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/TBDATA.2016.2599933, IEEE Transactions on Big Data

Accelerating MapReduce on Commodity Clusters: An SSD-Empowered Approach

Bo Wang, *Member, IEEE,* Jinlei Jiang, *Member, IEEE,* Yongwei Wu, *Member, IEEE,*
Guangwen Yang, *Member, IEEE,* Keqin Li, *Fellow, IEEE*

Abstract—MapReduce, as a programming model and implementation for processing large data sets on clusters with hundreds or thousands of nodes, has gained wide adoption. In spite of the fact, we found that MapReduce on commodity clusters, which are usually equipped with limited memory and hard-disk drive (HDD) and have processors of multiple or many cores, does not scale as expected as the number of processor cores increases. The key reason for this is that the underlying low-speed HDD storage cannot meet the requirement of frequent IO operations. Though in-memory caching can improve IO, it is costly and sometimes cannot get the desired result either due to memory limitation.

To deal with the problem and make MapReduce more scalable on commodity clusters, we present mpCache, a solution that utilizes solid-state drive (SSD) to cache input data and localized data of MapReduce tasks. In order to make a good trade-off between cost and performance, mpCache proposes ways to dynamically allocate the cache space between the input data and localized data and to do cache replacement. We have implemented mpCache in Hadoop and evaluated it on a 7-node commodity cluster by 13 benchmarks. The experimental results show that mpCache can gain an average speedup of 2.09x when compared with Hadoop, and can achieve an average speedup of 1.79x when compared with PACMan, the latest in-memory optimization of MapReduce.

Index Terms—Big data, data caching, MapReduce, scheduling

1 Introduction

1.1 Motivation

The human society has stepped into the big data era where applications that process terabytes or petabytes of data are common in science, industry and commerce. Usually, such applications are termed IO-intensive applications, for they spend most time on IO operations. Workloads from Facebook and Microsoft Bing data centers show that IO-intensive phase constitutes 79% of a job's duration and consumes 69% of the resources [2].

MapReduce [6] is a programming model and an associated implementation for large data sets processing on clusters with hundreds or thousands of nodes. It adopts a data parallel approach that first partitions the input data into multiple blocks and then processes them independently using the same program in parallel on a certain computing platform (typically a cluster). Due to its scalability and ease of programming, MapReduce has been adopted by many companies, including Google [6], Yahoo [15], Microsoft [18] [46], and Facebook [43]. Nowadays we can see MapReduce applications in a wide range of areas such as distributed sort, web link-graph reversal, term-vector per host, web log analysis, inverted index construction, document clustering, collaborative filtering, machine learning, and statistical machine translation, to name but just a few. Also, the MapReduce implementation has been adapted to computing environments other than traditional clusters, for example, multi-core systems [34] [17], desktop grids [42], volunteer computing environments [23], dynamic cloud environments [25], and mobile environments [7].

Along with the evolution of MapReduce, great progress has also been made with hardware. Nowadays it is common for commodity clusters to have processors of more and more in-chip cores (referred to as **many-core cluster** hereafter) [36] [26]. While MapReduce scales well with the increase of server number, its performance improves less or even remains unchanged with the increase of CPU cores per server. Fig. 1 shows the execution time of *self-join* with varying number of CPU cores per server on a 7-node many-core cluster, where the line with pluses denotes the time taken by Hadoop and the line with squares denotes the time in an ideal world. As the number of CPU cores increases, the gap between the two lines gets wider and wider.

The fundamental reason (refer to Section 2 for a detailed analysis) behind this is that the underlying low-speed HDD storage cannot meet the requirements of MapReduce frequent IO operations: in the Map phase, the model reads raw input data to generate sets of intermediate key-value pairs, which are then written back to disk; in the Shuffle phase, the model reads the intermediate data out from the disk once again and sends it to the nodes to which Reduce

- *B. Wang and G. Yang are with the Department of Computer Science and Technology, Tsinghua National Laboratory for Information Science and Technology, Ministry of Education Key Laboratory for Earth System Modeling, Center for Earth System Science, Tsinghua University, Beijing 100084, China.*
  E-mail: bo-wang11@mails.tsinghua.edu.cn, ygw@tsinghua.edu.cn
- *J. Jiang and Y. Wu are with 1) the Department of Computer Science and Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China, 2) Research Institute of Tsinghua University in Shenzhen, Shenzhen 518057, China, and 3) Technology Innovation Center at Yinzhou, Yangtze Delta Region Institute of Tsinghua University, Ningbo 315000, China.*
  E-mail: {jjlei, wuyw}@tsinghua.edu.cn
- *K. Li is with the Department of Computer Science, State University of New York at New Paltz, New York 12561, USA.*
  E-mail: lik@newpaltz.edu

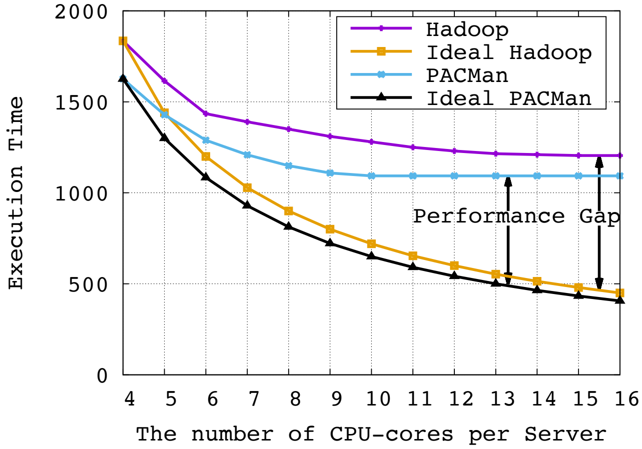


Fig. 1. Execution time of *self-join* with varying number of CPU cores per server using the settings in Section 4. The Input Data is of 60GB.

tasks are scheduled. In addition, during the whole execution of jobs, temporary data is also written to local storage when memory buffer is full. Although more tasks can run concurrently in theory as more CPU cores are equipped, the IO speed of the storage system which backs MapReduce remains unchanged and cannot meet the IO demand of high-concurrent tasks, resulting in slightly improved or even unchanged MapReduce performance.

Indeed, the IO bottleneck of hard disk has long been recognized and many efforts have been made to eliminate it. The research work can be roughly divided into two categories.

The first category tires to cache *hot* data in the memory [10] [11] [12] [32]. Since the IO speed of memory is orders of magnitude faster than that of HDD, data in memory can be manipulated more quickly. Only hot data is cached because only limited volume of memory is available due to the high cost (compared with HDD). For parallel computing, memory is also a critical resource. Many parallel computing frameworks (e.g., Apache YARN [27]) use self-tuning technology to dynamically adjust task parallelism degree (TPD, which is the number of concurrent running tasks) according to available CPU-cores and memory. Caching data in memory inevitably occupies memories and makes the available memory for normal tasks operation drop, thus reducing the TPD and the performance. For memory-intensive machine-learning algorithms such as *k-means* and *term-vector*, which consume very large volume of memory during execution, the thing would get even worse — their TPD would drop significantly due to reduced memory for normal operation, leaving some CPU cores idle. Fig. 1 also illustrates this point by the case of PACMan [2], which is the latest work that utilizes memory caching to improve MapReduce. Although adding more memory could alleviate the situation, the volume of data grows even faster, meaning more memory is needed to cache data to get the benefit. Taking cost into consideration, it is not cost-effective to improve IO speed by in-memory caching.

The second category tries to use new storage medium of high-IO speed to replace HDD [24] [48] [22] [21]. Flash-based SSD is such a most popular storage medium. Since SSD does not have mechanical components, it has lower access time and less latency than HDD, making it an ideal storage medium for building high performance storage systems. However, the cost of building a storage system totally with SSDs is often the budget of most commercial data centers. Even considering the trend of SSD price dropping, the average per GB cost of SSD is still unlikely to reach the level of hard disks in the near future [16]. Thus, we believe using SSD as a cache of hard disks is a good choice to improve IO speed as did in [4], [35], [33], and [19].

1.2 Our Contributions

Taking both performance and cost into consideration, this paper presents mpCache (a preliminary version has been published in [44]), a solution that tries to accelerate MapReduce on commodity clusters via SSD-based caching. mpCache not only boosts the speed of storage system (thus eliminating the IO bottleneck of HDD) for IO-intensive applications but also guarantees TPD of memory-intensive jobs. The contributions of our paper are as follows.

- We identify the key cause of the poor performance of MapReduce applications on many-core clusters as the underlying low-speed HDD storage system cannot afford high concurrent IO operations of MapReduce tasks.
- We propose mpCache, an SSD-empowered cost-effective cache solution that caches both Input Data and Localized Data of MapReduce jobs in SSD to boost IO operations. In order to get the best benefit of caching, a mechanism is also put forward to dynamically adjust the SSD allocation between Input Cache and Localized Cache.
- We present a cache replacement scheme that takes into consideration not only replacement cost, data set size, and access frequency, but also the all-or-nothing characteristic of MapReduce caching [2].
- Extensive experiments are conducted to evaluate mpCache. The experimental results shows that mpCache can get an average speedup of 2.09x when compared with standard Hadoop and an average speedup of 1.79x when compared with PACMan.

The rest of this paper is organized as follows. Section 2 gives a brief introduction to MapReduce and analyzes the reasons why MapReduce applications perform poorly on many-core clusters. Section 3 describes the key ideas and algorithms of mpCache. Section 4 shows the experimental results. Section 5 reviews the related work and the paper ends in Section 6 with some conclusions.

2 Problem Analysis

In this section, we first give a brief introduction to MapReduce, and then set out to find out the bottleneck of MapReduce applications on many-core clusters.

2.1 Overview of MapReduce

A MapReduce [6] program is composed of a Map function and a Reduce function, where the Map function is used to process the key-value pairs associated with the input data (supplied by a certain distributed file system or database) to generate a set of intermediate key-value pairs and the Reduce function is used to merge all intermediate values associated with the same intermediate key. The program is executed by a runtime, the core part of a MapReduce framework that is in charge of such things as reading in and partitioning the input data, scheduling tasks across the worker nodes, monitoring the progress of tasks execution, managing all the communications between nodes, tolerating the fault encountered, and so on.

There are many MapReduce frameworks available and in this paper, we base our prototype on YARN, the latest version of Apache Hadoop, which is probably the most popular open-source implementation of MapReduce model.

The execution of a MapReduce program consists of three phases, that is, the Map phase, the Shuffle phase, and the Reduce phase. Fig. 2 shows the execution of a MapReduce job from the perspective of IO operations. Details are as follows.

In the Map phase, each map task **Read**s in the data block (from the source specified by the job), and runs the user-providing Map function to generate some key-value pairs (called intermediate results) that are stored first in a memory buffer and then flushed to local disk as a file (called **spill** file in Hadoop) when the buffer runs out. The spill procedure repeats until the end of the Map task, generating multiple spill files. After that, spill files of the same Map task are **Merge**d (denoted by **M** in the figure) into a single file and written back to local disks for the purpose of fault-tolerance.

In the Reduce phase, the reduce task first **Fetch**es input data from all the Map nodes and **Merge**s (denoted by **M** in the figure) the fetched data into a single file. Then the user-providing Reduce function is executed to process the data. Since all the temporary results of **Spill** and **Merge** procedures and the outputs of Map function are written to local storage, they are called **Localized Data** in Hadoop.

Between Map phase and Reduce phase is the Shuffle phase that is employed to sort the Map-generating results by the key and pipeline data transfer between Mappers and Reducers. Since Reduce tasks in a MapReduce job will not execute until all Map tasks finish, the pipelining mechanism in the Shuffle phase would save a large part of data transfer time and thus improve performance.

All the three phases involve IO operations multiple times. For example, disk manipulation occurs two times (reading data from and writing data to disks) in the Map phase, while during the Shuffle phase, disk operations will happen at both the Mapper and the Reducer sides — data is read out from the disks of the Mappers, sent over the network, and then written to the disks of the Reducers. Since the speed of hard disks cannot match that of CPU, IO operations are time-consuming and thus limit tasks throughput. With IO operations improved by SSD-based caching at both Mappers and Reducers, the computation process will be accelerated accordingly. That is the basic idea behind our work here.

2.2 Bottleneck Analysis

With the development of hardware technology, many-core servers get common in data centers [36] [26]. For example, each server in our experiment has 16 CPU cores. More cores on a node usually means the node could process more tasks concurrently. In a typical Hadoop configuration, one CPU core corresponds to one Map/Reduce task. Thus, one node could run concurrently as many tasks as the number of CPU cores in theory. We define the number of concurrently running tasks (i.e., TPD) as *wave-width* since tasks in MapReduce are executed wave by wave. Then we have $wave\# = ceil(tasks\#/wave\text{-}width)$. Obviously, the bigger the *wave-width*, the smaller the *wave#* and the shorter the job execution time.

We examines the job execution time by varying the wave-width. As shown in Fig. 1, the execution time of the job reaches the minimum value when the wave-width is 12 and this value remains unchanged even if the wave-width increases. Consider a job consisting of 288 Map tasks and running on a many-core cluster of 6 nodes. Obviously, each node should process $288/6 = 48$ Map tasks. When each node is equipped with 12 CPU cores, the number of concurrently running Map tasks (i.e., the *wave-width*) is 12. In this case we get $48/12 = 4$ waves for the job. On the contrary, when each node is equipped with 16 CPU cores, we get *48/16=3* waves for the same job. Ideally, if the node could provide sufficient resources such as CPU, memory, and IO, job execution in 3 waves should take 3/4 time of that running in 4 waves. But as shown in Fig. 2, the Map time remains unchanged when the wave-width increases from 12 to 16. Please note that the execution time of different waves might be different in the real world and here we just use Fig. 2 to simplify the problem description.

The reason for unchanged execution time is that the IO-intensive operations (i.e., **Read**, **Spill**, **Merge**) slow down due to the IO bottleneck of the underlying storage system. For commodity clusters, they are usually equipped with HDDs. Since the MapReduce job performance is bounded by the low IO speed of HDD, it is natural that the performance remains unchanged with the increase of CPU cores. This phenomenon was also reported by PACMan [2] — the authors found that *"the client saturated at 10 simultaneous tasks and increasing the number of simultaneous tasks beyond this point results in no increase in aggregate throughput"*.

In summary, the bottleneck of MapReduce applications running on many-core clusters is the IO speed of storage system. As mentioned in Section 1, caching data in memory and building total SSD storage systems have several disadvantages, impeding them to be used for memory-intensive applications. Therefore, we propose mpCache, an SSD-based cache solution that caches both Input Data and Localized Data to provide high IO speed and thus speed up all the critical operations — **Read**, **Spill**, and **Merge**. Besides, mpCache also allows dynamically adjusting the space between Input Cache and Localized Cache to make full use of the cache to get the best benefit.

3 mpCache Design

This section details the design of mpCache.

3.1 Architecture

In accordance with the distributed file system that backs the MapReduce framework up, mpCache adopts a master-slave

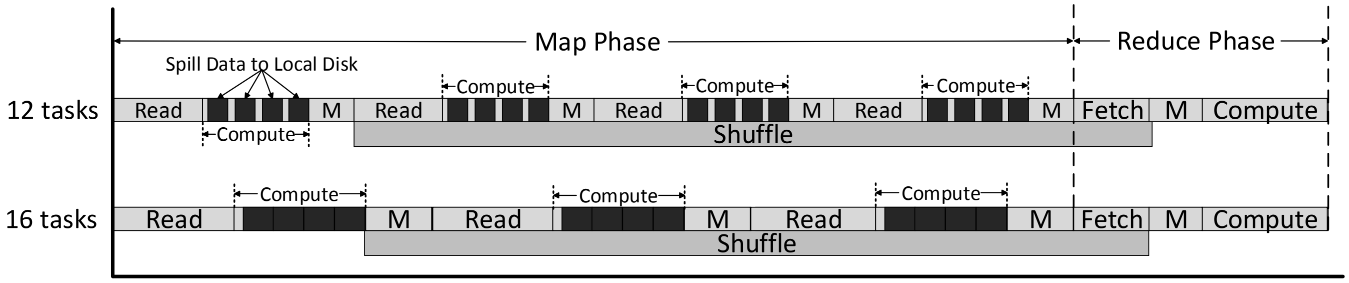


Fig. 2. Diagrammatic sketch of MapReduce performance issues with different concurrent tasks. **M** in the figure denotes **Merge**.

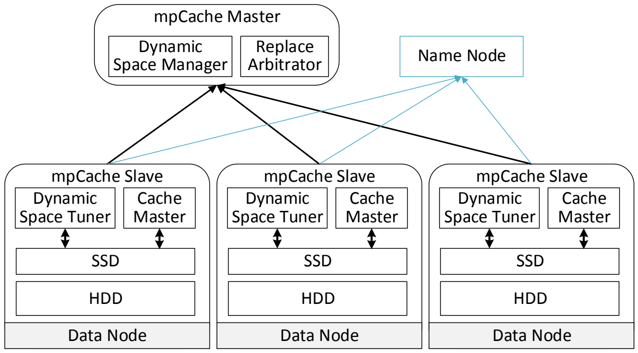


Fig. 3. mpCache architecture. It adopts a master-slave architecture with mpCache Master managing mpCache Slaves locating on every data node. Thin lines represent control flow and thick arrows denote data flow. Such an architecture is in accordance with that of the underlying distributed file system that backs up the MapReduce framework.

architecture, as shown in Fig. 3, with one mpCache Master and several mpCache Slaves. mpCache Master acts as a coordinator to globally manage mpCache slaves to ensure that a job's input data blocks, which are cached on different mpCache slaves, present in an all-or-nothing manner, for some prior research work [2] found that a MapReduce job can only be speeded up when inputs of all tasks are cached. We can see from the figure that SSD-based cache space locates in each data node of the underlying distributed file system of the MapReduce framework. It is a distributed caching scheme.

mpCache Master consists of two components – *Dynamic Space Manager* and *Replace Arbitrator*. *Dynamic Space Manager* is responsible for collecting the information about dynamic cache space allocation from each mpCache Slave and recording into history the job type and input data set size. *Replace Arbitrator* leverages the cache replacement scheme.

mpCache Slave locates on each data node and also consists of two components, that is, *Dynamic Space Tuner* and *Cache Master*. *Dynamic Space Tuner* is deployed to adjust the space allocation between Input Cache (for caching Input Data) and Localized Cache (for caching Localized Data). *Cache Master* is in charge of serving cached data blocks and caching new ones.

During job execution, *Cache Master* on each data node intercepts the data reading requests of Map tasks, and checks whether the requested data block is cached. If so, *Cache Master* servers the data blocks from cache and informs *Replace Arbitrator*, which resides with mpCache Master, of the block hit. Otherwise, the data block will be cached. In the case that there is no enough cache space, *Cache Master* will send cache replacement request to *Replace Arbitrator* and do cache replacement according to the information returned by *Replace Arbitrator* to make room for the newly requested data block.

3.2 Dynamic Space Allocation

As described in Section 2.1, **Read**, **Spill**, and **Merge** are all IO-intensive procedures, imposing restrictions on performance when running on many-core clusters. Since both reading in the job input data and reading/writing **Localized Data** involve IO operation, mpCache caches both data. Because the cache space is limited and different jobs may have different characteristics in terms of input data size and localized data size, we must smartly allocate the space between Input Cache and Localized Cache to get the best benefit. Fig. 4 illustrates this, where the *x-axis* represents the Localized Cache size and *y-axis* represents the total benefit of caching.

As shown in the figure, the Input Cache size as well as the corresponding caching benefit decreases as the Localized Cache size increases. With a larger Localized Cache, the cost of writing/reading Localized Data reduces and the cache performance improves. At a certain point, the two lines cross and total benefit of caching reaches the best value. Please note that this figure is just an illustration. In the real world, the optimal point may vary between jobs, for different jobs may produce quite different volumes of Localized Data, according to which jobs can be categorized into *shuffle-heavy*, *shuffle-medium*, and *shuffle-light*. Therefore, we must dynamically adjust the space allocation to ensure the best benefit of caching.

It is in this sense that *Dynamic Space Tuner* is introduced. As shown in Fig. 5, *Dynamic Space Tuner* divides the whole cache space into three parts, that is, Input Cache, Dynamic Pool, and Localized Cache. Since the distributed file systems (e.g., GFS [14] and HDFS [39]) that back MapReduce applications up store data in the unit of block, we also divide Dynamic Pool into many blocks. Blocks in Dynamic Pool will be used on demand as Input Cache or Localized Cache. During job execution, *Dynamic Space Tuner* constantly monitors the utilization of the Localized Cache. When the

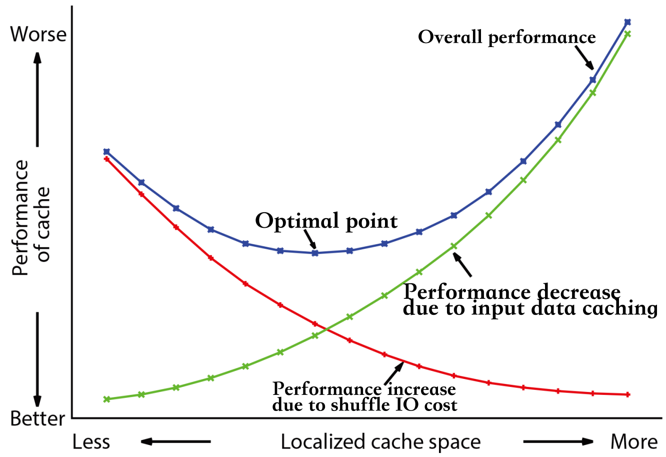


Fig. 4. Balancing the size of the Input Cache and the Localized Data Cache is necessary to get best benefit of caching.

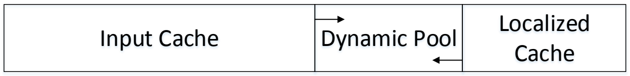


Fig. 5. The whole cache space is divided into three parts, namely Input Cache, Dynamic Pool and Localized Cache. Blocks in Dynamic Pool are used on demand as Input Cache or Localized Cache depending on the workload to get the most benefit of caching.

cache space runs out, *Dynamic Space Tuner* checks if there are free blocks in Dynamic Pool. If not, *Dynamic Space Tuner* will remove some cached input data from Dynamic Pool using the same scheme described in Section 3.3. Then the just freed blocks are used as Localized Cache one by one. If the utilization of Localized Cache is below the *guard value*, which is set to 0.5 in our implementation, all blocks used as Localized Cache in Dynamic Pool are reclaimed.

3.3 Input Data Cache Model

Since the cache size is limited, it is necessary to do cache replacement to guarantee the desired benefit. Here we explain the cache model used for input data.

3.3.1 Admission Control Policy

We use an admission control policy in the first place to decide whether or not an object should be cached. Since the cache space is limited and input data size varies job by job, caching input data of one job may mean purging the data of the other jobs from the cache. Too frequent cache replacement may result in the case that some cached data will never be used during the whole lifetime in the cache, reducing the benefit of caching. It is the duty of admission control policy to avoid this happening.

The admission control policy utilizes an auxiliary facility to maintain the *identities* of input data sets of different jobs. For each data set recorded in this facility, its access number and the last access time are also maintained. Each time the data set is accessed, the corresponding access number is increased by 1 and the record is updated. The auxiliary facility is kept in memory, for it just maintains metadata about the data sets rather than the data sets themselves and will not consume too much memory.

Using the admission control policy, we would like to ensure that, at a certain time when some data is accessed, the potential incoming input data jd_i gets popular enough so that it can be loaded into the cache to get more benefit. The process is as follows.

If there is enough free space for jd_i, we simply load jd_i into the main cache. Otherwise, we check to see if jd_i has been recorded by the auxiliary facility. If not, we will record the related information with the auxiliary facility rather than put jd_i itself into the main cache. In the case that jd_i does occur in the auxiliary facility, we proceed to see if some cache replacement is necessary. By necessary we mean the cache replacement is profitable, or in other words, it can bring in some benefit for the performance. This is done by comparing the value $1/(Size(jd_i)\Delta_{jd_i})$ with the sum $\sum_j 1/(Size(jd_j)\Delta_{jd_j})$, where jd_j is the candidate data sets to be replaced that is determined by the replacement scheme described in Section 3.3.2, $Size(jd)$ is the number of blocks in data set jd, and Δ_{jd} is the data set access distance, which is defined as the number of data set accesses between the last two times that the data set jd was accessed. In the case that jd is accessed for the first time, Δ_{jd} is defined as the number of all data set accesses before that, and in the case of successive accesses, Δ_{jd} is set to 0.01. A candidate data set jd_i can be loaded into the main cache if and only if $1/(Size(jd_i)\Delta_{jd_i}) > \sum_j 1/(Size(jd_j)\Delta_{jd_j})$. It is easy to see that the data set access distance defined in such a way ensures those data sets being frequently accessed (and thus, of smaller data set access distance) have greater chance to be loaded into the main cache.

3.3.2 Main Cache Replacement Scheme

We now describe the cache replacement scheme adopted by the main cache. For each data set in the main cache we associate it with a *frequency* $Fr(jd)$, which is the number of times that jd has been accessed since it was loaded into the main cache. Besides, a priority queue is maintained. When a data set of a certain job is inserted into the queue, it is given the priority $Pr(jd)$ using the following way:

$$Fr(jd) = Blocks\_Access(jd)/Size(jd) \tag{1}$$

$$Pr(jd) = Full + Clock + Fr(jd) \tag{2}$$

where *Blocks_Access(jd)* is the total number of times that all blocks of data set *jd* are accessed; *Full* is a constant *bonus* value assigned to the data set whose blocks are all in the main cache (in favor of the *all-or-nothing* characteristic of MapReduce cache [2]); *Clock* is a variable used by the priority queue that starts at 0 and is set to $Pr(jd_{evicted})$ each time a data set $jd_{evicted}$ is replaced.

Once the mpCache Master receives a data access message from an mpCache Slave, Algorithm 1 is used to update $Pr(jd)$ of the corresponding data set indicated by the message. Since *Clock* increases each time a data set is replaced and the priority of a data set that has not been accessed for a long time was computed using an old (hence small) value of *Clock*, cache replacement will happen on that data set even if it has a high *frequency*. This "aging" mechanism avoids the case that a once frequently-accessed data set, which will never be used in the future, unnecessarily occupies the cache and thus degrades performance. To_Del in Algorithm 1 is a list of tuples that have the format $< data\_node, blocks_{evicted} >$. It is introduced for *Replace Arbitrator* to record those data blocks on each data node that have already been selected by *Replace Arbitrator* as outcasts but the corresponding *Cache Master* is not notified of.

Algorithm 1 Main Cache Replacement Scheme.

```
if the requested block $bk$ is in the cache then
    $jd \leftarrow$ the data set to which $bk$ belongs
    $Blocks\_Access(jd) \leftarrow Blocks\_Access(jd)$+1
    update $Pr(jd)$ using Equation (1)-(2) and move $jd$ accordingly in the queue
else
    if no cache replacement is necessary then
        cache $bk$
    else
        $mpSlave \leftarrow$ the source of the data access request
        $data\_node \leftarrow$ the data node that $mpSlave$ is seated
        if $To\_Del$.hasRecord($data\_node$) then
            send $blocks_{evicted}$ to $mpSlave$, and replace $blocks_{evicted}$ with $bk$ at $mpSlave$
        else
            $jd_{evicted} \leftarrow$ the data set with lowest priority in the queue
            $Clock \leftarrow Pr(jd_{evicted})$
            $blocks_{evicted} \leftarrow$ all the blocks of $jd_{evicted}$
            send $blocks_{evicted}$ to $mpSlave$, and replace $blocks_{evicted}$ with $bk$ at $mpSlave$
            $allnodes \leftarrow$ all the data nodes that store $blocks_{evicted}$
            for $dn \in allnodes$ do
                $To\_Del$.addRecord($< dn, blocks_{evicted} >$)
            end for
        end if
    end if
    $Blocks\_Access(jd) \leftarrow Blocks\_Access(jd)$+1
    if all the blocks of $jd$ are cached then
        $Full = BONUS\_VALUE$
    else
        $Full = 0$
    end if
    compute $Pr(jd)$ using Equation (2) and put $jd$ into the queue accordingly
end if
```

4 Evaluation

We implement mpCache by modifying Hadoop distributed file system HDFS (version 2.2.0) [15] and use YARN (version 2.2.0) to execute the benchmarks.

4.1 Platform

The cluster used for experiments consists of 7 nodes. Each node has two eight-core Xeon E5-2640 v2 CPUs running at 2.0GHz, 20MB Intel Smart Cache, 32GB DDR3 RAM, one 2TB SATA hard disk and two 160GB SATA Intel SSDs configured as RAID 0. All the nodes run Ubuntu 12.04, have a Gigabit Ethernet card connecting to a Gigabit Ethernet switch. Though we have *160*2=320GB* SSD on each node, we only use 80GB as cache in our experiment to illustrate the benefit of mpCache. Such a value is selected because the data sets used for experiments are not large (the maximum volume of data manipulated during our experiments is about 169GB in the case of *tera-sort*) and too large cache space would hold all data, making cache replacement unnecessary. In the real world, the input data sets of MapReduce may be of terabytes or even petabytes, well beyond the SSD capacity.

TABLE 1
Input data size of benchmarks. (k=1,2,. . . ,20)

Data Source	Data Size	Benchmarks
wikipedia	k*4.3G	grep
		word-count
		inverted-index
		term-vector
		sequence-count
netflix data	k*3.0G	histogram-rating
		histogram-movies
		classification
		k-means
PUMA-I	k*3.0G	self-join
PUMA-II	k*3.0G	adjacency-list
PUMA-III	k*4.2G	ranked-inverted-index
PUMA-IV	k*3.0G	tera-sort

4.2 Benchmarks

We use 13 benchmarks released in PUMA [1], covering shuffle-light, shuffle-medium, and shuffle-heavy jobs. We vary the input data size of each benchmark from 1 to 20 times of the original data set. Input data size of each benchmark is shown in Table 1. *grep*, *word-count*, *inverted-index*, *term-vector*, and *sequence-count* use the same input data, which is a text file downloaded from wikipedia. *histogram-rating*, *histogram-movies*, *classification*, and *k-means* use the same data set, which is classified movie data downloaded from Netflix. *self-join*, *adjacency-list*, *ranked-inverted-index*, and *tera-sort* use data set downloaded from PUMA.

Since the input data size has Zipf-like frequency distribution [20], we associate a probability with each data size using Equation (3).

$$f(k;s,N) = \frac{1/k^s}{\sum_{i=1}^{N} 1/i^s} \tag{3}$$

Since 20 times of data size are generated, we set N to 20. For the Zipf parameter s, we set it to 1 if not specially mentioned. Table 2 summarizes the characteristics of the benchmarks in terms of input data size (take *k=10* for example), data source, the number of Map/Reduce tasks, shuffle size, and execution time on Hadoop.

Shuffle-light jobs, including *grep*, *histogram-ratings*, *histogram-movies*, and *classification*, have very little data transfer in the shuffle phase. Shuffle-heavy jobs, which have a very large data size to be shuffled (as shown in Table 2, almost the same as the input data size), include *k-means*, *self-join*, *adjacency-list*, *ranked-inverted-index*, and *tera-sort*. The shuffle data size of shuffle-medium jobs is between that of shuffle-light and shuffle-heavy ones, including *word-count*, *inverted-index*, *term-vector*, and *sequence-count*.

When submitting a job to the cluster, we randomly select one from the 13 benchmarks, and set the input data size according to the attached probability. Each time we submit a job, we use ”echo 1 > /proc/sys/vm/drop_caches” command to clear memory cache and make sure the data is read from mpCache other than memory.

4.3 Experimental Results

Our experiment consists of 5 parts: i) Section 4.3.1 compares mpCache with standard Hadoop and PACMan, the state-of-the-art way of MapReduce optimization by in-memory

TABLE 2
Characteristics of the benchmarks used in the experiment

Benchmark	Input size(GB)	Data source	#Maps & #Reduces	Shuffle size(GB)	Map&Reduce time on Hadoop(s)
grep	43	wikipedia	688 & 40	$6.9 * 10^{-6}$	222&2
histogram-ratings	30	netflix data	480 & 40	$6.3 * 10^{-5}$	241&5
histogram-movies	30	netflix data	480 & 40	$6.8 * 10^{-5}$	261&5
classification	30	netflix data	480 & 40	$7.9 * 10^{-3}$	286&5
word-count	43	wikipedia	688 & 40	0.318	743&22
inverted-index	43	wikipedia	688 & 40	0.363	901&6
term-vector	43	wikipedia	688 & 40	0.384	1114&81
sequence-count	43	wikipedia	688 & 40	0.737	1135&27
k-means	30	netflix data	480 & 4	26.28	450&2660
self-join	30	puma-I	480 & 40	26.89	286&220
adjacency-list	30	puma-II	480 & 40	29.38	1168&1321
ranked-inverted-index	42	puma-III	672 & 40	42.45	391&857
tera-sort	30	puma-IV	480 & 40	31.96	307&481

caching; ii) Section 4.3.2 compares mpCache with traditional cache replacement policies such as LRU (Least Recently Used) and LFU (Least-Frequently Used); iii) Section 4.3.3 shows mpCache behavior with different numbers of CPU cores per server; iv) Section 4.3.4 shows the adaptability of mpCache to the cache size; v) Section 4.3.5 shows the adaptability of mpCache to the Input Data size.

4.3.1 Comparison with Hadoop and PACMan

We compare the execution time of benchmarks on mpCache with that on both Hadoop and PACMan. We run the benchmarks with mpCache, Hadoop, and PACMan respectively and get the average value. PACMan uses memory to cache input data and the bigger the cache size, the more the cached data and thus the faster the Map phase. However, the concurrent running tasks number in YARN is tightly related to the available CPU cores and the free memory, and consuming too much memory for data caching would decrease the parallelism degree of the tasks. We set the volume of memory used for cache to 12GB as did in PACMan [2].

Fig. 6 shows the normalized execution time of the Map and Reduce phase. For shuffle-light jobs such as *grep*, *histogram-movies*, *histogram-ratings*, and *classification*, their execution time is short (about 241s, 253s, 279s, and 304s on Hadoop when *k=10*) and most time is spent on data IO. Input data caching supplied by mpCache can accelerates the Map phase significantly (2.42x faster on average). In the Reduce phase, the speedup is not notable for three reasons: i) The Reduce phase of shuffle-light jobs is very short (about 2s, 4s, 4s, and 5s when *k=10*); ii) Shuffle-light jobs have very little shuffle data (less than 10 MB); iii) The localized data size is so small (less than 1 MB) that caching localized data results in little acceleration. In all, mpCache gets a speedup of 2.23 times over Hadoop for shuffle-light jobs. When running the jobs with PACMan, each task performs well with 1GB memory. PACMan and mpCache get the same parallelism degree of the tasks. Although in-memory caching could provide faster IO than SSD-based caching as mpCache does, the larger cache size provided and cache replacement scheme supplied ensure **a higher hit ratio** of mpCache than that of PACMan does (61.7% vs. 38.5%). Therefore, mpCache performs even better than PACMan.

For shuffle-medium jobs such as *word-count*, *inverted-index*, *term-vector*, and *sequence-count*, their execution time is longer than that of shuffle-light jobs(about 779s, 932s, 1209s, and 1174s), caching Map input data only results in a speedup of 1.25 times averagely. The shuffle data size of these jobs is about 318∼737MB; the size of localized data is 1∼3GB; caching localized data would produce great benefit — the average speedup of the Reduce phase is 1.60 times. In all, mpCache can averagely get a speedup of 1.25 times over Hadoop for shuffle-medium jobs. With PACMan, *word-count* and *inverted-index* run well using 1GB memory and the speedup got is almost the same as in the case of mpCache. For *term-vector* tasks that need at least 3GB memory, the parallelism degree is 10 in Hadoop and 6 in PACMan. As a result, the performance of PACMan drops to 0.762 of the performance of Hadoop. The parallelism degree for *sequence-count*, whose task needs at least 2GB memory, is 16 in Hadoop and 10 in PACMan, making the performance of PACMan drop to 0.868 of the performance of Hadoop.

For shuffle-heavy jobs such as *k-means*, *self-join*, *adjacency-list*, *ranked-inverted-index*, and *tera-sort*, both the shuffle data size and the localized data size are very big. Thus, caching Map input data and localized data reduces the time of Map and Reduce phases greatly. The Map phase time of *k-means*, *self-join*, *ranked-inverted-index*, and *tera-sort* is shorter than that of *adjacency-list* (1168s). Thus the speedup got for the former three jobs is 1.82∼2.69 times, whereas the speedup got for the latter job is 1.04 times. Caching localized data also brings in great benefit — a speedup of 3.87 times would be got in the Reduce phase. In all, mpCache results in an average speedup of 2.65 times over Hadoop. For PACMan, the parallelism degree with *self-join*, *adjacency-list*, *ranked-inverted-index*, and *tera-sort*, each task of which needs 2 GB memory, is 10, resulting in the performance of PACMan dropping to 0.981 of the performance of Hadoop. As for *k-means*, the number of Reduce tasks is set to 4 (because it clusters the input data into 4 categories) and each task needs at least 8GB memory. Since less memory is left for normal operation, PACMan spends 2.46x longer time in the Map phase than Hadoop does. In addition, it does no help to the heavy Reduce phase (2660s, taking about 86.2% of the whole job execution time). As a result, the performance of PACMan drops to 0.808 of the performance of Hadoop.

PACMan used 12GB memory for data cache and got considerable performance improvement over Hadoop MapReduce v1 [15], the TPD of which is determined by the ”slots”

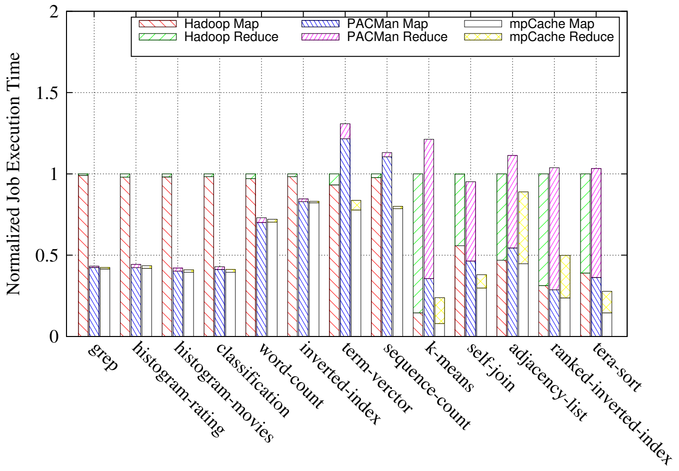


Fig. 6. Job execution time comparison with Hadoop and PACMan.

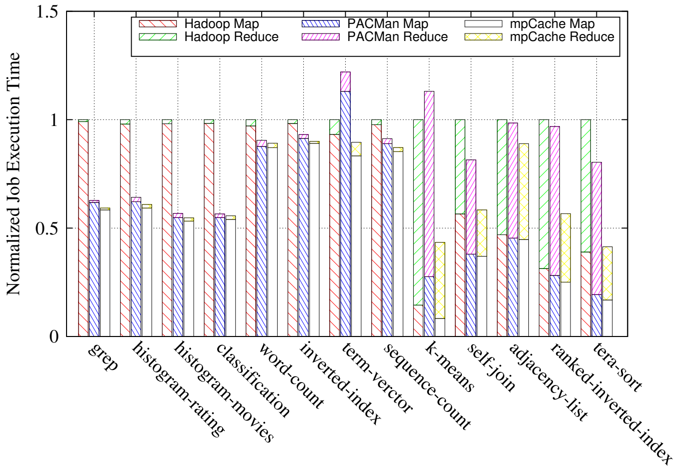


Fig. 7. Job execution time comparison with Hadoop and PACMan on the same cluster of 8 CPU cores.

number in the configuration file. Usually it is set to a constant value. Since both Hadoop and PACMan use the same configuration, they are of the same TPD. However, in MapReduce v2 (i.e., YARN [27]), the number of concurrent running tasks is determined by the number of free CPU cores and free memory, allocating memory for data cache inevitably reduces the TPD of some jobs.

In our cluster, each node has 16 CPU cores and 32GB memory. Since PACMan used 12GB memory for cache, the memory left for computing is 20GB. When running ”1GB jobs” (jobs with each task consuming 1GB memory, including *grep*, *histogram-rating*, *histogram-movies*, *classification*, *word-count*, and *inverted-index*) with PACMan, the TPD is 16, the same as that of Hadoop and mpCache. Therefore, PACMan gets a better performance than Hadoop and mpCache performs almost the same as PACMan. For other jobs, each task needs at least 2GB memory (3GB for *term-vector*, and 6GB for *k-means*), and therefore the TPD of PACMan drops to 10 (6 of *term-vector*, and 3 of *k-means*). Although in-memory caching could significantly speedup the Map phase, the dropping of TPD slows down the job worse: as illustrated in Fig. 6, PACMan performs even worse than Hadoop for these ”at least 2 GB” jobs.

For all these benchmarks, mpCache gains an average speedup of 2.09x when compared with the Hadoop, and an average speedup of 1.79x when compared with PACMan. Such improvements come from the speedup of IO operations. Since more data is read from the SSD-based cache rather than hard disks, computing resources waste due to lack of data is lowered and thus tasks can progress faster. Though the speed of SSD is slower than that of memory, the volume of SSD cache is much larger than that of memory cache. As a result, SSD-based cache also shows advantage over memory-based cache. This is why mpCache performs better than PACMan.

In order to better illustrate the in-memory caching effect of PACMan, we also do an experiment where only 8 CPU cores are used on each node for Hadoop, PACMan, and mpCache.

As shown in Fig. 7, for the case of 8 CPU cores, most benchmarks can run with the same TPD on Hadoop, mpCache, and PACMan except *term-vector* and *k-means*. For shuffle-light jobs, mpCache and PACMan run with the same TPD, getting 1.74x and 1.67x speedup over Hadoop respectively. For shuffle-medium jobs, in the 1GB job case (*word-count* and *inverted-index*), the speedup got over Hadoop is 1.12x and 1.08x respectively; in the 3GB job case (*term-vector*), Hadoop and mpCache run with TPD=8 whereas PACMan runs with TPD=6. Thus PACMan has a longer Map phase time than Hadoop and the whole performance of PACMan is even worse than that of Hadoop. For shuffle-heavy jobs, the localized data size is also big. mpCache caches both input data and localized data, resulting in an average speedup of 1.63 times in the Map phase and 2.09 times in the Reduce phase. In contrast, PACMan gets an average speedup of 1.35 times in the Map phase and introduces no benefit in the Reduce phase. Totally, for all the benchmarks, mpCache gets an average speedup of 1.62 times, whereas PACMan gets an average speedup of 1.25 times.

4.3.2 Comparison with Traditional Cache Replacement Policies

We implement two traditional cache replacement policies, namely LRU and LFU. In our settings, mpCache gets an average hit ratio of 61.7%, while LRU gets an average hit ratio of 53.9% and LFU gets an average hit ratio of 68.3%. The resulted performance is shown in Fig. 8. Although LFU gets a higher hit ratio than mpCache does, mpCache takes all-or-nothing characteristic of MapReduce caching into consideration and deploys an auxiliary facility to prevent too frequent replacements, and therefore gets a higher speedup than LFU does. Compared with LRU, mpCache gets both higher hit ratio and speedup. With the cache space better utilized, it is natural that the IO operations and the consequent tasks execution are speeded up.

4.3.3 Adaptability to the Number of CPU Cores per Server

Fig. 9 shows mpCache’s adaptability to the number of CPU cores per server, where the line with pluses denotes the execution time of Hadoop, the line with squares denotes

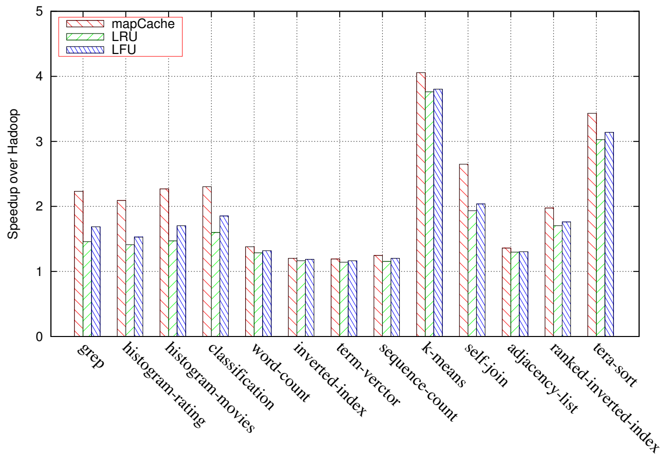


Fig. 8. Performance comparison with LRU and LFU.

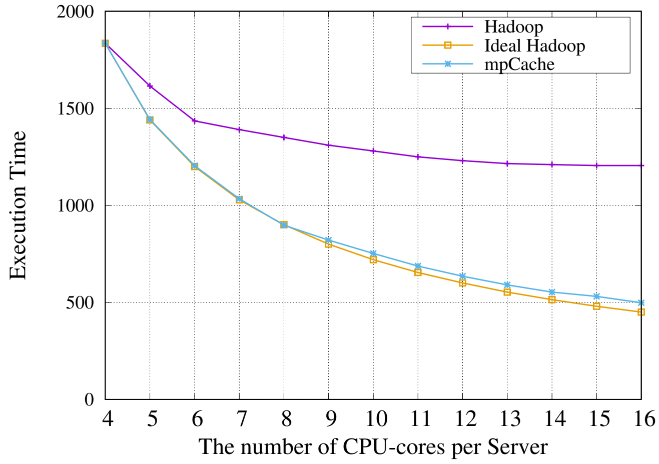


Fig. 9. Execution time of *self-join* varies with the number of CPU cores per server changing.

the execution time of mpCache, and the line with asterisks denotes the execution time of Hadoop in an ideal world (i.e., with no constraint). mpCache scales well when the number of CPU cores per server increase. Its behavior is almost the same as the ideal case.

4.3.4 Adaptability to Cache Size

We now evaluate mpCache's adaptability to cache size by varying the available cache size of each mpCache Slave between 5GB and 160GB. The experimental results are shown in 3 sub-figures, i.e., Fig. 10(a), Fig. 10(b), and Fig. 10(c), in accordance with the 3 categories of benchmarks.

Fig. 10(a) shows the impact of cache size on *shuffle-light* benchmarks. All these benchmarks have very little shuffle date and very short Reduce phase (the Reduce phase is no greater than 2.1% of the whole time). Therefore, the Localized Cache occupies less space and most space is used as Input Space. The speedup of these benchmarks mainly comes from Input Data caching. When the cache size is 5GB per node, the speedup is very small due to insufficient space to hold Input Data. As the cache size increases, the speedup grows significantly and a maximum value is obtained when the cache size is about 90GB.

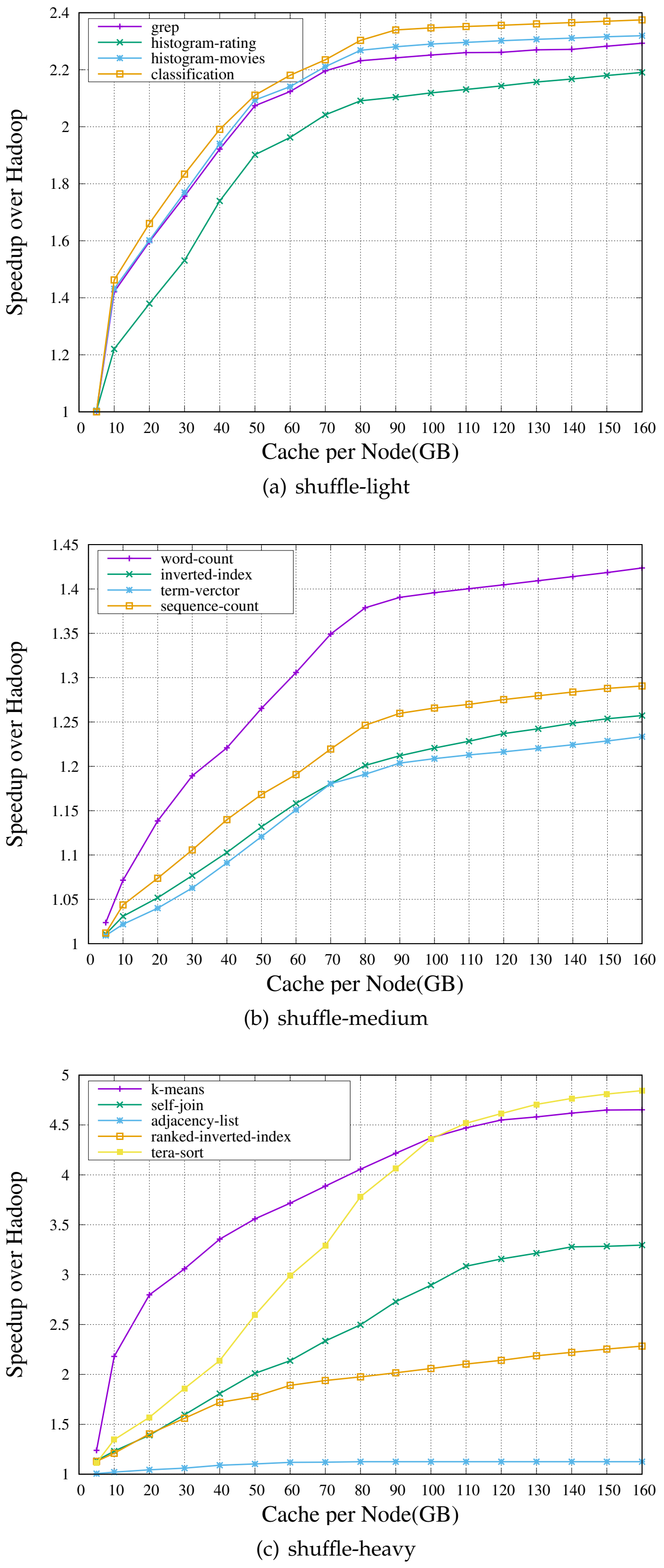
Fig. 10(b) shows the impact of cache size on *shuffle-medium* benchmarks. These benchmarks have some volume of shuffle data (no more than 1GB), both Map and Reduce phase could be accelerated by caching. When the cache size per node is 5GB, all Localized Data is cached, resulting in an average speedup of 59.99% in the Reduce phase. However, since the Reduce phase only takes 3.43% of the whole time, this only contributes 1.40% of the whole job speedup. As the cache size increases, the speedup grows due to the reduction of the Map phase time and a maximum value is reached when the cache size is about 100GB.

Fig. 10(c) shows the impact of cache size on *shuffle-heavy* benchmarks. These benchmarks have very large volume of shuffle data. When *tera-sort* runs with 30GB input data, the localized data occupies as large as 32GB space. Thus, when the cache size is below 40GB, most cache is allocated to cache Localized Data, which is the main contribution of the speedup. As depicted in the figure, the *k-means* job gets higher speedup than *tera-sort* does when the cache size is below 100GB and *tera-sort* gets higher speedup when the cache size is larger than 100GB. The reason behind this is: the Reduce phase of *k-means* takes a very large portion of the whole execution time (85.53%) and larger volume of Localized Data is *spilled* than the case of *tera-sort*. Therefore, caching Localized Data accelerates *k-means* faster than the case of *tera-sort*. When the cache size is below 40GB, the gradient of *k-means* is bigger than that of *tera-sort*. When the cache size is above 40GB, the increase of speedup is due to Input Data caching and the reduction of the Map phase time. Since *tera-sort* has smaller Map phase time than *k-means* (as shown in Table 2, when the input data size is 30GB, the Map phase time of *tera-sort* is 307s, while that of *k-means* is 450s), caching Input Data accelerates *tera-sort* faster than *k-means*, resulting in the same speedup at 100GB and greater speedup beyond 100GB. All the shuffle-heavy benchmarks get the maximum speedup when the cache size is about 130∼140GB. Among these benchmarks, the speedup of *adjacency-list* is smallest. The reason behind this is that both Map phase and Reduce phase are compute-intensive and take a long time. Since the critical resource of this benchmark is CPU, accelerating IO only improves the performance a little.

4.3.5 Adaptability to Input Data Size

We now evaluate mpCache's adaptability to the input data size by *ranked-inverted-index*. As described in Section 4.2, we attach a selection probability to each input data size using Zipf distribution, which is indicated by parameter s in Equation (3). By varying s between 0.2 and 2, we get different distributions of input data size. Fig. 11 shows input data size distribution with varying Input Data Size Coefficient, where the X-axis represents the Input Data Size Coefficient k and Y-axis indicates the CDF (cumulative distribution function) distribution probability. It can be found that the bigger the s, the higher the probability of small Input Data Size. For example, when s=2, more than 80% of the Input Data size coefficient is below 3. In other words, more than 80% of the Input Data has a size below 12.6GB.

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/TBDATA.2016.2599933, IEEE Transactions on Big Data



(a) shuffle-light

(b) shuffle-medium

(c) shuffle-heavy

Fig. 10. The impact of cache size on mpCache performance. For shuffle-light and shuffle-medium jobs, the cache space is mainly used for caching input data. Good benefit can be got when the cache size is 80GB.

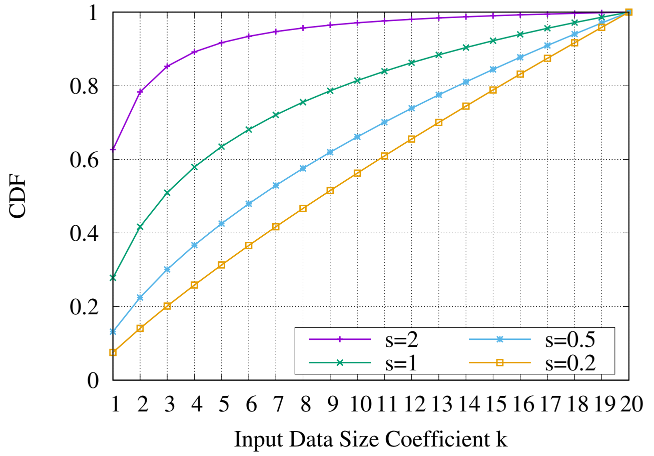


Fig. 11. Input data size distribution varies with Zipf parameter *s*. The greater the parameter, the larger the input data size.

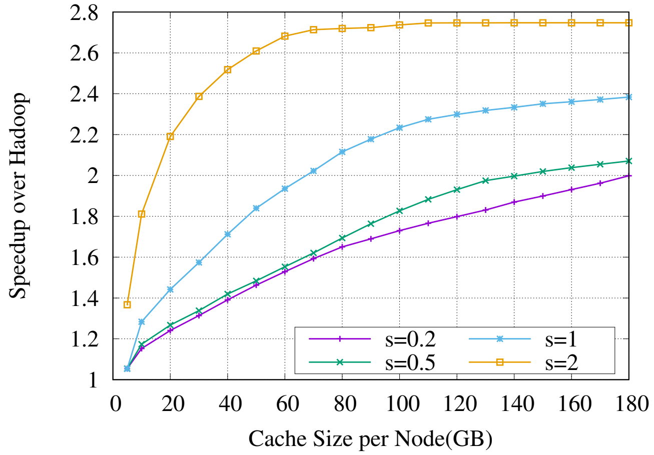


Fig. 12. The impact of Zipf parameter *s* on mpCache performance.

Fig. 12 shows the average speedups of the benchmarks with varying s. It is easy to see that mpCache works well in all cases. With the same cache size, the bigger the s, the greater the speedup (a maximum value exists as illustrated by Fig. 4). With Fig. 11 the reason behind this is obvious: a bigger s means small input data size and thus less space is needed to cache all the data to get the same speedup.

5 Related Work

There is a lot of work about MapReduce. Below is the work most related to ours.

5.1 MapReduce implementations

Due to the high impact, MapReduce, since the first release by Google [6], has been re-implemented by the open-source community [15] and ported to other environments such as desktop grids [42], volunteer computing [23], dynamic cloud [25], and mobile systems [7]. Besides, some MapReduce-like systems [45] [18] [46] [5] and high-level facilities [43] [31] [13] were proposed. In addition, MapReduce has expanded its application from batch processing to iterative computation [38] [47] and stream processing [28] [29]. Our solution can do help to these systems when hard disks are used and many cores are involved.

5.2 MapReduce Optimization on Multi-Core Servers

This can be seen in [34] [41] [17] [9] [40]. All these frameworks are designed for a single server, of which [17] [9] [40] mainly focused on graphics processors and [34] [41] were implemented on symmetric-multiple-processor server. Obviously, a single node with the frameworks could only process gigabytes of data at most and cannot afford the task of handling terabytes or petabytes of data. Besides, they still suffer from the IO bottleneck as could also be seen from Fig. 2 of [34] when the number of cores is greater than 8. Our solution is a distributed caching scheme covering each node of the MapReduce cluster. Therefore, it cannot only accelerate data processing on a single server but also on clusters.

5.3 In-Memory MapReduce

In-memory MapReduce borrows the basic idea of in-memory computing — data put in memory can be processed faster because memory is accessed much more quickly — and place job-related data in random access memory (RAM) to boost job execution. Typical systems include Spark [47], HaLoop [3], M3R [38], Twister [8], and Mammoth [37]. Spark, HaLoop, M3R, and Twister are specially designed for iterative computation and they reduce the IO cost (and thus boost computation) by placing in RAM the data to be processed multiple rounds. Such a way costs more because more memory is needed to hold the data and memory is more expensive than SSD. Mammoth is a comprehensive solution trying to solve inefficiencies in both memory usage and IO operations. To achieve the purpose, it devises various mechanisms to utilize memory more smartly, including rule-based prioritized memory allocation and revocation, global memory scheduling algorithm, memory-based shuffling, and so on. Mammoth can benefit from mpCache especially in a memory-constrained environment where only limited memory can be used for data caching. With mpCache introduced, more memory can be released to support computation and thus the task parallelism degree is improved, which means faster job execution.

5.4 IO Optimization via SSD-based Cache

With the emergence of NAND (Negative-AND) Flash memory, much research work has been reported that utilized SSD to improve storage performance. Yongseok et al. [30] proposed a way to balance cache size and update cost of flash memory so that better performance can be obtained in the HDD-SSD hybrid storage system. Hystor [4], Proximal IO [35], SieveStore [33], and HybridStore [19] also used SSD as a cache of hard disks as we do. But these methods only focus on a single node, with an aim to boost small files (typical size is below 200KB) manipulation by caching. mpCache can work across many nodes in a coordinated way. In addition, it devises a relatively complex and efficient cache replacement scheme to better support MapReduce applications.

5.5 MapReduce Optimization via In-Memory Cache

PACMan [2] cached input data in memory to reduce the high IO cost of hard disks so as to improve performance. Since the task parallelism degree of new generation of MapReduce (e.g., YARN) is more concerned with free memory. Caching data in memory, as shown in Section 4.3.1, would cut down the task parallelism and lead to low performance for some memory-intensive jobs (e.g., shuffle-heavy jobs in our benchmarks), for the memory left for normal task operations reduces. It is on account of only limited memory available and the large volume of Localized Data that PACMan only has Input Data cached. As a result, it just improves the Map phase. For those shuffle-heavy MapReduce jobs (e.g., *k-means* and *tera-sort*), they cannot benefit from in-memory caching in the Reduce phase. Unfortunately, the number of shuffle-heavy jobs is large in the real world. Our SSD-based caching solution can solve the problem and accelerate both phases.

6 Conclusion

In this paper we presented mpCache, a solution that utilizes SSD to cache MapReduce Input Data and Localized Data so that all the costly IO operations—**Read**, **Spill**, and **Merge**—are boosted and the whole job is accelerated as a result. Caching in such a way is cost-effective and can solve the performance degradation problem caused by in-memory caching as mentioned in Section 1. Given the fact that data will continue growing exponentially, this is especially important. We have implemented mpCache in Hadoop and evaluated it on a 7-node commodity cluster. The experimental results show that mpCache can get an average speedup of 2.09 times over Hadoop, and 1.79 times over PACMan, the latest work about MapReduce optimization by in-memory data caching.

Acknowledgment

This work is co-sponsored by Natural Science Foundation of China (61572280, 61433008, 61373145, U1435216), National High-Tech R&D (863) Program of China (2014AA01A302), National Key Research and Development Plan (2016YFB1000504), and National Basic Research (973) Program of China (2014CB340402).

References

[1] F. Ahmad, S. Lee, M. Thottethodi, and T. Vijaykumar, “Puma: Purdue mapreduce benchmarks suite,” 2012, http://web.ics.purdue.edu/˜fahmad/benchmarks.htm.

[2] G. Ananthanarayanan, A. Ghodsi, A. Wang, D. Borthakur, S. Kandula, S. Shenker, and I. Stoica, “Pacman: Coordinated memory caching for parallel jobs,” in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, NSDI'12*. USENIX, 2012, pp. 20–20.

[3] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst, “Haloop: Efficient iterative data processing on large clusters,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 285–296, 2010.

[4] F. Chen, D. A. Koufaty, and X. Zhang, “Hystor: making the best use of solid state drives in high performance storage systems,” in *Proceedings of the international conference on Supercomputing, ICS'11*. ACM, 2011, pp. 22–32.

[5] P. Costa, A. Donnelly, A. Rowstron, and G. O???Shea, “Camdoop: Exploiting in-network aggregation for big data applications,” in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, NSDI'12*. USENIX, 2012, pp. 3—3.

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/TBDATA.2016.2599933, IEEE Transactions on Big Data

[6] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[7] A. Dou, V. Kalogeraki, D. Gunopulos, T. Mielikainen, and V. H. Tuulos, "Misco: a mapreduce framework for mobile systems," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments, PETRA'10*. ACM, 2010, p. 32.

[8] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox, "Twister: a runtime for iterative mapreduce," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. ACM, 2010, pp. 810–818.

[9] W. Fang, B. He, Q. Luo, and N. K. Govindaraju, "Mars: Accelerating mapreduce with graphics processors," *IEEE Transactions on Parallel and Distributed Systems, TPDS'11*, vol. 22, pp. 608–620, 2011.

[10] M. J. Feeley, W. E. Morgan, E. Pighin, A. R. Karlin, H. M. Levy, and C. A. Thekkath, *Implementing global memory management in a workstation cluster*. ACM, 1995.

[11] M. J. Franklin, M. J. Carey, and M. Livny, *Global memory management in client-server DBMS architectures*. University of Wisconsin-Madison, Computer Sciences Department, 1992.

[12] H. Garcia-Molina and K. Salem, "Main memory database systems: An overview," *IEEE Transactions on Knowledge and Data Engineering, TKDE'92*, vol. 4, no. 6, pp. 509–516, 1992.

[13] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, "Building a high-level dataflow system on top of map-reduce: the pig experience," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1414–1425, 2009.

[14] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *ACM SIGOPS Operating Systems Review*, vol. 37. ACM, 2003, pp. 29–43.

[15] A. Hadoop, "Hadoop," 2014, http://hadoop.apache.org.

[16] J. Handy, "Flash memory vs. hard disk drives - which will win?" 2014, http://http://www.storagesearch.com/semico-art1.html.

[17] B. He, W. Fang, Q. Luo, N. K. Govindaraju, and T. Wang, "Mars: a mapreduce framework on graphics processors," in *Proceedings of the 17th international conference on Parallel architectures and compilation techniques, PACT'08*. ACM, 2008, pp. 260–269.

[18] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: distributed data-parallel programs from sequential building blocks," *ACM SIGOPS Operating Systems Review*, vol. 41, no. 3, pp. 59–72, 2007.

[19] Y. Kim, A. Gupta, B. Urgaonkar, P. Berman, and A. Sivasubramaniam, "Hybridstore: A cost-efficient, high-performance storage system combining ssds and hdds," in *2011 IEEE 19th International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems, MASCOTS'11*. IEEE, 2011, pp. 227–236.

[20] D. E. Knuth, *The art of computer programming, vol. 3, Addison-Wesley, Reading Mass.* Pearson Education, 2005.

[21] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable dram alternative," *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3, pp. 2–13, 2009.

[22] S.-W. Lee, B. Moon, C. Park, J.-M. Kim, and S.-W. Kim, "A case for flash memory ssd in enterprise database applications," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD'08*. ACM, 2008, pp. 1075–1086.

[23] H. Lin, X. Ma, J. Archuleta, W.-c. Feng, M. Gardner, and Z. Zhang, "Moon: Mapreduce on opportunistic environments," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, HPDC'10*. ACM, 2010, pp. 95–106.

[24] Y. Lu, J. Shu, and W. Wang, "Reconfs: a reconstructable file system on flash storage," in *Proceedings of the 12th USENIX Conference on File and Storage Technologies, FAST'14*. USENIX, 2014, pp. 75–88.

[25] F. Marozzo, D. Talia, and P. Trunfio, "P2p-mapreduce: Parallel data processing in dynamic cloud environments," *Journal of Computer and System Sciences, JCSS'12*, vol. 78, no. 5, pp. 1382–1402, 2012.

[26] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards characterizing cloud backend workloads: insights from google compute clusters," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 34–41, 2010.

[27] A. C. Murthy, C. Douglas, M. Konar, O. Malley, S. Radia, S. Agarwal, and V. KV, "Architecture of next generation apache hadoop mapreduce framework," Tech. rep., Apache Hadoop, Tech. Rep., 2011.

[28] nathanmarz, "Storm," 2014, https://github.com/nathanmarz/storm.

[29] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, "S4: Distributed stream computing platform," in *2010 IEEE International Conference on Data Mining Workshops, ICDMW'10*. IEEE, 2010, pp. 170–177.

[30] Y. Oh, J. Choi, D. Lee, and S. H. Noh, "Caching less for better performance: Balancing cache size and update cost of flash memory cache in hybrid storage systems," in *Proceedings of the 10th USENIX Conference on File and Storage Technologies, FAST'12*. USENIX, 2012, pp. 25–25.

[31] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig latin: a not-so-foreign language for data processing," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1099–1110.

[32] J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazières, S. Mitra, A. Narayanan, G. Parulkar, M. Rosenblum *et al.*, "The case for ramclouds: scalable high-performance storage entirely in dram," *ACM SIGOPS Operating Systems Review*, vol. 43, no. 4, pp. 92–105, 2010.

[33] T. Pritchett and M. Thottethodi, "Sievestore: a highly-selective, ensemble-level disk cache for cost-performance," in *Proceedings of the 37th Annual International Symposium on Computer Architecture, ISCA'10*. ACM, 2010, pp. 163–174.

[34] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating mapreduce for multi-core and multiprocessor systems," in *IEEE 13th International Symposium on High Performance Computer Architecture, HPCA'07*. IEEE, 2007, pp. 13–24.

[35] J. Schindler, S. Shete, and K. A. Smith, "Improving throughput for small disk requests with proximal i/o." in *Proceedings of the 9th USENIX Conference on File and Storage Technologies, FAST'11*. USENIX, 2011, pp. 133–147.

[36] L. Seiler, D. Carmean, E. Sprangle, T. Forsyth, M. Abrash, P. Dubey, S. Junkins, A. Lake, J. Sugerman, R. Cavin *et al.*, "Larrabee: a many-core x86 architecture for visual computing," in *ACM Transactions on Graphics, TOG'08*, vol. 27. ACM, 2008, p. 18.

[37] X. Shi, M. Chen, L. He, X. Xie, L. Lu, H. Jin, Y. Chen, and S. Wu, "Mammoth: Gearing hadoop towards memory-intensive mapreduce applications," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 26, no. 8, pp. 2300–2315, 2015.

[38] A. Shinnar, D. Cunningham, V. Saraswat, and B. Herta, "M3r: increased performance for in-memory hadoop jobs," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1736–1747, 2012.

[39] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST'10*. IEEE, 2010, pp. 1–10.

[40] J. A. Stuart and J. D. Owens, "Multi-gpu mapreduce on gpu clusters," in *2011 IEEE International Parallel & Distributed Processing Symposium, IPDPS'11*. IEEE, 2011, pp. 1068–1079.

[41] J. Talbot, R. M. Yoo, and C. Kozyrakis, "Phoenix++: modular mapreduce for shared-memory systems," in *Proceedings of the second international workshop on MapReduce and its applications*. ACM, 2011, pp. 9–16.

[42] B. Tang, M. Moca, S. Chevalier, H. He, and G. Fedak, "Towards mapreduce for desktop grid computing," in *2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC'10*. IEEE, 2010, pp. 193–200.

[43] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1626–1629, 2009.

[44] B. Wang, J. Jiang, and G. Yang, "mpcache: Accelerating mapreduce with hybrid storage system on many-core clusters," in *Network and Parallel Computing, NPC'14*. Springer, 2014, pp. 220–233.

[45] H.-c. Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker, "Map-reduce-merge: simplified relational data processing on large clusters," in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD'07*. ACM, 2007, pp. 1029–1040.

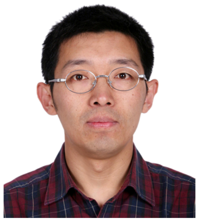
[46] Y. Yu, M. Isard, D. Fetterly, M. Budiu, Ú. Erlingsson, P. K. Gunda, and J. Currey, "Dryadlinq: A system for general-purpose distributed data-parallel computing using a high-level language," in *Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation, OSDI'08*. USENIX, 2008, pp. 1–14.

[47] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, HotCloud'10*, 2010, pp. 10–10.

[48] M. Zheng, J. Tucek, F. Qin, and M. Lillibridge, "Understanding the robustness of ssds under power fault," in *Proceedings of the*

11th USENIX Conference on File and Storage Technologies, FAST'13. USENIX, 2013, pp. 271–284.

Bo Wang received a BS degree in computer science and technology from Tsinghua University, China in 2008 and a MS degree in computer applications from North China Institute of Computing Technology in 2011. He is currently a PhD candidate in the Department of Computer Science and Technology at Tsinghua University, China, working on Hadoop optimization. His research interests include distributed systems, big data computing, storage and file systems, and virtualization. He is a student member of IEEE.



Jinlei Jiang received a PhD degree in computer science and technology from Tsinghua University, China in 2004 with an honor of excellent dissertation. He is currently an associate professor in the Department of Computer Science and Technology at Tsinghua University, China. His research interests include distributed computing and systems, cloud computing, big data, and virtualization. He is currently on the editorial boards of *KSII Transactions on Internet and Information Systems*, *International Journal on Advances in Intelligent Systems*, and *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*. He is a winner of Humboldt Research Fellowship and an IEEE member.

Yongwei Wu received the PhD degree in applied mathematics from the Chinese Academy of Sciences in 2002. He is currently a professor in computer science and technology at Tsinghua University, China. His research interests include parallel and distributed processing, mobile and distributed systems, cloud computing, and storage. He has published over 80 research publications and has received two Best Paper Awards. He is currently on the editorial boards of *IEEE Transactions on Cloud Computing*, *Journal of Grid Computing*, *IEEE Cloud Computing*, and *International Journal of Networked and Distributed Computing*. He is an IEEE member.

Guangwen Yang is a professor in the Department of Computer Science and Technology and the director of the Institute of High Performance Computing, Ministry of Education Key Laboratory for Earth System Modeling at Tsinghua University, China. His research interests include parallel and distributed algorithms, cloud computing, and the earth system model. He received a PhD degree in computer architecture from Harbin Institute of Technology, China in 1996 and a MS degree in applied mathematics from the same university in 1987. He is an IEEE member.

Keqin Li is a SUNY Distinguished Professor of computer science. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, CPU-GPU hybrid and cooperative computing, multicore computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file sharing systems, mobile computing, service computing, Internet of things and cyber-physical systems. He has published over 390 journal articles, book chapters, and refereed conference papers, and has received several best paper awards. He is currently or has served on the editorial boards of *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Computers*, *IEEE Transactions on Cloud Computing*, *Journal of Parallel and Distributed Computing*. He is an IEEE Fellow.