks. In *Proc. of IC2E*, 2016.
[29] Z. Ren, X. Xu, J. Wan, W. Shi, and M. Zhou. Workload characterization on a production Hadoop cluster: A case study on Taobao. In *Proc. of IISWC*, 2012.
[30] N. B. Rizvandi, A. Y. Zomaya, A. J. Boloori, and J. Taheri. On modeling dependency between mapreduce configuration parameters and total execution time. *arXiv preprint arXiv:1203.0651*, 2012.
[31] A. Rowstron, D. Narayanan, A. Donnelly, G. O'Shea, and A. Douglas. Nobody ever got fired for using hadoop on a cluster. In *Proc. of HotCDP*, 2012.
[32] G. Wang, A. Khasymski, K. Krish, and A. R. Butt. Towards improving mapreduce task scheduling using online simulation based predictions. In *Proc. of ICPADS*, pages 299–306, 2013.
[33] L. Wang, J. Zhan, C. Luo, Y. Zhu, Q. Yang, Y. He, W. Gao, Z. Jia, Y. Shi, S. Zhang, et al. Bigdatabench: A big data benchmark suite from internet services. In *Proc. of HPCA*, 2014.
[34] Y. Wang, C. Xu, X. Li, and W. Yu. Jvm-bypass for efficient hadoop shuffling. In *Proc. of IPDPS*, 2013.

Zhuozhao Li received the B.S. degree in Optical Engineering from Zhejiang University, China in 2010, and the M.S. degree in Electrical Engineering from University of Southern California in 2012. He is currently a Ph.D. candidate in Electrical and Computer Engineering at Clemson University. His research interests include distributed computer systems, with an emphasis on cloud computing, resource management in cloud networks, and big data processing.

Haiying Shen received the B.S. degree in Computer Science and Engineering from Tongji University, China in 2000, and the M.S. and Ph.D. degrees in Computer Engineering from Wayne State University in 2004 and 2006, respectively. She is currently an Associate Professor in the Department of Electrical and Computer Engineering at Clemson University. Her research interests include distributed computer systems and computer networks, with an emphasis on P2P and content delivery networks, mobile computing, wireless sensor networks, and cloud computing. She is a Microsoft Faculty Fellow of 2010, a senior member of the IEEE and a member of the ACM.



Walter Ligon received the B.S. and M.S. degrees in Information and Computer Science from Georgia Institute of Technology in 1987 and 1988, respectively, and the Ph.D. degree in Computer Science from Georgia Institute of Technology in 1992. He is currently an Associate Professor in the Department of Electrical and Computer Engineering at Clemson University. His research interests include parallel I/O systems, parallel computing environments, and high performance computing.

Jeffrey Denton received the B.S. degree in Computer Science from Clemson University in 2011. He is currently a Deputy Director of Data Science with the CITI group of CCIT and a system administrator of both Palmetto and Cypress Clusters at Clemson University. His research interests include distributed systems, storage, cloud computing, and programming languages.