Accumulative Computation on MapReduce

Yu Liu\textsuperscript{1,4,a)} \quad Kento Emoto\textsuperscript{2,b)} \quad Kiminori Matsuzaki\textsuperscript{3,c)} \quad Zhenjiang Hu\textsuperscript{4,d)}

Received: July 4, 2013, Accepted: October 9, 2013

Abstract: MapReduce programming model attracts a lot of enthusiasm among both industry and academia, largely because it simplifies the implementations of many data parallel applications. In spite of the simplicity of the programming model, there are many applications that are hard to be implemented by MapReduce, due to their innate characters of computational dependency. In this paper we propose a new approach of using the programming pattern \textit{accumulate} over MapReduce, to handle a large class of problems that cannot be simply divided into independent sub-computations. Using this \textit{accumulate} pattern, many problems that have computational dependency can be easily expressed, and then the programs will be transformed to MapReduce programs executed on large clusters. Users without much knowledge of MapReduce can also easily write programs in a sequential manner but finally obtain efficient and scalable MapReduce programs. We describe the programming interface of our \textit{accumulate} framework and explain how to transform a user-specified \textit{accumulate} computation to an efficient MapReduce program. Our experiments and evaluations illustrate the usefulness and efficiency of the framework.

Keywords: parallel programming, MapReduce, accumulative computation, automatic parallelization, algorithmic skeleton

1. Introduction

MapReduce \cite{10} is a popular parallel programming framework, which is designed for parallel data processing such as clustering, mining or statistical analysis on large-scale data. The programming model of MapReduce is inspired by the functional programming languages and algorithms in MapReduce model are mainly restricted to using \textit{map} and \textit{reduce} functions\textsuperscript{1}. The MapReduce framework executes such functions in a massively parallel manner while dealing with failures automatically.

In spite of the simplicity, many problems are still difficult to be expressed in MapReduce model. As an example, consider the \textit{elimSmallers} problem of eliminating all the smaller elements of a list to produce an ascending list (if an element is less than someone in the previous, it is \textit{smaller}). For instance, given a list \([11, 15, 8, 9, 20, 25, 12, 23]\), then \(8, 9, 12\) and \(23\) are smaller ones, and thus the result is \([11, 15, 20, 25]\). A recursive function that solves this problem can be defined as follows, in Haskell \cite{5}:

\begin{align*}
\text{elimSmallers} & \colon \mathbb{Z} \\
\text{elimSmallers} & \colon (\mathbb{Z} \times \mathbb{Z}) \to \mathbb{Z}
\end{align*}

\begin{align*}
\text{elimSmallers} \ [\ c \ ] \ & = \ [\ ] \\
\text{elimSmallers} \ (x : xs) \ & = \ (\text{if } x < c \ \text{then } [\ ] \ \text{else } [x]) \ + \ \\
& \quad \text{elimSmallers} \ xs \ (\text{if } x < c \ \text{then } c \ \text{else } x).
\end{align*}

In this function, recursively, numbers of input are compared with an accumulative parameter \(c\) (with initial value \(-\infty\)). The accumulative parameter \(c\) always holds the current maximum value and is used in the next recursion. If a number is no less than \(c\) then it is appended to the tail of the result, and otherwise dropped. In functional programming, such kind of computation pattern with accumulative parameters is called accumulative computation \cite{15, 16, 18}.

The recursive function \textit{elimSmallers} clearly describes the computation (in \(O(n)\) work, \(n\) is the length of input), but it cannot be easily mapped to MapReduce, because in the recursive function \textit{elimSmallers}, every inner step of the recursion relies on the current maximum value, which is computed at the outer step. Such kind of recursive functions do not correspond to a simple divide-and-conquer algorithm. Developing an \(O(n)\) work MapReduce algorithm for \textit{elimSmallers} needs to resolve such computational dependency and avoid unnecessary and expensive I/O, which is not easy for many programmers. There are many applications (e.g., the \textit{prefix-sum/scan} related problems \cite{6}) having similar characters with \textit{elimSmallers} and thus are also difficult to be resolved in MapReduce model.

In this paper, we propose a new programming framework for simplifying MapReduce programming on accumulative computations that cannot be expressed by using \textit{map} and \textit{reduce} functions in only one iteration of MapReduce processing. The programming interface\textsuperscript{2} is designed to help users defining recursive functions in the accumulative form, and then efficient and scalable MapReduce solutions can be automatically gained. Our main technical contributions can be summarized as follows.

\textsuperscript{1} The \textit{map} and \textit{reduce} functions in MapReduce are inspired by but not equivalent to those in Lisp or Haskell.

\textsuperscript{2} The \textit{accumulate} skeleton has been implemented using MPI that is more flexible in programming model and does not have same constants as MapReduce has.
• We implemented a parallel programming framework modeled by the accumulate computation pattern [15], [16], [18] for accumulative computations. The framework can produce efficient and scalable MapReduce jobs for dealing with large data. Two important technical points in the implementations are: (1) dealing with accumulative computations on ordered lists, in spite of the massively parallel execution manner of MapReduce and (2) providing generic and high-level programming interfaces for wrapping the low-level MapReduce APIs.

• We evaluated the framework with many interesting examples, e.g., tag match, elimSmaller, maximum prefix sum, and line-of-sight, on MapReduce clusters dealing with big input data. The experimental results show good efficiency and scalability of our accumulation framework.

The organization of this paper is as follows. In Section 2, we briefly explain the MapReduce model and accumulative computation patterns. In Section 3, we introduce our MapReduce algorithms for accumulative computations and describe the library we developed on Hadoop. We present the experimental results in Section 4, introduce the related work in Section 5, and conclude the paper in Section 6.

2. Background

2.1 Notations

To make our descriptions precise and clear, notations in this paper are mainly based on the functional language Haskell [5]. Function application is denoted with a space with its argument without parentheses, i.e., \( f \ a \) equals to \( f(a) \). Functions are carried and bound to the left, and thus \( f \ a \ b \) equals to \( (f \ a) \ b \). Function application has higher precedence than using operators, so \( f \ a \ @ \ b = (f \ a) \ @ \ b \). We use the operator \( \circ \) over functions and \( (f \circ g) \ x = f \ (g \ x) \). Function \( \text{id} \) is the identity function. Tuples are written like \((a, b)\) or \((a, b, c)\). Function \( \text{fst} \) (\( \text{snd} \)) extracts the first (the second) element of the input tuple, and \( \text{top} \) \( x \) returns the first element of a stack \( x \). Function \( \text{drop} \ m \ x \) drops the first \( m \) elements from a list \( x \). The binary operator \( \top \) applies on two numbers and returns the larger one. We denote lists with square brackets, and use \( [ ] \) to denote an empty list, and \( + \) to denote the list concatenation: \([3, 1, 4] + [1, 5] = [3, 1, 4, 1, 5]\). Function \( [] \) takes a value and returns a singleton list with the value. The scan, map, reduce, zip are standard skeletons in the Bird-Meertens formalism [4], [26]. To distinguish the map / reduce functions in MapReduce form above skeletons, we use \( f_{\text{MAP}} \) and \( f_{\text{REDUCE}} \) for the parameter functions used in the MapReduce.

2.2 MapReduce

Google’s MapReduce [10] is a popular programming model for processing large data sets in a massively parallel manner. In the MapReduce programming model, parallel computations are represented in the paradigm of a parallel Map processing followed by a Reduce processing\(^{33}\). Between the Map and Reduce phases, there is a Shuffle/Sort phase. Figure 1 shows the typical data-processing flow of MapReduce. Note that Map tasks are executed independently (no direct way for one Map task to communicate/synchronize with another one, and so do the Reduce tasks. The only global synchronization in MapReduce is the barrier between Map and Reduce. Usually an instance of Map (Reduce) task is also called a mapper (reducer). The types of the two basic functions of MapReduce are defined as follows.

• Function \( f_{\text{MAP}} \):

\[
f_{\text{MAP}} : (k_1, v_1) \rightarrow (k_2, v_2)
\]

This function is invoked during the Map phase, and it is applied on each key-value pair of input and returns an intermediate key-value pair.

• Function \( f_{\text{REDUCE}} \):

\[
f_{\text{REDUCE}} : ([k_2, [v_2]],) \rightarrow (k_3, v_3)
\]

This function is invoked during the Reduce phase, and it takes a key and a list of values associated to the key and merges the values.

Nowadays, several free, realistic implementations of MapReduce are available. In particular, Hadoop [2] is a famous open-source implementation using Java as its primitive language. Our implementation is based on Hadoop.

2.3 Accumulative Computations

Accumulative computation [15] plays an important role in describing a computation on an ordered list from left or right, when a later computation depends more or less on this computation. The innate character of data dependency can be captured by using an accumulative parameter that holds and delivers some information through the whole computation.

2.4 General Accumulative Computation Pattern

The accumulate skeleton abstracts a typical pattern of recursive functions with an accumulative parameter, which can be defined as a function \( h \) in the following form.

\[
h \ [] \ c = g \ c
\]

\[
h \ (x : xs) \ c = p \ (x, c) \oplus h \ xs \ (c \oplus q \ x).
\]

This definition provides a natural way to describe computations with data dependencies and can be understood as follows.

• If the input list is empty, the result is computed by applying

\(^{33}\) http://code.google.com/p/diffusion-mapreduce/

\(^{34}\) The Reduce processing can be done in parallel or sequentially.
some function \( g \) to accumulative parameter \( c \).

- If the input list is not empty and its head and tail parts are \( x \) and \( xs \) respectively, then the result is generated by combining the following two values using some binary operator \( \circ \): the result of applying \( p \) to \( x \) (head value) and \( c \) (the accumulative parameter), and the recursive call of \( h \) to \( xs \) (the rest part of the input list) with its accumulative parameter updated to \( c \circ q \). x.

Because \( h \) is uniquely defined by \( g \), \( p \), \( \circ \), \( q \), and \( \@ \), so we write \( h \) with special parentheses \([ \[] \) \() \] as:

\[
\text{elimSmaller} \, \text{xs} \, c = \langle g, \langle p, \circ \rangle, \langle q, \@ \rangle \rangle \, \text{xs} \, c.
\]

Note that \( \langle p, \circ \rangle \) and \( \langle q, \@ \rangle \) correspond to two basic recursive forms \( \text{foldr} \) and \( \text{foldl} \) \cite{14} respectively. The \( \text{elimSmaller} \) discussed in the introduction can be also written as follows.

\[
\text{elimSmaller} \, \text{xs} \, c = \langle g, \langle p, \circ \rangle, \langle q, \@ \rangle \rangle \, \text{xs} \, c
\]

where \( g \) \( c \) \( = \) \( [ \[] \)

\[
p \, (x, c) = \begin{cases} x < c \ & \text{then} \ [ \[ \text{x} \] \] \ , \ \circ = \oplus , \\ \text{id} \ , \ \circ = \uparrow . \end{cases}
\]

Since the function \( h \) in the above form represents the most natural recursive definition on lists with a single accumulative parameter, it is general enough to capture many algorithms \cite{15} as seen below.

Scan

Given a list \([ x_1, x_2, x_3, x_4] \) and an associative binary operator \( \circ \) with an identity element \( i_0 \), a function \( \text{scan} \) computes all its prefix sums yielding

\[
[i_0, x_1, x_1 \circ x_2, x_1 \circ x_2 \circ x_3, x_1 \circ x_2 \circ x_3 \circ x_4].
\]

As mentioned in the introduction, \( \text{scan} \) can be defined in terms of \( \text{accumulate} \) (by giving an initial value \( i_0 \) to the accumulative parameter \( c \)).

\[
\text{scan} \, [ \[] \, c = [ \[] \, c
\]

\[
\text{scan} \, (x : xs) \, c = ([ ] \circ \text{snd})(x, c) \oplus \text{scan} \, xs \, (c \circ (\text{id} \, x)).
\]

The function \( \text{scan} \) is very useful in algorithm design and is also a primitive operator in lots of parallel computations \cite{6}. For example, lexical analysis, quick sort, and regular-expression matching can be implemented by using \( \text{scan} \).

Line-of-Sight Problem

The well known line-of-sight problem \cite{6} is that given a terrain map in the form of a grid of altitudes and an observation point, find which points are visible along a ray originating at the observation point. For instance, we use a pair \((d,a)\) to represent a point, where \( a \) is the altitude of the point and \( d \) is its distance from the observation point. The function \( i \, (d,a) = a/d \) computes the tangent of an angle. If the list is \([([1,1],[2,2]),(3,2),(4,10])\), then the point \((3,2)\) is invisible. The function \( \text{los} \) \cite{18} solves a simplified line-of-sight problem which counts the number of visible points.

\[
\text{los} \, \text{xs} \, c = \langle g, \langle p, \oplus \rangle, \langle q, \uparrow \rangle \rangle \, \text{xs} \, c
\]

where \( g \) \( c \) \( = \) \( 0 \)

\[
p \, (x, c) = \begin{cases} c \leq \uparrow \, x \ & \text{then} \ 1 \ \text{else} \ 0 \\ q \, x = \uparrow \, x . \end{cases}
\]

Maximum Prefix Sum Problem

Intuitively, the maximum prefix sum problem is to compute the maximum sum of all the prefixes of a list. Given a list \([3,-4,9,2,-6]\) the maximum of the prefix sums is \(10\), to which the underlined prefix corresponds. We can define a function \( \text{mps} \) that solves this problem, in terms of \( \text{accumulate} \).

\[
\text{mps} \, \text{xs} \, c = \langle \text{id}, (\text{snd}, \uparrow) \rangle, (\text{id}, \uparrow) \rangle \, \text{xs} \, c
\]

Tag Matching Problem

The tag matching problem is to check whether the tags are well matched or not in a document, e.g., an XML file. There is an accumulative function \( \text{tagmatch} \) introduced by Ref. \cite{18} for the tag matching problem.

\[
\text{tagmatch} \, \text{xs} \, cs = \langle \text{isEmpty}, (\text{id}, \land), (\text{q}, \circledast) \rangle \, \text{xs} \, cs
\]

where

\[
p \, (x, cs) = \begin{cases} \text{if} \ \text{isOpen} \ x \ & \text{then} \ True \\ \text{else if} \ \text{isClose} \ x \ & \text{then} \ \text{notEmpty} \ cs \land \ \text{match} \ x \ (\text{top cs}) \\ \text{else} \ True \\ q \, x = \begin{cases} \text{if} \ \text{isOpen} \ x \ & \text{then} \ ([x], 1, 0) \\ \text{else if} \ \text{isClose} \ x \ & \text{then} \ ([], 0, 1) \\ \text{else} \ ([], 0, 0) \\ (s_1, n_1, m_1) \circledast (s_2, n_2, m_2) = \begin{cases} \text{if} \ n_1 \leq m_2 \ & \text{then} \ (s_2, n_2, m_1 + n_2 - n_1) \\ \text{else} \ (s_2 \ast \text{drop} \ m_2 \ s_1, \ n_1 + n_2 - m_2, m_1). \end{cases} \end{cases}
\]

3. Parallel Accumulation on MapReduce

Due to the different infrastructures of MapReduce and MPI, doing parallel accumulative computation on MapReduce is quite different and challenging. In typical MapReduce programming environments like Hadoop, or some MapReduce-like ones such as Dryad \cite{17} and Spark \cite{28}, there is no option for users to use peer-to-peer communication like MPISend or MPIRecv, and the synchronization of parallel processes can be only implemented by making use of the barrier between Map phase and Reduce phase.

There are two strategies for implementation of the accumulative skeleton. One is extending the existing MapReduce framework, by adding new peer-to-peer communication functions and barrier functions, so that we can do implementation in a similar way as Ref. \cite{18} did. The other way is just making use of existing high-level API of a MapReduce framework such as map and reduce functions (and the necessary API of the distributed file system).

Actually we have implemented and evaluated both, based on the state-of-the-art open-source MapReduce framework Hadoop. The prior one has advantages in the performance (several times faster) but significantly affected the fault tolerance mechanisms of MapReduce and it is not compatible with vanilla MapReduce frameworks [10]. On the contrary, the latter way, although it is not as fast as the prior one, enjoys all features of MapRede’s system design and has good portability (between different instances of \( \text{mps} \).
MapReduce frameworks). In this paper we select the latter way as our main solution and introduce the algorithm and implementation of it. The readers who are interested in the solution of prior way can refer the source code in the package of our framework.

Users of our framework can directly use the accumulate as a building block to solve their computations. Our implementation is based on Hadoop but could be easily ported to other MapReduce engines such as Spark[28].

3.1 Input Data Model

The accumulative computations take lists as input and obey the order of elements in the input list, while the input data of MapReduce are represented as a set of records (key-value pairs) stored in the distributed file system. This means we need a file-based list data structure for computations of accumulate. One issue for processing a file-based list is that the massively parallel processing manner of MapReduce could cause the accumulative computations being out-of-order.

For simplicity of discussion, we suppose the input data to an accumulative computation are stored as a list of records in one binary file (could be very huge) in the distributed file system (DFS). Each record of the file is a serialized Java object\(^*5\) that represents an element of input list, and when detrialized to memory, each record will be transformed to a key-value pair. The key part of the pair is the corresponding Java object and the value part is a null object.

If big enough, the input file will be split to several splits by DFS (each split is called a chunk in the DFS) [13] and distributed to several DataNodes, and the DFS knows the offsets of each splits [10], [13], which can be seen as the indices of the splits. When data are loaded to multiple mappers, users of MapReduce cannot control which mapper to load which splits. In order to keep the total order of computation, the output of each mapper must be associated with the offset of its input, so that when merging the results from mappers, the order can be carefully manipulated by making use of such offsets and the sorting function.

3.2 A 2-pass MapReduce Algorithm for Accumulation

From the diffusion theorem [15], [18], an accumulative function \(h = \ll g, (p, \oplus), (q, \otimes) \rr\) can be transformed into the following compositional form using the parallel skeletons scan, map, reduce and zip.

\[
h \left( xx \right) c = \text{reduce} \left( \otimes \right) \left( \text{map} \left( p, \otimes \right) q \left( xx \right) \right) \oplus \text{zip} \left( xx, bs \right)
\]

where \( bs \oplus [b] = \text{map} \left( c \otimes \left( \text{scan} \left( \otimes \right) \left( \text{map} \left( q, xx \right) \right) \right) \right) \text{as} = \text{zip} \left( xx, bs \right) \).

In this form, \( \text{map} \left( c \otimes \left( \text{scan} \left( \otimes \right) \left( \text{map} \left( q, xx \right) \right) \right) \right) \) can be firstly computed to get \( bs \oplus [b] \), then \( xx \) and \( bs \) to obtain \( ax \), and finally \( \text{reduce} \left( \otimes \right) \left( \text{map} \left( p, \otimes \right) q \left( xx \right) \right) \oplus \) to get the result. However, directly doing in this way will generate a lot of intermediate data such as \( bs \oplus [b], \) as (these are much bigger than the input), so that it is uncomputable in the MapReduce-like environments where input data are usually in terabytes. The previous MPI implementa-

\[^*5\] In practice, serialization systems like Avro[1] can be used to improve performance.

3.2.1 The MapReduce Implementation for General Accumulation

Our approach is to divide the computation into two MapReduce phases and restrain the data transportation between the two. Suppose input list \( xx \) is split to \( p \) sublists, i.e., \( xx = chk_1 \oplus chk_2 \oplus \ldots \oplus chk_p \). The \( k \)-th split \( chk_k \) has \( m \) elements \([x_1,x_2, \ldots, x_m]\) and its offset is \( seg_k \). Our two-pass MapReduce algorithm (shown in Fig. 2) actually avoids generating large intermediate data and thus it is efficient. We introduce the details in the following paragraphs.

The first MapReduce job

There are \( p \) Map tasks spawned for each split, in the first MapReduce job. In general, for each sublist \( chk_k (k \in [1,p]) \), the first MapReduce computes:

\[
\text{mapRed}_k = \text{reduce} \left( \otimes \right) \left( \text{map} \left( q, chk_k \right) \right).
\]

We do the above computation during Map phase and just use one reducer to collect the result. In detail, each Map task iterates over the elements of its input and applies the following \( f_{\text{MAP}} \) function on each input record \( (x_i^k, \_ \_ ) (i \in [1,m]) \).

\[
f_{\text{MAP}} \left( x_i^k, \_ \_ \right) = \left( \left( 0, \text{seg}_k \right), q(x_i^k) \right).
\]

\[
\text{HDFS} \quad \text{seg}^1 \quad x_1^1, x_2^1, x_3^1, x_4^1, x_5^1 \ldots \quad \ldots \quad \text{seg}^p \quad x_1^p, x_2^p, x_3^p, x_4^p, x_5^p \ldots x_m^p
\]

\[
\text{Mapper1} \\
\text{JOB-1}
\]

\[
\text{Mapper2p}
\]

\[
\text{Reducre1} \quad \text{shuffle & sort}
\]

\[
\text{DistributedCache} \quad \text{v1, v2, \ldots, v_p}
\]

\[
\text{Mapper1} \quad \text{JOB-2}
\]

\[
\text{Reducre1} \quad \text{shuffle & sort}
\]

\[
\text{output}
\]

\[
\text{Fig. 2 The 2-pass MapReduce Accumulation.}
\]

\[\text{2014 Information Processing Society of Japan} \quad 21\]
Different with general MapReduce applications, here once the \( f_{\text{MAP}} \) function was applied on an input pair \((x_{k}, \cdot, \cdot)\), the output did not be emitted immediately, but aggregated to a value \( v_{k} \):

\[
v_{k} = t_{0} \odot q(x_{1}) \odot q(x_{2}) \odot \cdots \odot q(x_{m-1}).
\]

After the iterations, each Map task emits only one key-value pair: \((0, \text{seg}k) , v_{k})\). Here the key itself is a pair consisting of a constant value 0 and the offset \(\text{seg}k\). The outputs of Map tasks are grouped (by the constant value) and sorted by the offset.

In the reduce phase we only spawn one reducer. We use a special group function that groups records by first element of keys, so that the reducer collects all \(( (0, \text{seg}k) , v_{k})\) \((k \in [1, p])\) and sort them by the offsets \(\text{seg}k\). Then the reducer just emits a key-value pair \((v_{0}, v_{1}, \ldots, v_{p})\) \((\cdot)\) (the key is a list and value part is useless) to the distributed file system. The \( f_{\text{REDUCE}} \) function is defined as follows.

\[
f_{\text{REDUCE}} (0, [v_{0}, v_{1}, \ldots, v_{p}]) = ([v_{1}, v_{2}, \ldots, v_{p}]) \cdot (0, \cdot, \cdot).
\]

Then the final result of the first MapReduce is a list that contains \(p\) elements, say \(v_{s} = [v_{1}, v_{2}, \ldots, v_{p}]\), and \(v_{s}\) is guaranteed being in the correct order.

The second MapReduce job

After the first MapReduce, we initialize the second MapReduce: each second-Map task reads \(v_{s} = [v_{1}, v_{2}, \ldots, v_{p}]\) (the result list of the first reducer\(^{6}\)) from HDFS, in addition to the same input data as the first Map task. After initialization, in general, for each sublist \(\text{ch}k_{k}\) each Map task in the second MapReduce computes:

\[
\text{mapRed}_{\text{map}} \text{ch}k_{k} = \text{map} p (\text{zip} \text{ch}k_{k} ws)
\]

where

\[
\text{ws} = \text{map} \left( \text{tok} \odot (\text{scan} \odot (\text{map} q \text{ch}k_{k})) \right)
\]

\[
\text{vk} = \text{reduce} (\odot) \left( c, v_{1}, v_{2}, \ldots, v_{k-1} \right).
\]

The only one Reduce task in the second MapReduce computes:

\[
\text{mapRed}_{\text{red}} ss = \left( \text{reduce} (\odot) (ss) \right) \odot g(vp)
\]

where

\[
ss = [s_{1}, s_{2}, \ldots, s_{p}]
\]

\[
vp = \text{reduce} (\odot) \left( [v_{1}, v_{2}, \ldots, v_{p}] \right).
\]

In detail, each Map task computes in a loop \(s_{k} = p(s_{1}, w_{k}) \odot p(s_{2}, w_{k}) \odot q(x_{1}) \odot \cdots \odot p(s_{m-1}, w_{k}) \odot q(x_{m-1})\), where \(w_{k} = c \odot v_{1} \odot \cdots \odot v_{k-1}\).

The output of a Map task is a nested key-value pair whose key is the same \((0, \text{seg}k)\), and the value is \((v_{1}, s_{k})\). The \( f_{\text{MAP}} \) function is defined as follows.

\[
f_{\text{MAP}} \left( (x_{k}, \cdot, \cdot) \right) = (0, \text{seg}k) \cdot p(x_{1}, w_{k}) \odot q(x_{k-1})
\]

Similar to the first pass MapReduce, the outputs of all Map tasks are grouped/sorted, and we spawn a single reducer in the second MapReduce. The final result is \(s_{1} \odot s_{2} \odot \cdots \odot s_{p} \odot g(c \odot v_{1} \odot v_{2} \odot \cdots \odot v_{p})\). The \( f_{\text{REDUCE}} \) function is defined as follows.

\[
f_{\text{REDUCE}} (0, ss) = \left( \text{reduce} (\odot) (ss) \odot g(vp) \right) \cdot (0, \cdot, \cdot)
\]

Here \(ss = [s_{1}, s_{2}, \ldots, s_{p}]\), and \(vp = \text{reduce} (\odot) \left( [v_{1}, v_{2}, \ldots, v_{p}] \right)\).

\(^{6}\) We use the \(\text{DistributedCache}\) function of Hadoop to implement such initialization.

An example

As a concrete example, let us demonstrate the above algorithm to compute the \emph{elimSmallers} problem on a two-nodes cluster. An input list is given as \([11, 15, 8, 9, 20, 25, 12, 23]\), the initial value of parameter \(c = -\infty\), and the list is split to two (with the offset 0 and 10, respectively). The processing is represented in the following tabular form, step by step.

<table>
<thead>
<tr>
<th></th>
<th>node(^{1})</th>
<th>node(^{2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>input</td>
<td>0 : [11, 15, 8, 9]</td>
<td>10 : [20, 25, 12, 23]</td>
</tr>
<tr>
<td>1(^{\text{st}}) Map</td>
<td>(\infty \downarrow 11 \uparrow 15 \uparrow 8 \uparrow 9 = 15)</td>
<td>(\infty \downarrow 20 \uparrow 25 \uparrow 12 \uparrow 23 = 25)</td>
</tr>
<tr>
<td>output</td>
<td>((0,0,15))</td>
<td>((0,10,25))</td>
</tr>
<tr>
<td>1(^{\text{st}}) Reduce</td>
<td>emit directly</td>
<td></td>
</tr>
<tr>
<td>output</td>
<td>((15,25))</td>
<td>N/A</td>
</tr>
<tr>
<td>2(^{\text{nd}}) Map</td>
<td>(p(11 \uparrow -\infty \downarrow 15 \uparrow 11 + p(15 \uparrow 11 + 15 \uparrow 15 \uparrow 0 = 15)</td>
<td>(p(20 \uparrow -\infty \downarrow 20 \uparrow 12 \uparrow 25 \uparrow 20 \uparrow 12 \uparrow 25)</td>
</tr>
<tr>
<td>output</td>
<td>((0,0,15,11,15))</td>
<td>((0,10,25,20,20))</td>
</tr>
<tr>
<td>2(^{\text{nd}}) Reduce</td>
<td>(11,15 \uparrow 20,25) (\uparrow -\infty \downarrow 15 \uparrow 25)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Discussions on efficiency

Our two-pass MapReduce algorithm for accumulate \([g, (p, \odot), (q, \odot)]\) has two parallel Map phases and two sequential Reduce phases, and it only generates \(p\) intermediate data \((v_{1}, v_{2}, \ldots, v_{p})\) and duplicates \(p\) times through networks (copied to \(p\) Map tasks using parallel-copy). Consider that \(p\), i.e., the number of input splits, is not a very huge value (for \(1\) TB data, if the chunk size of HDFS is \(128\) MB then the \(p\) is \(7813\)), so that all \(v_{i}\) and \(x_{i}\) are in small constant size, then the two Reduce phases will not be bottlenecks and also the communication cost is low. This algorithm has been proved to be efficient and scalable by our evaluations shown in Section 4.2. However, there is still a restriction on operators \(\odot\) and \(\odot\), in practice. Under the assumption that the input data are larger than the storage capability of any single node in the cluster, that \(\oplus\) must not be \(\oplus\) (or any other that has similar effect), otherwise in the Map phases of the first MapReduce job, the result of \(v_{k} = t_{0} \odot q(x_{1}) \odot q(x_{2}) \odot \cdots \odot q(x_{m-1})\) may be too large to be stored in the \(\text{DistributedCache}\) nor be transported via networks, unless function \(g\) can filter out (returns an empty list) most of the elements of the input. If \(\odot\) is not \(\oplus\) but \(\oplus\), then whether the accumulate is efficient depends on the size of \(x_{k}\). Here \(x_{k} = p(x_{1}, w_{k}) \odot p(x_{2}, w_{k}) \odot q(x_{1}) \odot \cdots \odot p(x_{m-1}, w_{k}) \odot q(x_{m-1})\), and \(w_{k} = c \odot v_{1} \odot \cdots \odot v_{k-1}\). For function \(p\), it can filter out most of its input then using only one reducer in the second MapReduce will not be a big problem, otherwise we have to do special optimization for such case by using multiple reducers.

3.2.2 The Optimized MapReduce Implementation for Specialized Accumulation

If the emitted intermediate data are small enough, then they can be efficiently transferred to one reducer via network otherwise the computation will be very costive or the data are too large to be manipulated by only one reducer.

In order to improve the performance for some special cases such that (in the Map phase of the second MapReduce job), \(x_{k} = p(x_{1}, w_{k}) \oplus p(x_{2}, w_{k}) \odot q(x_{1}) \odot \cdots \odot p(x_{m-1}, w_{k}) \odot q(x_{m-1})\) is a long list (suppose the input is split to \(p\) sub-
lists), we have optimized the implementation. We do not group all output of Map phase to one reducer but use multiple reducers instead. The number of reducers can be adjusted to fit the practical problems and data. Same as the general case in Subsection 3.2.1, output of Map are sorted by $seg_k (k \in [1, p])$, but grouped to $t$ reducers. Intuitively, let $r = p/t$, reducer$k$ receives $[x_{k \cdot r + 1}, x_{k \cdot r + 2}, \ldots, x_{k \cdot r + r}]$, and emits $x_{k \cdot r + 1} \oplus x_{k \cdot r + 2} \oplus \cdots \oplus x_{k \cdot r + r} (k \in [0, r - 1))$. The reducer$k$ receives $[x_{0 \cdot r + 1}, x_{0 \cdot r + 2}, \ldots, x_{0}]$ and reads $[v_1, \ldots, v_0]$ from $\text{DistributedCache}$, and computes $w = c \odot v_1 \odot v_2 \odot \cdots \odot v_0$. At last the reducer$k$ emits $x_{k \cdot r + 1} \oplus x_{k \cdot r + 2} \oplus \cdots \oplus x_{k \cdot r + r} \oplus g(w)$. Each output from the $t$ reducers contains part of the final result.

### The Optimized Implementation for Scan

The scan skeleton is a special case of accumulate: $\text{scan} = [\{ \}. \odot \text{snd}, \oplus ]$, i.e., $g = [\}. \odot \text{snd}, \oplus = \oplus$, and $q = \text{id}$. The MapReduce implementation of scan can be optimized and efficiently computed, if $\odot$ is not $\oplus$. Because the result of $\text{scan}$ is $s_1 \oplus s_2 \oplus \cdots \oplus s_k$ we do not need the Reduce phase in the second $\text{MapReduce}$, and just let each mapper emit $(seg_k, s_k) (k \in [1, p])$. The $seg_k$ denotes the offset of sublist handled by the $k$th mapper, so that these pairs can be sorted by $seg_k$ and form the final result.

### 3.3 The Programming Interfaces

We provide two parallel skeletons $\text{scan}$ and $\text{accumulate}$ in our framework. These two skeletons and also the related binary operators are represented as Java classes, in the object-oriented style.

#### 3.3.1 Scan Interface

Listing 1 shows the representation Java class for $\text{scan}$. Users need to define the associative binary operator to create an instance of the $\text{scan}$ computation. There is an example of an associative binary operator $\text{Plus}$ in Listing 2, which adds two integers and returns the sum. The $\text{evaluate}$ method takes two arguments and returns one value. The $\text{method, the \text{runScanMR}}$ which takes two arguments — one is an instance of $\text{scan}$, and the other is the $args$ (the input, output paths given by users) from the $\text{main}$, — will execute the scan computation on the Hadoop cluster.

#### 3.3.2 Accumulation Interface

An accumulate can be defined by implementing the abstract class $\text{Accumulation}$ (as shown in Listing 3). There are five functions/operators according to the definition of the accumulate, and an accumulative parameter $c$. The Java class $\text{MapReduceExample}$ (Listing 4) which extends $\text{MRAccHelper}$ shows how to write the client code. Similarly to $\text{scan}$, the method $\text{createAccuIns}$ needs to be override, in which an instance of $\text{accumulate}$ is created. In the main function, the $\text{runAccMR}$ method should be invoked to execute the accumulate. The Listing 5 shows an example to define the $\text{elimSmaller}$ computation.

### 4. Programming Examples and Evaluations

We have developed several examples by using the parallel skeletons $\text{scan}$ and $\text{accumulate}$ provided by our framework, and evaluated them on Hadoop clusters.

#### 4.1 Example Programs

Table 1 lists five examples developed on our framework. More examples of accumulative computations can be found in the source packages of the framework. The $\text{scan} (+)$ is an application of prefix-sum on numbers using the binary operator $\oplus$. The

---

**Listing 1: Scan Representation**

```java
public class Scan<T> {
    public AssociativeBinaryOP<T> oplus;
    public Scan(AssociativeBinaryOP<T> op) {
        this.oplus = op;
    }
}
```

**Listing 2: An Example of Using the Scan Programming Interface**

```java
public class ScanExample extends MRScanHelper<Int> {
    public static void main(String[] args) throws Exception {
        public Int id() {
            return new Int(a.val + b.val);
        }
    }
}
```

**Listing 3: Accumulation Representation**

```java
public class Accumulation<T0, T1, T2> {
    public T2 accumulate( );
    public Int id() {
        return new Int(new Int(50));
    }
}
```

**Listing 4: An Example of Using the Accumulation Programming Interface**

```java
public class MapReduceExample extends MRAccHelper<Int, Int,Integer> {
    public void createAccuIns() {
        this.accumulate = new ElimSmallers(new T1);
        this.id() = new T1(5);}
    }
}
```

---

© 2014 Information Processing Society of Japan
Listing 5: Definition of Accumulation for ElimSmallers

```java
public class ElimSmallers extends Accumulation<Int, Int, IntList> {
    public ElimSmallers(Int x) {
        c = x; // new Int(50);
        oplus = new AccociBinaryOP<IntList>() {
            @Override
            public IntList evaluate(IntList left, IntList right) {
                if (left == null || left.get() == null) {
                    left = new IntList(new ArrayList<Int>());
                } else if (right == null && right.get() != null) {
                    left.get().addAll(right.get());
                }
                return left;
            }
            @Override
            public IntList id() {
                return new IntList();
            }
        };
        otimes = new AccociBinaryOP<Int>() {
            @Override
            public Int evaluate(Int left, Int right) {
                if (right < left) {
                    return left;
                } else {
                    return right;
                }
            }
            @Override
            public Int id() {
                return new Int(Integer.MIN_VALUE);
            }
        };
        g = new UnaryFunction<Int, IntList>() {
            @Override
            public IntList evaluate(Int obj) {
                return new IntList();
            }
        };
        p = new BinaryOperator<Int, Int, IntList>() {
            @Override
            public IntList evaluate(Int x, Int c) {
                ArrayList<Int> val = new ArrayList<Int>();
                if (x.get() < c.get()) {
                    return null;
                } else {
                    val.add(x);
                }
                return new IntList(val);
            }
        };
        q = new UnaryFunction<Int, Int>() {
            @Override
            public Int evaluate(Int da) {
                return da;
            }
        };
    }
}
```

elimSmallers, los, tagmatch and mps are those introduced in Section 2. Each program requires a particular type of input list, such as list of numbers, list of tags or list of pairs. Table 1 also gives the type of the input lists for each program.

4.2 Evaluation

We evaluated the performance and scalability of the example programs with manually generated data sets shown in Table 1.

We configured Hadoop (cdh3u5) clusters with up to 32 virtual machines (VMs) inside the EdibaseCloud system in National Institute of Informatics. Each VM has 2 CPUs (a CPU is one core of the Xeon E5530@2.4 GHz), 6 GB RAM. The total parallel-task slots in Hadoop are configured to be equal to the total number of CPUs in the cluster.\(^7\)

### Scalability

The experiment results are summarized in Figs. 3 and 4. Letting the input data be fixed size (shown as Table 2) while increasing the working nodes of the cluster (i.e., increasing VMs), all programs have almost twice speedup when the number of CPUs increases from 8 to 16. This indicates the good scalability of our framework. When the number of CPUs keeps increasing, the running-time becomes shorter and approaches to a constant value.

\(^7\) We made this configuration in order to simplify the analysis of scalability. In fact, optimizing the configurations of the Hadoop cluster, e.g., allowing more mappers running simultaneously in each VM, can obtain much better performance.
which is the time of fixed sequential parts computation and system overhead of Hadoop. As a summary, the relation between speedup \( y \) and number of CPUs \( x \) approximately fits to a linear curve \( y = Ax + B \), e.g., in case of scan (+), \( A = 6.49 \times 10^{-2} \), \( B = 0.779 \).

**In Comparison with Vanilla Hadoop Programs**

Finally we discuss the programmability and relative efficiency by comparing the programs written by using our framework and those written by directly using the Hadoop API. As mentioned before, programs written by using our \textit{accumulate} API will be transformed to programs that are exactly equivalent to those manually written by using vanilla Hadoop (i.e., without using our framework). The comparison we gave just shows the benefits that how much our framework saves programmers efforts and low overhead of the high-level abstraction in our framework.

For each problem in Table 1, we made a new version using only vanilla Hadoop, which implemented the same two-phases MapReduce algorithm in Section 3.2. Table 4 shows the comparison of length of source code and running times of the two classes of programs (on the same 64-CPU cluster). The source code is formatted by the Eclipse code formatter, and counted by using Google CodePro Analytix\(^8\). We used much larger data sets listed in Table 3 as input for the evaluation.

In Table 4, the column \textit{Lines (vanilla)} is for the lengths of programs implemented without using our framework, and the column \textit{Lines} is for the lengths of programs implemented by using our framework. The column \textit{Time (vanilla)} and the column \textit{Time} are running times of the two versions.

The results show that all vanilla Hadoop programs are much longer than programs written by using our \textit{accumulate} API (3.2–5.6 times longer). The system overhead caused by the generic abstraction and wrapping of Hadoop API can be almost negligible. In addition, programs implemented by using our framework can still handle the larger input data (nearly 20 times larger compared to each data set in Table 2) very well.

Generally, the main difficulty for a Hadoop programmer to implement a MapReduce algorithm for problems such as \textit{elimSmallerex}, is about finding the scalable divide-and-conquer algorithm. Furthermore, even when he knows the algorithm, the implementation of the cumbersome Hadoop code is still probably very time consuming.

### Table 3

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Length</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>(1 \times 10^6 \times 2^{20})</td>
<td>195.32 GB</td>
</tr>
<tr>
<td>Pairs</td>
<td>(1 \times 10^6 \times 2^{20})</td>
<td>193.21 GB</td>
</tr>
<tr>
<td>Tags</td>
<td>(1 \times 10^3 \times 2^{20})</td>
<td>195.32 GB</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Problems</th>
<th>\textit{Lines (vanilla)}</th>
<th>\textit{Lines}</th>
<th>\textit{Time (vanilla)}</th>
<th>\textit{Time}</th>
</tr>
</thead>
<tbody>
<tr>
<td>scan (+)</td>
<td>163</td>
<td>29</td>
<td>1,995 s</td>
<td>1,988 s</td>
</tr>
<tr>
<td>\textit{elimSmallerex}</td>
<td>348</td>
<td>81</td>
<td>1,583 s</td>
<td>1,587 s</td>
</tr>
<tr>
<td>los</td>
<td>346</td>
<td>107</td>
<td>2,902 s</td>
<td>2,903 s</td>
</tr>
<tr>
<td>tagmatch</td>
<td>347</td>
<td>75</td>
<td>1,793 s</td>
<td>1,791 s</td>
</tr>
<tr>
<td>mps</td>
<td>347</td>
<td>75</td>
<td>1,793 s</td>
<td>1,791 s</td>
</tr>
</tbody>
</table>

*8 https://developers.google.com/java-dev-tools/codepro/

5. Related Work

Algorithmic skeletons for parallel programming have been well studied from 1989 [9], and a lot of frameworks have been developed to provide those algorithmic skeletons [3], [7], [8], [21], [24]. Not only the programming frameworks, we have studied a systematic way to develop parallel programs using those skeletons. In particular, \textit{scan} is a very useful skeleton because it enables us to reuse the partial results in reduce. For example, in [19] we showed that we can solve a set of maximum marking problems on lists with \textit{scan} skeletons. We also showed that we can perform the matching of a regular expression over a single large document in parallel [20]. With the \textit{scan} or \textit{accumulate} computations developed in this paper, we can extend the technique to retrieve substrings.

The MapReduce was firstly introduced by Google to handle very large raw data form Internet. The programming model of MapReduce is inspired by the concepts from functional programming [10], [22]. MapReduce gains a big success in both industry and academia, because of its functionality and simplicity. But there is still a gap between a nontrivial problem and the MapReduce paradigm because MapReduce programming model is relatively low-level that leads many difficulties in practical programming. Many studies such as Sawzall [25], PigLatin [12], and DryadLINQ [27] tried to provide more user-friendly domain specific languages to address the programmability problem, and our previous work [11], [23] introduced the approaches of calculation theorems for list homomorphisms into MapReduce, for the similar purpose but in different methodology. The \textit{accumulate} as an algorithmic parallel computation pattern has been implemented by using MPI [18] and now is a part of the Sketo library [24] that provides a simple programming interface and efficient parallel implementation. To our knowledge, there was no completable MapReduce implementation for the accumulative computing\(^9\) before present study.

### 6. Conclusion

The research on parallel skeletons [9], [15], [18] and list homomorphisms [23] illustrates a systematic and constructive way to high-level parallel programming, by which this work was inspired. In this paper, we have described how to implement and use the two parallel skeletons \textit{scan} and \textit{accumulate}, in MapReduce. We provide a Hadoop-based framework with high-level programming interfaces. A large class of computations that are originally difficult to be programmed directly with MapReduce APIs can be easily implemented on our framework and enjoy the merits of MapReduce. The implementation is efficient and scalable, because it is based on the result of applying the fusion transformation to \textit{accumulate}, which eliminates unnecessary intermediate data structures.

Although we limited our discussion to lists in this paper, the diffusion theorem can be extended to trees [16] and other general recursive data types. We plan to implement more algorithmic

---

\(^9\) Some MapReduce frameworks, Pig [12] and Spark [28] also have so-called “\textit{accumulate}” interface, but such “\textit{accumulate}” is not in the same sense of the functional programming pattern \textit{accumulate}.
skeletons on MapReduce to simplify the parallel programming and large-scale data processing.

Acknowledgments We would like to thank Hideya Iwasaki, Shigeyuki Sato, and Le Duc Tung, for their helpful advices and constructive comments on this paper. This work was partially supported by JSPS KAKENHI Grant Number 24700025, and JST/ANR Grant Numbers 10102704/ANR-2010-INTB-0205-02.

References

Yu Liu is a Ph.D. student in the Graduate University for Advanced Studies (Japan), and is a Research Assistant of National Institute of Informatics (NII). He received his B.S. from Xi’an Jiao Tong University (China) in 2004, and began doctoral study since 2009. Before entered Ph.D. course of the Graduate University for Advanced Studies, he worked as a sofer engineer in China and Japan. His current research interest is parallel programming and parallel algorithms. He is a member of JSSST.

Kento Emoto is an Assistant Professor of Kyushu Institute of Technology (Kyutech) in Japan. He received his B.E., M.S. and Ph.D. from The University of Tokyo in 2004, 2006 and 2009, respectively. He was a Researcher (2007–2009) and an Assistant Professor (2009–2013) in The University of Tokyo, before joining Kyutech as an assistant professor in 2013. His main interest is in programming languages and software engineering in general, and parallel programming and algorithm synthesis in particular. He is a member of ACM, JSSST, IPSJ.

Kiminori Matsuoka is an Associate Professor of Kochi University of Technology in Japan. He received his B.E., M.S. and Ph.D. from The University of Tokyo in 2001, 2003 and 2007, respectively. He was an Assistant Professor (2005–2009) in The University of Tokyo, before joining Kochi University of Technology as an Associate Professor in 2009. His research interest is in parallel programming and algorithm derivation. He is a member of ACM, JSSST, IEEE.
Zhenjiang Hu is a Professor of National Institute of Informatics (NII) in Japan. He received his B.S. and M.S. from Shanghai Jiao Tong University in 1988 and 1991 respectively, and Ph.D. degree from The University of Tokyo in 1996. He was a Lecture (1997–1999) and an Associate Professor (2000–2007) in The University of Tokyo, before joining NII as a Full Professor in 2008. His main interest is in programming languages and software engineering in general, and functional programming, parallel programming and bidirectional model-driven software development in particular. He is a member of ACM, JSSST, IPSJ.