Future Generation Computer Systems 26 (2010) 1200–1214

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A data placement strategy in scientific cloud workflows

Dong Yuan*, Yun Yang, Xiao Liu, Jinjun Chen

Faculty of Information and Communication Technologies, Swinburne University of Technology, Hawthorn, Melbourne 3122, Australia

ARTICLE INFO

Article history:
Received 30 May 2009
Received in revised form
26 November 2009
Accepted 1 February 2010
Available online 18 February 2010

Keywords:
Data management
Scientific workflow
Cloud computing

ABSTRACT

In scientific cloud workflows, large amounts of application data need to be stored in distributed data centres. To effectively store these data, a data manager must intelligently select data centres in which these data will reside. This is, however, not the case for data which must have a fixed location. When one task needs several datasets located in different data centres, the movement of large volumes of data becomes a challenge. In this paper, we propose a matrix based k-means clustering strategy for data placement in scientific cloud workflows. The strategy contains two algorithms that group the existing datasets in k data centres during the workflow build-time stage, and dynamically clusters newly generated datasets to the most appropriate data centres – based on dependencies – during the runtime stage. Simulations show that our algorithm can effectively reduce data movement during the workflow's execution.

1. Introduction

Running scientific workflow applications usually need not only high performance computing resources but also massive storage [1]. In many scientific research fields, like astronomy [2], high-energy physics [3] and bio-informatics [4], scientists need to analyse terabytes of data either from existing data resources or collected from physical devices. During these processes, similar amounts of new data might also be generated as intermediate or final products [1]. Workflow technologies are facilitated to automate these scientific applications. Scientific workflows are typically very complex. They usually have a large number of tasks and need a long time for execution. Nowadays, popular scientific workflows are deployed in grid systems [3] because they have a high performance and massive storage. However, building a grid system is extremely expensive and it is not available for scientists all over the world to use.

The emergence of cloud computing technologies offers a new way to develop scientific workflow systems. Since late 2007 the concept of cloud computing was proposed [5] and it has been utilised in many areas with some success [6–9]. Cloud computing is deemed as the next generation of IT platforms that can deliver computing as a kind of utility [10]. Foster et al. made a comprehensive comparison of grid computing and cloud computing [11]. Some features of cloud computing also meet the requirements of scientific workflow systems. First, cloud computing systems can provide a high performance and the massive storage required for scientific applications in the same way as grid systems, but with a lower infrastructure construction cost among many other features, because cloud computing systems are composed of data centres which can be clusters of commodity hardware. Second, cloud computing systems offer a new paradigm so that scientists from all over the world can collaborate and conduct their research together. Cloud computing systems are based on the Internet, and so are the scientific workflow systems deployed on the cloud. Dispersed computing facilities (like clusters) at different institutions can be viewed as data centres in the cloud computing platform. Scientists can upload their data and launch their applications on scientific cloud workflow systems from anywhere in the world via the Internet. As all the data are managed on the cloud, it is easy to share data among scientists. Research into doing science on the cloud has already commenced such as early experiences like the Nimbus [12] and Cumulus [13] projects. The work by Deelman et al. [14] shows that cloud computing offers a cost-effective solution for data-intensive applications, such as scientific workflows [15].

By taking advantage of cloud computing, scientific workflow systems could gain a wider utilisation; however they will also face some new challenges, where data management is one of them. Scientific applications are data intensive and usually need collaborations of scientists from different institutions [16], hence the application data in scientific workflows are usually distributed and very large. When one task needs to process data from different data centres, moving the data becomes a challenge [1]. Some application data are too large to be moved efficiently, some may have fixed locations that are not feasible to be moved and some may have to be located at fixed data centres for processing, but this is only one aspect of this challenge. For the application data that are flexible to be moved, we also cannot move them whenever and wherever we want, since in the cloud computing platform, data centres may belong to different cloud service providers so that data movement would result in costs. Furthermore, the infrastructure of the

* Corresponding author.
E-mail addresses: dyuan@swin.edu.au (D. Yuan), yyang@swin.edu.au (Y. Yang), xliu@swin.edu.au (X. Liu), jchen@swin.edu.au (J. Chen).

Denotations

d_i	dataset
D	set of datasets
D_i	set of datasets in a partition
fd_i	fixed location dataset
FD	set of fixed location datasets
t_i	workflow task
T	set of workflow tasks
T_i	set of workflow tasks that will use dataset d_i
dc_i	data centre
DC	set of data centres
p_i	partition of datasets
P	set of partitions
s_i	size of a dataset
cs	size of a data centre
ds	size of a partition
ps	size of a set of partitions
FP	set of partitions that have fixed location datasets
NFP	set of partitions that do not have fixed location datasets
DM	dependency matrix
CM	clustered dependency matrix
CM_i	sub clustered dependency matrix
CM_T	the top sub clustered dependency matrix after one binary partition
CM_B	the bottom sub clustered dependency matrix after one binary partition
GM	global measure of BEA transformation
PM	global measure of binary partition
dep_{ij}	dependency between datasets d_i and d_j
dc_dep_{ij}	dependency between dataset d_i and data centre dc_j
K	set of data centres with placement of datasets
λ_{ini}	initial storage usage parameter of data centres
$\lambda_{\max}$	maximum storage usage parameter of data centres

cloud computing systems is hidden from their users. They just offer the computation and storage resources required by users for their applications. The users do not know the exact physical locations where their data are stored. This kind of model is very convenient for users, but remains a big challenge for data management to scientific cloud workflow systems.

In this paper, we propose a matrix based *k*-means clustering strategy for data placement in scientific cloud workflow systems. Scientific workflows can be very complex, one task might require many datasets for execution; furthermore, one dataset might also be required by many tasks. If some datasets are always used together by many tasks, we say that these datasets are dependant on each other. In our strategy, we try to keep these datasets in one data centre, so that when tasks were scheduled to this data centre, most, if not all, of the data they need are stored locally.

Our data placement strategy has two algorithms, one for the build-time stage and one for the runtime stage of scientific workflows. In the build-time stage algorithm, we construct a dependency matrix for all the application data, which represents the dependencies between all the datasets including the datasets that may have fixed locations. Then we use the BEA algorithm [17] to cluster the matrix and partition it that datasets in every partition are highly dependent upon each other. We distribute the partitions into *k* data centres, where the partitions have fixed location datasets are also placed in the appropriate data centres. These *k* data centres are initially as the partitions of the *k*-means algorithm at the runtime stage. At the runtime, our clustering algorithm deals with the newly generated data that will be needed by other tasks. For every newly generated dataset, we calculate its dependencies with all *k* data centres, and move the data to the data centre that has the highest dependency with it.

By placing data with their dependencies, our strategy attempts to minimise the total data movement during the execution of workflows. Furthermore, with the pre-allocation of data to other data centres, our strategy can prevent data gathering to one data centre and reduces the time spent waiting for data by ensuring that the relevant data are stored locally.

The remainder of the paper is organised as follows. Section 2 presents the related work. Section 3 gives an example and analyses the research problems. Section 4 introduces the basic strategy of our algorithms. Section 5 presents the detailed steps of the algorithms in our data placement strategy. Section 6 demonstrates the simulation results and the evaluation. Finally, Section 7 addresses our conclusions and future work.

2. Related work

Data placement of scientific workflows is a very important and challenging issue. In traditional distributed computing systems, much work about data placement has been conducted. In [18], Xie proposed an energy-aware strategy for data placement in RAID-structured storage systems. Stork [19] is a scheduler in the Grid that guarantees that data placement activities can be queued, scheduled, monitored and managed in a fault tolerant manner. In [20], Cope et al. proposed a data placement strategy for urgent computing environments to guarantee the data's robustness. At the infrastructure level, NUCA [21] is a data placement and replication strategy for distributed caches that can reduce the data's access latency. However, none of them focuses on reducing the data's movement between data centres on the Internet. As cloud computing has become more and more popular, new data management systems have also appeared, such as the Google File System [22] and Hadoop [23]. They all have hidden infrastructures that can store the application data independent of the users' control. The Google File System is designed mainly for Web search applications, which are different from workflow applications. Hadoop is a more general distributed file system, which has been used by many companies, such as Amazon and Facebook. When you push a file to a Hadoop File System, it will automatically split this file into chunks and randomly distribute these chunks in a cluster. Furthermore, the Cumulus project [13] introduced a scientific cloud architecture for a data centre. And the Nimbus [12] toolkit can directly turn a cluster into a cloud and it has already been used to build a cloud for scientific applications. Within a small cluster, data movement is not a big problem, because there are fast connections between nodes, i.e. the Ethernet. However, the scientific cloud workflow system is designed for scientists to collaborate, where large scale and distributed applications need to be executed across several data centres. The data movement between data centres may cost a lot of time, since data centres are spread around the Internet with limited bandwidths. In this work, we try to place the application data based on their dependencies in order to reduce the data movement between data centres.

Data transfer is a big overhead for scientific workflows [24]. Though popular scientific workflow systems have their data management strategies, they do not focus on reducing data movement. For the build-time stage, these systems mainly focus on the data modelling methods. For example, Kepler [3] has an actor-oriented data modelling method that works for large data in a grid environment, Taverna [4] and ASKALON [25] have their own process definition language to represent their data flows. For the runtime stage, most of the scientific workflow systems adopt some data grid systems for their data management. For examples, Kepler uses the SRB [26] system, while Pegasus [2] and Triana [27] adopt the RLS system [28], Gridbus [29] has a grid service broker [30] where all data are deemed as important resources. Data grids primarily deal with providing services and an infrastructure for distributed data-intensive applications that need to access, transfer, and modify massive datasets stored in distributed storage resources [31]. However, these systems do not consider the dependencies between data in scientific workflows either at the build-time or the runtime and they also cannot reduce the data's movement. Some researches in grid computing have addressed the importance of the data's dependency for the large-scale scientific applications, although they did not focus on the workflow data management. The Filecules project [32] groups the files based on the dependencies. Using real workload experiments data, the authors demonstrated that the filecules grouping is a reliable and useful abstraction for data management in a science Grid. BitDew [33] is a distributed data management system for a desktop Grid. Different from data centres in the cloud that aim to provide services to users, the desktop Grid aims to make use of the idle computing and storage resources in the desktop computers. In BitDew, the data placement dependency is denoted by a data attribute called "affinity", which is pre-defined by users. However, in cloud computing, all the application data are hosted in the data centres, where anyone can use the cloud services and upload their data. Letting users define the data dependencies for the scientific cloud workflows is clearly impractical.

The closest workflow research to ours is the Pegasus workflow system which has proposed some data placement strategies [34,35] based on the RLS system. The strategies are: first, pre-allocate the required data to the computation resource where the task will be executed; second, dynamically delete the data that will no longer be used by tasks. These strategies are only for the runtime stage of scientific workflows and can effectively reduce the overall execution time and the storage usage of the workflows. Furthermore, in [36], the authors proposed a data placement scheduler for the distributed computing systems. It guarantees the reliable and efficient data transfer with different protocols. These works mainly focus on how to move the application data, and they cannot reduce the total data movement of the whole system. However, our work aims to reduce the data's movement. Our strategy is for both the build-time and the runtime stages of scientific workflows and we design specific algorithms to automatically place and move the application's data.

In cloud computing systems, the infrastructure is hidden from users. Hence, for most of the application's data, the system will decide where to store them. Dependencies exist among these data. In this paper, we initially adapt the clustering algorithms for the data movement based on the data's dependency. Clustering algorithms have been used in pattern recognition since 1980s [37], which can classify patterns into groups without supervision. Today they are widely used to process data streams [38]. In many scientific workflow applications, the intermediate data movement is in the data stream format and the newly generated data must be moved to the destination in real-time. We adapt the k-means clustering algorithm for the data's placement. When new data is generated by a task, we dynamically calculate the dependencies of the new data with the K data centres, and move the new data to the centre with highest dependency. The simulation results of this paper show that with our data placement strategy, the data movement between the data centres is significantly reduced compared to a random data placement.

3. Scientific cloud workflow data management

3.1. A motivating example

Scientific applications often need to process terabytes of data. For example, the ATNF[1] Parkes Swinburne Recorder (APSR) [39] is a next-generation baseband data recording and processing system currently under development in collaboration by the Swinburne University of Technology and the ATNF. The data from the APSR streams at a rate of one gigabyte per second. The researchers at Parkes process the data with a local cluster of servers and do their research. All the data are stored locally at Parkes and they are not available to other institutions. If researchers at other institutions need the data resources from the Parkes Radio Telescope, they have to contact the researchers at Parks and request for the data. Researchers at Parkes will check the local repositories to see if the existing data resources could fulfill the requirements. In this situation communications often suffer from low efficiency because researchers are from different projects and the requirements are usually complex. Sometimes researchers even have to go to Parkes and bring back the data that they need on hard disks. Sharing data resources in this manner is obviously inefficient and hence not desirable.

With cloud computing technologies, we can turn the Parkes cluster into a data centre on a cloud computing platform that can offer services to researchers all over the world. The cloud computing platform is built on the Internet, which is how the data centres are connected to each other. All the data are managed by the cloud data management system. The researchers can access the existing data resources, upload application data and launch their applications via the cloud service. By doing this, the resources at Parkes will be fully utilised, since data can be sent to other data centres for different applications as needed. On the other hand, the researchers at Parkes will be able to do more scientific research by retrieving useful data from other data centres around the world. All these data sending and retrieving operations are hidden from the researchers. In other words, via a cloud computing platform, researchers can utilise data resources from other institutions without knowing where the data are physically stored. Hence, on a cloud computing platform, data centres should have the ability to host each other's data. For example, if some particular data at Parkes are frequently retrieved by another data centre, the system will store these data on that data centre instead. Furthermore, if many applications at Parkes need the same data from another data centre, the system will also move those data to Parkes for storage.

The Parkes Radio Telescope was set up in 1961. For over 40 years, the Parkes cluster has accumulated a large amount of data resources in different formats and sizes. Normally, data can be moved to other data centres, but if the size of the data is very large, moving them via the Internet will be inefficient. To transport terabytes of data, the most efficient way is for a delivery company to ship the hard disks [40]. If an application needs the majority of its data from Parkes, it is preferable that it is executed locally and retrieves data from elsewhere. For example, some research projects may need to process the raw data recorded from the telescope by the APSR, in order to get some specific results.

3.2. Problem analysis

Scientific cloud workflows run on the cloud platform, which is composed of many distributed data centres on the Internet (like the Parkes cluster) and each connection between the data centres has a limited bandwidth. Tasks sometimes need to process more than one dataset that may be stored in different data centres. Because of the bandwidth constraints, the movement of datasets between data centres would be the bottle-neck of the system. In [41], the authors proposed a new protocol for data transportation that could provide gigabits of bandwidth. However, it has not been widely supported by the Internet. The popular cloud systems, such as Amazon EC2 [42], still have a limited bandwidth [43]. It charges \$0.10 to \$0.15 per gigabyte to move data in to and out of Amazon Web Services over the Internet. Another approach to deal with the bottle-neck of large data transfer is to divide the tasks,

[1] ATNF refers to the Australian Telescope National Facility.

i.e. for the tasks that need to process many distributed datasets, we split them to many smaller and parallel sub-tasks, and schedule them to different datasets. Map-Reduce technology [44] is a typical and successful paradigm. It has gained great success in the Google File System and Hadoop, as well as in scientific applications [45]. However, Map-Reduce is more applicable to be used within one data centre, since it needs a huge interconnected bandwidth, such as the shuffle step that occurs between the Map procedure and the Reduce procedure. Furthermore, in scientific applications, many tasks must use more than one dataset together and cannot be further divided, such as the All-Pairs problem [9]. Therefore, data movement is inevitable. In light of this, we have to place the datasets that are needed by the same task in the same data centre as much as possible, so as to minimise data movement when the task is executed. The placement of datasets among data centres is not trivial.

Normally, a cloud computing system needs to decide in which data centres the application's data are stored. Most datasets are flexible about where they are stored since they are independent of users. The cloud computing system can automatically store the application's data based on some data placement strategies. However, in scientific cloud workflow systems, some data are not so flexible. They have to be stored in some particular data centres due to different reasons. Some common scenarios are demonstrated below.

First, some data may need to be processed by special equipment. In some scientific projects, many special types of equipment are utilised. Some data can only be processed by particular equipment since they are in certain formats, e.g. the signal from the Parkes Radio Telescope can only be processed by the equipment at Parkes, such as the ASPR. These data have to be stored where the required equipment is located.

Second, some data are naturally distributed and too large to be moved efficiently. For example, the raw data files recorded by the ASPR are usually terabytes or even petabytes in size. They are naturally stored in Parkes, and impossible to move to other locations via the Internet.

Another reason that some data must be placed at a particular data centre is about the ownership. Data are considered as an important and valuable resource in many scientific projects. The cloud computing platform offers a new paradigm for cooperation that institutions can easily share their valuable data resources by placing a charge on them. So the data with limited access rights have to be stored in particular data centres.

No matter what the reason that the data must be stored in a particular data centre, we call these datasets as *fixed location datasets* in general. As such, we call the datasets that the system can flexibly decide where to store *flexible location datasets*. The data placement strategy not only has to place the flexible location datasets, but also has to take into account the impact of the fixed location datasets. Some challenges exist in the data placement strategy as discussed below.

First, in scientific workflows, both tasks and datasets could be numerous and make up a complicated many-to-many relationship. One task might need many datasets and one dataset might be needed by many tasks. Furthermore, new datasets will be generated during the workflow execution. One dataset generated by a task might be used by several later tasks. So the data placement strategy should be based on these data dependencies.

Second, the scientific cloud workflow system is a dynamic computing environment. Many workflow instances will run in the system simultaneously. Some instances might need a long time execution and some might be short. New workflow instances could be deployed to the system and completed instances could be removed from the system at any time. So the relationships between datasets and tasks will change often and the placement of datasets has to be changed accordingly.

Third, the data management in scientific cloud workflow systems is opaque to users, that means users do not know where and how the data been stored. In the cloud environment, users only pay for the computation and storage resources that they need and give the application data to the system for processing. Because the cloud systems are built on the service oriented architecture (SOA), the users just use the dynamic cloud services and do not know the infrastructure of the system. Hence, the data placement has to be automatic.

4. Basic strategies for data placement

For scientific workflow data management, there are two types of data we have to deal with.

First is the *existing data* that exists before the workflow's execution starts. This type of data mainly includes the resource data from the existing file systems or databases and the application's data from users as input for processing or analysis.

Second is the *generated data* that are generated during the workflow's execution. This type of data mainly includes the newly generated mediate and result data, as well as the streaming data dynamically collected from scientific devices during the workflow's execution.

We propose this taxonomy because we will treat these two types of data at the workflow's build-time and runtime respectively with different algorithms. This taxonomy only indicates the generation time of the datasets. When the generated data moves to a data centre and is stored, it becomes existing data. The most important common feature is that both types of data might be very large. They cannot and should not be stored and moved wherever and whenever we want, since the cloud system has the bandwidth constraints.

The application's data of the scientific workflow could also have a variety of formats (e.g. XML data, complex objects, raw data files, tables in relational databases). But in this paper, we do not consider the structure of the data, since it is not the main focus of this paper and we will treat all data in the same way.

In scientific workflows, moving data to one data centre will cost more than scheduling tasks to that centre [23]. Hence, our basic strategy is to have a reasonable placement of data in distributed data centres first, so that when tasks are scheduled to the appropriate data centres, almost all the datasets they need are in local storage. In this work we analyse the dependencies between datasets. Based on this dependency, we adapt the *k*-means clustering algorithm to cluster datasets to the proper data centres.

In scientific cloud workflow systems, many workflow instances will run simultaneously, each of which has complex structures. Large numbers of tasks will access large numbers of datasets and produce large output data. In order to execute a task, all required datasets must be located on the same data centre, and this may require some movement of datasets. Furthermore, if two datasets are always used together by many tasks, they should be stored together in order to reduce the frequency of the data movement. Here, we say that these two datasets have a dependency. In other words, two datasets are said to be dependent on each other if they are both used by the same task. The more tasks there are that use the same datasets, the higher the dependency between those datasets. We denote the set of datasets as $D$and the set of tasks as T.[2] To represent this dependency, we give every dataset a task set in addition to its size. So, every dataset $d_i \in D$ has two attributes denoted as $\langle T_i, s_i \rangle$, where $T_i \subset T$ is the set of tasks that will use the dataset d_i, s_i denotes the size of d_i. Furthermore, we use $dependency_{ij}$ to denote the dependency between the datasets d_i and d_j. We say

2 All the denotations are listed at the end of the paper.

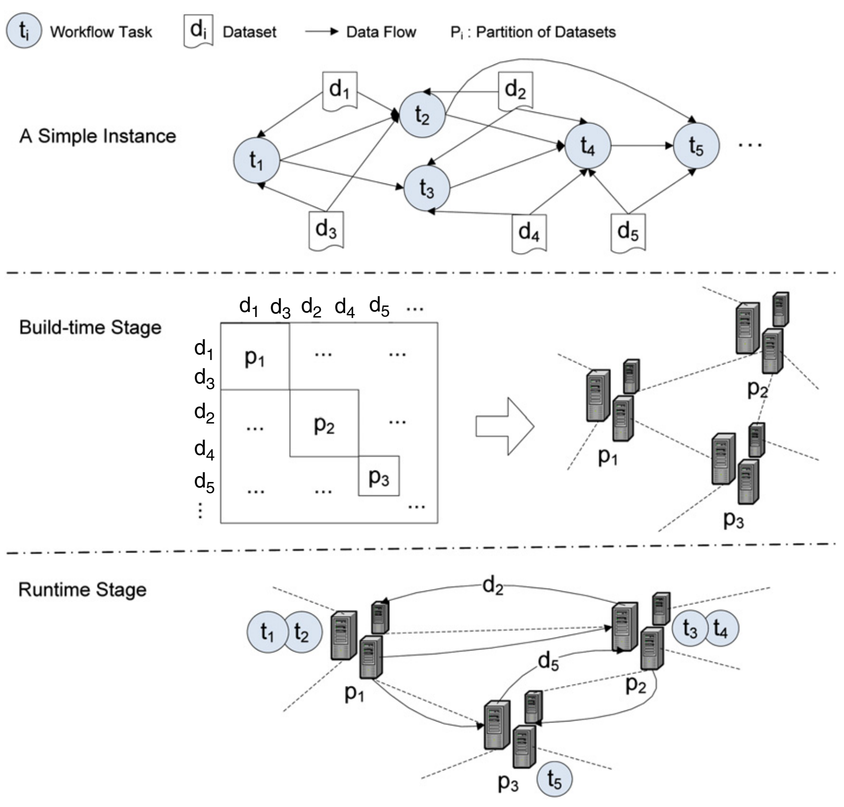


Fig. 1. Example of data placement.

that the datasets d_i and d_j have a dependency if there are tasks that will use d_i and d_j together and the quantity of this dependency is the number of tasks that use both d_i and d_j.

$$\text{dependency}_{ij} = \text{Count}\left(T_i \cap T_j\right).$$

In this work, our k-means clustering data placement strategy is based on this dependency that can cluster the datasets into different data centres. The strategy has two stages: build-time and runtime.

At the build-time stage, the main goal of the algorithm is to set up k initial partitions for the k-means algorithm. We use a matrix based approach to cluster the existing datasets into k data centres as the initial partitions.

At the runtime stage, the main goal of the algorithm is to cluster the newly generated datasets to one of the k data centres based on their dependencies, which will be calculated dynamically.

We have to design different algorithms for the build-time and runtime stages to treat the existing data and generated data respectively, mainly because of the dynamic nature of the cloud environment. Even though we know the size and related tasks of the datasets that will be generated during the workflow's execution, it is not practical to calculate their dependencies and assign them a data centre at the build-time stage. This is because the scientific workflows have a large number of tasks and need a long time for their execution. It is very hard to predict when a certain dataset will be generated in a dynamic cloud environment. If we assign the generated data a data centre at the build-time stage, then when the data are actually generated the data centre might have not enough available storage to store them. Furthermore, it is impractical and inefficient to reserve the storage for the generated data at the build-time stage. This is because the data might not be generated until the end of the scientific workflow and it would be a waste of the reserved storage space during this time.

5. Matrix based *k*-means clustering strategy for data placement

In this section we will intricately discuss our data placement strategy. In Fig. 1, there is an example of a simple workflow instance, and it shows the two stages of our strategy. The data flows in the workflow instance, for example, from dataset d_1 to tasks t_1 and t_2 mean that d_1 will be used by both t_1 and t_2; and data flows from t_1 to t_2 and t_3 mean that the dataset generated by t_1 will be used by both t_2 and t_3. During the build-time stage, we partition the existing datasets into several partitions, denoted as $p_1, p_2 \dots p_n$, based on their dependencies, and distribute these partitions into different data centres. During the runtime stage, tasks may retrieve datasets from other data centres as needed, and we also pre-allocate generated datasets to the appropriate data centres.

5.1. Build-time stage algorithm

During the build-time stage, we use a matrix model to represent the existing data. We pre-cluster the datasets by transforming the matrix, and then distributing the datasets to different data centres as the initial partitions for the k-means clustering algorithm, to be used during the runtime stage. The build-time stage algorithm has two steps and the pseudocode is shown in Fig. 4.

Step 1: Setup and cluster the dependency matrix.

First, we calculate the data dependencies of all the datasets and build up a dependency matrix DM (Line 3 in Fig. 4), where DM's element DM_{ij} = dependency$_{ij}$. dependency$_{ij}$ is the dependency value between datasets d_i and d_j, as we defined in the previous section. It can be calculated by counting the tasks in common between the task sets of d_i and d_j, which are denoted as T_i and T_j. Specially, for the elements in the diagonal of DM, each value means the number of tasks that will use this dataset. In our algorithm,

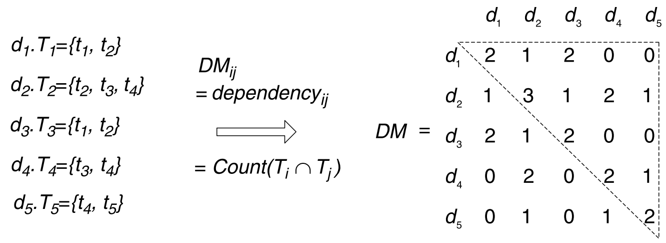


Fig. 2. Build up dependency matrix.

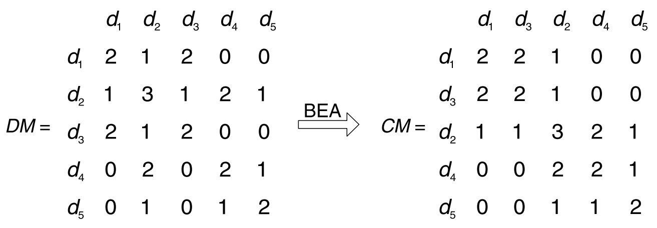


Fig. 3. BEA transformation of dependency matrix.

DM is an $n \times n$ symmetrical matrix where n is the total number of existing datasets. If we take the simple workflow instance in Fig. 1 as an example (with only 5 datasets, namely d_1 to d_5, in the system initially), the dependency matrix *DM* is shown in Fig. 2.

The dependency matrix (i.e. *DM*) is dynamically maintained at the runtime. When new datasets are generated by tasks or added to the system by users, we calculate their dependencies with all the existing datasets and add them to *DM*.

Next, we use the BEA (Bond Energy Algorithm) to transform the dependency matrix *DM* (Line 4 in Fig. 4). The BEA was proposed in 1972 [17] and has been widely utilised in distributed database systems for the vertical partition of large tables [46]. It is a permutation algorithm that can group similar items together in the matrix by permuting the rows and columns. In our work, it takes the dependency matrix (*DM*) as input, and generates a clustered dependency matrix (*CM*). In *CM*, the items with similar values are grouped together (i.e. large values with other large values, and small values with other small values). We define a global measure (*GM*) of the dependency matrix:

$$GM = \sum_{i=1}^{n}\sum_{j=1}^{n} DM_{ij}(DM_{i,j-1} + DM_{i,j+1}).$$

The permutation is done in such a way as to maximise this measure. The detailed algorithm of permutation can be found in [46]. Fig. 3 shows the *CM* of the example *DM* after the BEA transformation.

In this step, we do not consider the difference between fixed location datasets and flexible location datasets. If there are some fixed location datasets in the system, they will be arbitrarily scattered in the columns and rows of the dependency matrix, since we built up the matrix by calculating the dependencies between all the datasets. After the BEA transformation, all the datasets, including the fixed location datasets, are clustered by their dependencies.

Step 2: Partition and distribute datasets.

In this step we will distribute the datasets to the data centres as the initial k partitions for the k-means clustering algorithm at the runtime stage. We denote the set of data centres as *DC*. As shown in Fig. 1, we partition the clustered dependency matrix and place the corresponding datasets to different data centres. However, each dataset d_i has a size s_i and each data centre dc_j also has a storage capacity denoted as cs_j. To find the best partitioning of datasets matching the data centres' storage is an NP-hard problem, since it could be reduced to the Knapsack Packing Problem. Here, we develop a recursive binary partitioning algorithm to find the approximate best solution.

First, we partition *CM* into two parts $\{d_1, d_2 \dots d_p\}$ and $\{d_{p+1}, d_{p+2} \dots d_n\}$, which maximises the following measurement:

$$PM = \sum_{i=1}^{p}\sum_{j=1}^{p} CM_{ij} * \sum_{i=p+1}^{n}\sum_{j=p+1}^{n} CM_{ij} - \left(\sum_{i=1}^{p}\sum_{j=p+1}^{n} CM_{ij}\right)^2.$$

This measurement, *PM*, means that the datasets in each partition have higher dependencies with each other and lower dependencies with the datasets in the other partitions. Based on this measure we can simply calculate all the *PMs* for $p = 1, 2 \dots n-1$, and choose p such that it has the maximum *PM* value as the partition point.

After one partition, the *CM* forms two new clustered matrices, we denote the top one as CM_T, which contains the dependencies of datasets $D_T = \{d_1, d_2 \dots d_p\}$ and the bottom one as CM_B, which contains the dependencies of datasets $D_B = \{d_{p+1}, d_{p+2} \dots d_n\}$. Every clustered matrix represents a partition of datasets and we denote the total size of the datasets it contains as $ds = \sum_{i=1}^{n} s_i$. Hence the ds for CM_T and CM_B are $ds_T = \sum_{i=1}^{p} s_i$ and $ds_B = \sum_{i=p+1}^{n} s_i$ respectively.

Next, we distribute the datasets to the data centres by recursively partitioning the clustered dependency matrix.

For each of the data centres, we introduce a percentage parameter λ_{ini} to denote the initial usage of their storage capacity, which means that the initial size of the datasets in the data centre dc_i could not exceed $cs_i * \lambda_{ini}$. The reason we cannot fill the data centre with their maximum storage is that in scientific workflows, the generated data can also be very large. We have to reserve sufficient space in the data centres to store those data during the workflow's execution. λ_{ini} is an experience parameter. The value of λ_{ini} should depend on what kinds of applications are running on the system, because the generated data of different applications might have different sizes. Furthermore, we also assume that the data centres can host all the application data in the system, i.e. $\sum_{i=1}^{n} s_i < \sum_{i=1}^{m} (cs_i * \lambda_{ini})$.

To distribute the datasets, we have to examine whether there are fixed location datasets in the system (Line 5 in Fig. 4). If the system does not have fixed location datasets (Line 35 in Fig. 4), we will recursively partition the sub-matrices CM_T and CM_B until the size of the sub-matrix can fit into one of the data centres' initial storage size limits ($ds <= cs_i * \lambda_{ini}$). Then we distribute the datasets in this sub-matrix into this data centre, and add the reference of this data centre (dc_i) to K, where K is a set of data centres. When the partitioning of *CM* finishes, all the initial datasets are moved to proper data centres. We take the data centres in K as the initial partitions of the k-means clustering algorithm.

If there are fixed location datasets in the system, the distribution process is more complicated. For a fixed location dataset fd_i, we denote it as $\langle T_i, s_i, dc\rangle$, where the additional attribute dc is the data centre where this dataset has to be stored. And we use *FD* to denote the set of the fixed location datasets a data centre has. For a data centre that does not have fixed location datasets, *FD* is empty. The distribution is conducted as the three following sub-steps.

Sub-step 1 (Line 6–16 in Fig. 4), we classify fixed location datasets and flexible location datasets in different partitions. We also need to recursively partition the sub-matrices CM_T and CM_B. The stop condition is that the sub-matrix does not have fixed location datasets or all the fixed location datasets it has belong to one data centre. We add the partitions that do not have fixed location datasets to a set named *NFP* and the partitions have fixed location datasets to a set named *FP*.

Sub-step 2 (Line 17–34 in Fig. 4), we distribute the partitions with fixed location datasets in *FP*. We need to check the data centres' information. For the data centres that have fixed location

```
Build-time Stage Algorithm

Input:   D: set of existing datasets $d_1, d_2, \dots d_n$
         DC: set of data centres $dc_1, dc_2, \dots dc_m$
Output: K: set of data centres with initial datasets

01. K=Ø; FP=Ø; NFP=Ø;       //Initialization. FP: set of partitions that have fixed location datasets
                            //NFP: set of partitions that have not fixed location dataset
02. For (every $dc_i$ in DC) $i\_cs_i = cs_i * \lambda_{ini}$ ;    //Calculate initial available storage of all data centres
03. DM = $dependency_{ij}$ = Count ($T_i \cap T_j$) ;    //Step 1: setup DM
04. CM = BEA (DM) ;                 //Step 1: BEA transformation
05. if (CM contains fd)             //Step 2 starts.  Check the existence of fixed location datasets
06.   Partition&Classify (CM)       //Sub-step 1: partition CM and classify the partitions in to FP and NFP
07.     if ($CM_T$ contains fd & the fd belong to different dc)
08.       Partition&Classify ($CM_T$) ;  //Recursively partition and classify $CM_T$
09.     else if ($CM_T$ contains fd)
10.       add $CM_T$ to FP ;        //$CM_T$ has fixed location datasets, add to FP
11.     else add $CM_T$ to NFP ;    //$CM_T$ has not fixed location datasets, add to NFP
12.     if ($CM_B$ contains fd & the fd belong to different dc)
13.       Partition&Classify ($CM_B$) ;  //Recursively partition and classify $CM_B$
14.     else if ($CM_B$ contains fd)
15.       add $CM_B$ to FP ;        //$CM_B$ has fixed location datasets, add to FP
16.     else add $CM_B$ to NFP ;    //$CM_B$ has not fixed location datasets, add to NFP
17.   for (every data centre $dc_i$ in DC)    //Sub-step 2: distribute the partitions with fixed location datasets
18.     if ($dc_i$ has fd)          //Choose the data centre $dc_i$ that has fixed location datasets
19.       for (every $fd_j$ in $FD_i$)   //Go through all the fixed location datasets belong to $dc_i$
20.         find $CM_j$ in FP ;     //Pick out the partitions that contain these fixed location datasets from PF
21.         add $CM_j$ to $P_i$ ;   //Setup the partitions set P for $dc_i$
22.       calculate $ps_i = \sum_{cm_j \in P_i} ds_j$ ;    //The total size of the partitions in P
23.       while ($ps_i > i\_cs_i$)  //Further partition if the size of P is too large for $dc_i$
24.         find $CM_k$ in $P_i$ , where $ds_k = \max_{cm_i \in P_i} ds_i$ ;    //Largest partition in P
25.         remove $CM_k$ from $P_i$ ;
26.         BinaryPartition ($CM_k$) ;     //Partition $CM_k$ and update the partitions sets
27.         if ($CM_{kT}$ contains fd) add $CM_{kT}$ to $P_i$ ;
28.         else add $CM_{kT}$ to NFP ;
29.         if ($CM_{kB}$ contains fd) add $CM_{kB}$ to $P_i$ ;
30.         else add $CM_{kB}$ to NFP ;
31.         calculate $ps_i = \sum_{cm_j \in P_i} ds_j$ ;    //New size of P after partition
32.       distribute all $CM_j$ in $P_i$ to $dc_i$ ;    //Distribute datasets
33.       update $dc_i$ to K ;
34.       $i\_cs_i = i\_cs_i - ps_i$ ;
35. else add CM to NFP ;            //CM do not contain fixed location datasets
36. for (all the partitions $CM_i$ in NFP)  //Sub-step 3: distribute the partitions without fixed location datasets
37.   Partition&Distribute ($CM_i$)        //Partition and distribute $CM_i$
38.     if ($ds_T < \max_{j=1}^{m} cs_j$)  //Size of $CM_{iT}$ is small enough for some data centres
39.       find $dc_j$ from DC,             //Find the best data centre
40.         where $cs_i = \min_{j=1}^{m}(cs_j > ds_T)$ ;
41.       distribute $CM_{iT}$ to $dc_j$ ;   //Distribute datasets
42.       update $dc_j$ to K ;
43.       $i\_cs_j = i\_cs_j - ds_{iT}$ ;
44.     else Partition&Distribute ($CM_{iT}$) ;   //Recursively partition and distribute $CM_{iT}$
45.     if ($ds_B < \max_{j=1}^{m} cs_j$)  //Size of $CM_{iB}$ is small enough for some data centres
46.       find $dc_j$ from DC,             //Find the best data centre
47.         where $cs_i = \min_{j=1}^{m}(cs_j > ds_B)$;
48.       distribute $CM_{iB}$ to $dc_j$ ;   //Distribute datasets
49.       update $dc_j$ to K ;
50.       $i\_cs_j = i\_cs_j - ds_{iB}$ ;
51.     else Partition&Distribute ($CM_{iB}$) ;   //Recursively partition and distribute $CM_{iB}$
52. Return K ;
```

Fig. 4. Build-time stage algorithm.

datasets, we pick out the partitions that contain these fixed location datasets from *FP*, denote as *P*. Then, we calculate the total size of these partitions, denote as *ps*, where $ps = \sum_{CM_i \in P} ds_i$. If these partitions can fit into this data centre, we store them. If not, we recursively pick the largest partition from *P*, binary partition it and move the part that does not have fixed location datasets to NFP, until these partitions can fit into the data centre.

Sub-step 3 (Line 36–51 in Fig. 4), we distribute the partitions that only contain flexible location datasets in *NFP*. We start with the largest one and go through all the partitions in *NFP* by their size.

For every partition, we distribute it to the data centres by recursive binary partitioning.

5.2. Runtime stage algorithm

At the runtime stage, we use the *k*-means clustering algorithm to dynamically cluster the generated data to one of the *k* data centres based on their dependencies. And when new workflows are deployed to the system or some data centres become overloaded, we also have to adjust the data placement among data centres. The pseudocode of the runtime stage algorithm is shown in Fig. 5.

For the generated data, some of them could be valuable resources that can be utilised by other workflows, but most of them are temporal data. They are generated by the preceding tasks in the workflows and will be used by the subsequent tasks. They do not need permanent storage and will be deleted after the workflows have finished their execution. In many scientific applications, the temporal data are in large volumes [47]. Some researches demonstrated that the timely removal of these temporal data can save a lot of runtime storage space [35]. In our work, we dynamically check and delete the obsolete temporal data before every round of task scheduling. The runtime stage algorithm contains the following two steps.

Step 1: Data pre-allocation by the clustering algorithm.

In this step, the first thing we have to do is task scheduling (Line 2–3 in Fig. 5). Scheduling is a very important issue in scientific workflow systems, especially for computation intensive and/or data intensive applications. Much research has been done into scheduling workflows [48,49]. However, task scheduling is not the main focus of this paper. Therefore, our scheduling strategy is quite straightforward. We just follow the philosophy of "moving data to a data centre will cost more than scheduling tasks to that centre", and schedule tasks based on the placement of datasets. We periodically monitor the state of all the workflow's tasks and dynamically schedule the ready tasks to the data centre which has the most datasets they require. Here, a task is ready if all the datasets it needs are existing data (i.e. have been generated).

When tasks have been executed, new datasets will be generated. The system will then decide where to put these datasets: either store them locally or allocate them to other data centres. In our work, the system will cluster the newly generated datasets to the data centre that has the highest dependency with them (Line 4–12 in Fig. 5). We define the dependency between dataset d_i and data centre dc_j as dc_dep_{ij}, which is the sum of the dependencies of d_i with all the datasets in dc_j.

Suppose d_u is a new generated dataset and T_u is the set of tasks that will use d_u. First, we calculate the dependencies of d_u with all other datasets in the system and add the new row and column to DM for d_u, where

$$DM_{ui} = DM_{iu} = dependency_{ui} = Count\,\{T_u \cap T_i\} \quad i = 1, 2, \ldots, n.$$

Then we calculate the dependencies of d_u with all the k data centres, where

$$dc\_dep_{uj} = \sum_{d_m \in dc_j} dependency_{um}, \quad j = 1, 2, \ldots, k.$$

With these dependencies, we will select the data centre dc_h that has the highest dependency with d_u, where

$$dc\_dep_{uh} = \max_{j=1}^{k}(dc\_dep_{uj})$$

dc_h is the data centre in which we will store the dataset d_u. And we will check the available storage of dc_h, before we move d_u to it.

Here we will introduce a maximum storage usage parameter $\lambda_{\max}$ for the data centres, which is a percentage threshold indicating whether a data centre is overloaded or not. $\lambda_{\max}$ is also an experience parameter, just like the initial storage usage parameter λ_{ini}.

Run-time Stage Algorithm

```
Input:   T: set of tasks;
         DC: set of data centres;
Output:  All the tasks are finished;

01. for (every ready task $t_i \in T$)        //Step 1: data pre-stage by clustering algorithm
02.    schedule $t_i$ to $dc_j$ to execute,        //Tasks scheduling
03.       where $dc_j$ has the most data sets $t_i$ needs;
04.    for (every generated data set $d_u$ of $t_i$)     //New datasets generated
05.       add $d_u$ to DM,                    //Update dependency matrix
06.          where $dependency_{ui} = Count\{T_u \cap T_i\}, i = 1,2,\ldots n$ ;
07.       for (every data centre $dc_k \in DC$)          //Calculate dependency
08.          calculate $dc\_dep_{uk} = \sum_{d_m \in dc_k} dependency_{um}$ ;
09.       choose $dc_h$ ,                      //Choose data centre and pre-stage data
10.          where $dc\_dep_{uh} = \max_{j=1}^{k}(dc\_dep_{uj})$ ;
11.       if ($cs_h\lambda + s_u < cs_h\lambda_{\max}$)
12.          move $d_u$ to $dc_h$ ;
13.       else                     //Step 2: adjust data placement
14.          do adjustment
15.             CM'= BEA (DM);
16.             K'= Build-time Algorithm Step 2 ;   //Get new data placement
17.             for (all the dc in DC)
18.                choose $dc_i$,            //Choose the most overloaded data centre
19.                   where $\lambda_i = \max_{dc_j \in DC} \lambda_j$;
20.                Compare&Adjust ($dc_i$, $dc_i'$)      //$dc_i'$ is the reference in K'
21.                   for (all $d_j \in dc_i \wedge d_j \notin dc_i'$ )
22.                      send $d_j$ ;         //Send the datasets belong to other data centres
23.                   for (all $d_j \notin dc_i \wedge d_j \in dc_i'$ )
24.                      retrieve $d_j$ ;     //Retrieve the datasets it should have
25.    update T for new ready tasks with $d_u$ ;      //For next round scheduling
```

Fig. 5. Runtime stage algorithm.

Hence, the storage that the runtime data can use of a data centre dc_i is $cs_i * (\lambda_{\max} - \lambda_{ini})$. The value of $\lambda_{\max}$ depends on the overall workload of the system. If the system workload is heavy, $\lambda_{\max}$ has to be set to a larger value. Likewise, if the system workload is light, $\lambda_{\max}$ is set smaller to prevent too many datasets gathering in one data centre.

We will move the new generated dataset d_u to the selected data centre dc_h, if $cs_h\lambda + s_u < cs_h\lambda_{\max}$ is true, where s_u is the size of d_u and λ is the current storage usage percentage of dc_h. Otherwise, we go to the next step to adjust the data placement.

Step 2: Adjust data placement among data centres.

During the workflow's execution, there are two situations that trigger the need to adjust the data placement among the data centres.

The first is when the selected destination data centre dc_h for the new generated dataset does not have enough available storage. This means that dc_h is overloaded. Hence, we have to adjust the datasets placement to balance the overall workload of the system.

The second is when new workflows are deployed to the system. Together with the new workflows, new datasets and tasks will be added to the system. The dependencies of the original datasets will change, since the new tasks might use the existing data in the system. In this situation, we will calculate the dependencies between the new datasets and the existing datasets, and add them to the dependency matrix DM. If there are any new tasks which use the existing data, they will be added to the task set of the appropriate existing dataset. For every new dataset, we will find an appropriate data centre for it by following the procedure in step 1. If the selected data centre is overloaded, we have to adjust the datasets placement to balance the overall workload of the system.

To adjust the data placement, we need to run some functions from the build-time stage algorithms (Line 15–16 in Fig. 5). First, we do the BEA transformation to cluster the updated dependency matrix (DM) and get a new clustered dependency matrix (CM'). Next, we run the algorithm in step 2 of the build-time stage, but without the actual data distribution. We just calculate the new

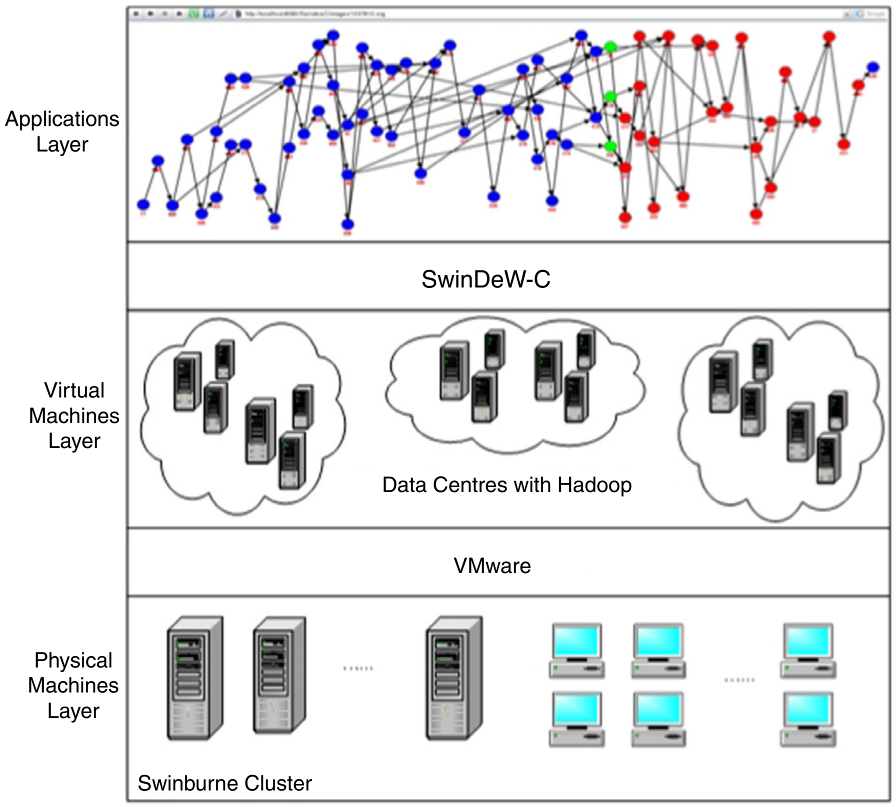


Fig. 6. Simulation environment of SwinDeW-C.

placement of datasets in the data centres and save the references in a new set of data centres, denoted as K'.

Then we can do the adjustment by comparing the old data placement with the new one in K' (Line 17–24 in Fig. 5). We start the adjustment from the data centre that has the highest storage load and go through all the data centres by the storage usage in decreasing order. For every data centre, we compare the datasets it currently has with the new datasets in K'. Then we send the datasets that do not belong to this data centre to the ones they now belong to and retrieve the datasets it should have from the other data centres.

Since $\lambda_{\max}$ represents a percentage of a data centre's total storage space, each data centre will still have some storage available $(100\% - \lambda_{\max})$ to facilitate the data movement during this redistribution. In the case that $\lambda_{\max}$ is set to 100%, additional temporary storage space may need to be acquired to serve as a buffer before the adjustment process can be completed. However, this situation rarely happens in the system, due to the following reasons: (1) in the adjustment process we always select the data centre with the highest storage usage to adjust as the priority, and send its datasets to other data centres first; (2) the total size of the datasets in the system is smaller than the total size of the available storage of all the data centres $(\sum_{i=1}^{n} s_i < \sum_{i=1}^{m} (cs_i * \lambda_{ini}))$, because we have the assumption that the data centres can host all the application data in the system; and (3) for every data centre we reserve some storage for the runtime generated datasets $(cs * (\lambda_{\max} - \lambda_{ini}))$, this storage space is not always highly utilised, because we delete obsolete datasets dynamically. In our system, for every data centre, we reserve the runtime storage for generated datasets as 40% of the initial storage for existing datasets i.e. $(\lambda_{\max} - \lambda_{ini}) / \lambda_{ini} = 40\%$. As addressed in Section 6 later, we have run tens of thousands of workflow instances for a simulation, and a situation where we lacked storage for data reallocation did not occur.

The data placement strategy in this section states that when a task is scheduled to one data centre during a workflow's execution, that data centre will have most input datasets for that task. Then, only a small number of datasets have to be retrieved from the remote data centres. The simulations in the next section will show that our data placement strategy can greatly reduce the total data movement during a workflow's execution.

6. Simulation

6.1. Simulation environment: SwinDeW-C

SwinDeW-C (Swinburne Decentralised Workflow for Cloud) [49] is developed based on SwinDeW [50] and SwinDeW-G [51]. It is currently running at the Swinburne University of Technology, which is composed of 10 servers and 10 high-end PCs. To simulate the cloud computing environment, we set up VMware [52] software on the physical servers and create virtual clusters as data centres. Fig. 6 shows our simulation environment.

Every data centre created is composed of 8 virtual computing nodes with storage, and we deploy an independent Hadoop file system on each data centre. SwinDeW-C runs on these virtual data centres that can send and retrieve data to and from each other. Through a user interface at the applications layer, which is a Web based portal, we can deploy workflows and upload application data.

SwinDeW-C is designed for large scale cloud applications. It has a novel architecture for the cloud computing environment. However, the presentation of the comprehensive system design of SwinDeW-C is not the main focus of this paper. In Fig. 7, we only illustrate the key system components of SwinDeW-C that relate to the data placement strategy.

User interface module: The cloud computing platform is built on the Internet and a Web browser is normally the only software

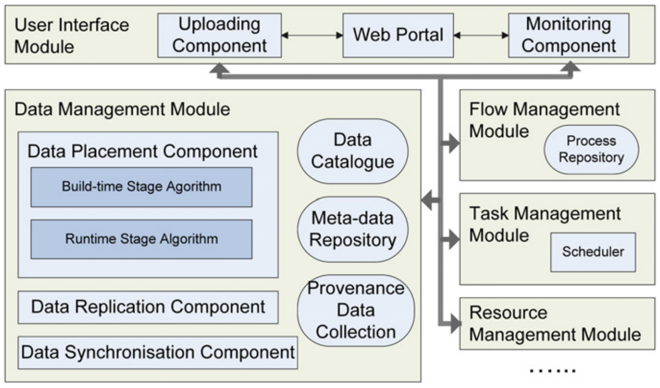


Fig. 7. Related key system components of SwinDeW-C.

needed at the client's side. This interface is a Web portal by which users can visit the system and deploy their applications. The *Uploading Component* is for users to upload application data and workflows, and the *Monitoring Component* is for users, as well as system administrators to monitor the workflow's execution.

Data management module: The *Data Placement Component* is the core component of data management in SwinDeW-C that facilitates the algorithms in our data placement strategy. The *Data Catalogue* is used to store the information of applications which, in a service oriented cloud platform, is a registry for the data services. By using the catalogue, the system can locate the data needed. Other components in this module, such as the *Data Replication Component, Data Synchronisation Component, Meta-data Repository*and *Provenance Data Collection* are also essential for cloud data management. Since they are not directly related to the data placement strategy, we do not give their details here.

Other modules: The *Flow Management Module* has a *Process Repository* that stores all the workflow instances running in the system. The *Task Management Module* has a *Scheduler* that schedules ready tasks to data centres during the runtime stage of the workflows. Furthermore, the *Resource Management Module* keeps the information of the data centres' usage, and can trigger the adjustment process in the data placement strategy. For other components in these modules, as well as other modules in SwinDeW-C, we do not give the details as the work presented here only focuses on the workflow data management.

6.2. Simulation strategies

The algorithms in our data placement strategy are for the build-time and runtime stages respectively. To evaluate their performance, we run each workflow instance through 4 simulation strategies:

Random: In this simulation, we randomly place the *existing data* during the build-time stage and store the *generated data* in the local data centre (i.e. where they were generated) at the runtime. This simulation represents the traditional data placement strategies in old distributed computing systems (i.e. clusters and early grid systems). At that time, data were usually stored in the local node naturally or in the nodes that had the available storage. The temporal intermediate data, i.e. generated data, were also naturally stored where they were generated waiting for the tasks to retrieve them.

Build-time only: This simulation shows the performance of our build-time algorithm. It is used to place the existing data at the build-time. During the runtime stage we will store the generated data in the local data centre, as with the Random simulation. In a cloud computing system, data are more flexible than they were in the past; this allows the system to decide where to store them. Our build-time algorithm places the application data based on their dependencies. This simulation will show the data movement reduction in the workflows' execution by using this algorithm.

Runtime only: This simulation shows the performance of the runtime algorithm by randomly placing the existing data at build-time and by pre-allocating the generated data with our runtime algorithm. This simulation represents the strategy that some popular grid scientific workflows used [34]. Their work shows that pre-allocating data to the computing node where the tasks will execute can reduce the total execution time of the workflow. However, this simulation will show that only pre-allocating data at runtime stage can not reduce the data movement in a workflow's execution.

Build & run: This simulation shows the overall performance of our algorithms both at the build-time and the runtime. Our algorithms are specifically designed for scientific cloud workflows. The strategy is based on data dependency and can automatically place existing data; and cluster generated data to the appropriate data centres. Comparisons with other strategies will be made with different aspects to show the performance of our algorithms.

The traditional way to evaluate the performance of a workflow system is to record and compare the execution time [34,35]. However, in our work we will count the total data movement instead. The execution time could be influenced by other factors beside data management, such as bandwidth, scheduling strategy and I/O speed. Our data placement strategy aims to reduce the data movement between data centres on the Internet. So we directly take the number of datasets that are actually moved during the workflow's execution as the measurement to evaluate the performance of the algorithms. In a cloud computing environment with a limited bandwidth based on the Internet, if the total data movement has been reduced, the execution time will be reduced correspondingly. Furthermore, the cost of data transfer will also decrease.

To make the evaluation as objective as possible, we generate test workflows randomly to run on SwinDeW-C. This would make the evaluation results independent of any specific applications. As we need to run the build-time and the runtime algorithms separately, we set the number of existing datasets and generated datasets to be the same for every test workflow. That means that we have the same number of existing datasets and tasks for every test workflow, and we assume that each task will only generate one dataset. We can control the complexity of the test workflow by changing the number of datasets. Every dataset will be used by a random number of tasks, and tasks that use generated datasets must be executed after the task that generates their input. We can control the complexity of the relationships between the datasets and tasks by changing the range of this random number. Another factor that would have an impact on the algorithms is the number of fixed location datasets. We can randomly choose some percentage of datasets from the existing data and randomly select some data centres for them. We will run new simulations to show their impact on the performance. Here we have only included graphs of the simulation results. The detailed configuration and result reports of the simulations, as well as the source code can all be found at http://www.swinflow.org/docs/DataPlacement.zip.

6.3. Simulation results

Fig. 8 shows the data movement when we run workflows with different complexities on different numbers of data centres. We can see the increases in data movement as the workflows become more complex and the number of data centres increases. All the values in the figure are the average of running 1000 test workflows with the same parameters.

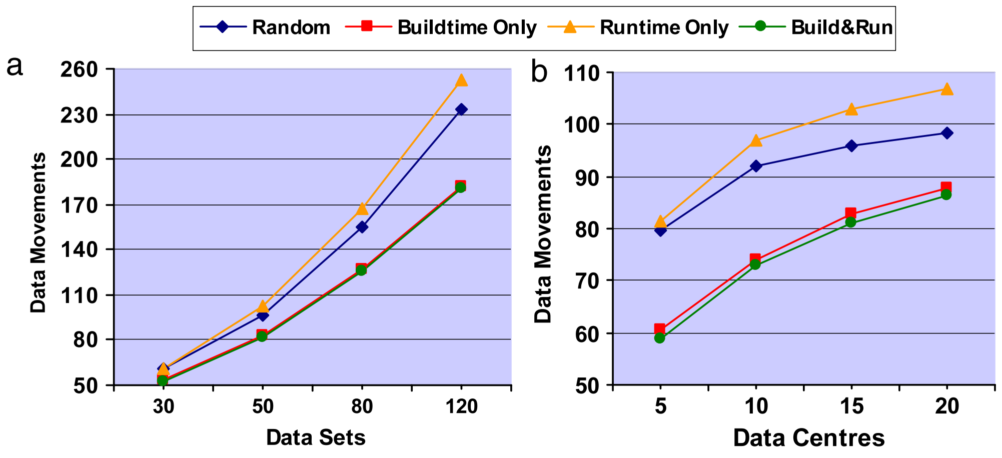


Fig. 8. Data movements without runtime storage limit and without fixed location datasets.

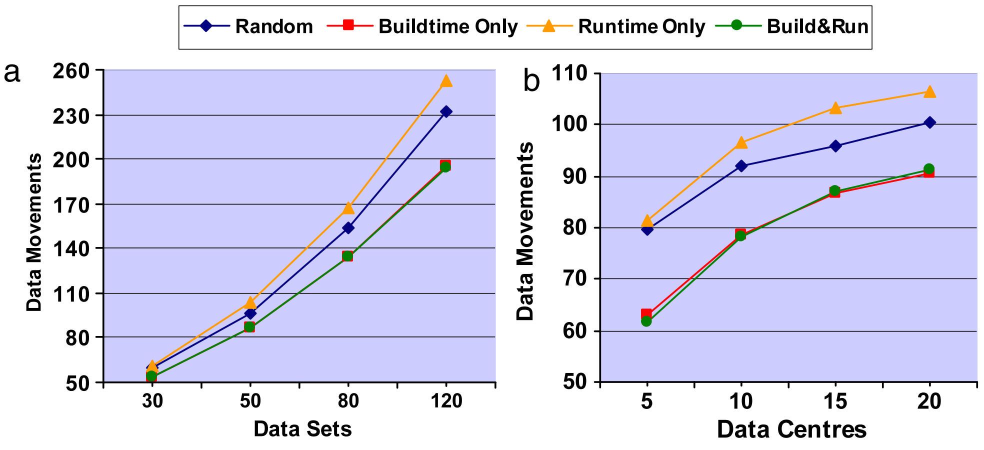


Fig. 9. Data movements without runtime storage limit and with 10% of fixed location datasets.

In Fig. 8(a), we ran the test workflows with different complexities on 15 data centres. We used 4 types of test workflows with different numbers of datasets. In Fig. 8(b), we fixed the test workflows' datasets count to 50, and ran them on different numbers of data centres. Then we changed 10% of the input datasets to fixed location datasets and ran the same simulation again. The results are shown in Fig. 9.

From the results, we could draw the conclusions that (1) the build-time algorithm can effectively reduce the total data movement of the workflow's execution; (2) the runtime algorithm does not reduce the total data movement, and even causes more data movement if the existing datasets are placed randomly and (3) with fixed location datasets added to the system, our algorithms can still work very well with the performance only degrading slightly. The runtime algorithm does not decrease the data movement because it pre-allocates datasets before scheduling tasks based on their data dependencies. If the existing datasets are randomly placed, the differing dependencies of the data centres are not obvious. The increase in data movement is caused by a pre-allocation of the datasets to the wrong data centres. However, if the existing datasets were clustered by the build-time algorithm, the performance of the runtime algorithm would be better.

However, in the simulation described above, we did not limit the amount of storage that the data centres had available during runtime. The reason for this is that we wanted to see how the tasks and datasets were distributed, which indicates the workload balance among data centres. During the execution of every test workflow instance, we recorded the number of datasets that moved to each data centre, as well as the tasks that scheduled to that data centre. We also calculated the standard deviation of the data centres' usage. Fig. 10 shows the average standard deviation of running 1000 test workflows on 15 data centres each having 80 existing datasets and 80 tasks, both with and without fixed location datasets.

From Fig. 10 we can see relatively high deviations in the data centres' usage in the two simulations without the runtime algorithm. This means that tasks and datasets are allocated to one data centre more frequently. This leads to a data centre becoming a super node that has a high workload. By contrast, in the other two simulations that use the runtime algorithm to pre-allocate the generated data to other data centres, the deviation of the data centre usage is low. This demonstrates that the runtime algorithm can make a more balanced distribution of the workload among data centres.

In a cloud computing environment, data centres normally have a limited storage, especially in some storage constrained systems. When one data centre is overloaded, we need to reallocate the data to other data centres. The reallocation will not only cause extra data movement, but will also delay the execution of the workflow. To count the reallocated datasets, we ran the same test workflows as in Fig. 10 with a storage limit in every data centre. We limited the runtime storage for generated datasets to 40% of the initial storage for existing datasets i.e. $(\lambda_{\max} - \lambda_{ini}) / \lambda_{ini} = 40\%$. In Fig. 11 we show the average data movement including the data reallocation.

From Fig. 11, we can see that a lot of data is reallocated in the simulations without the runtime algorithm. The least data reallocation occurred when we only use the runtime algorithm. However, the least data movement in total occurred when using the

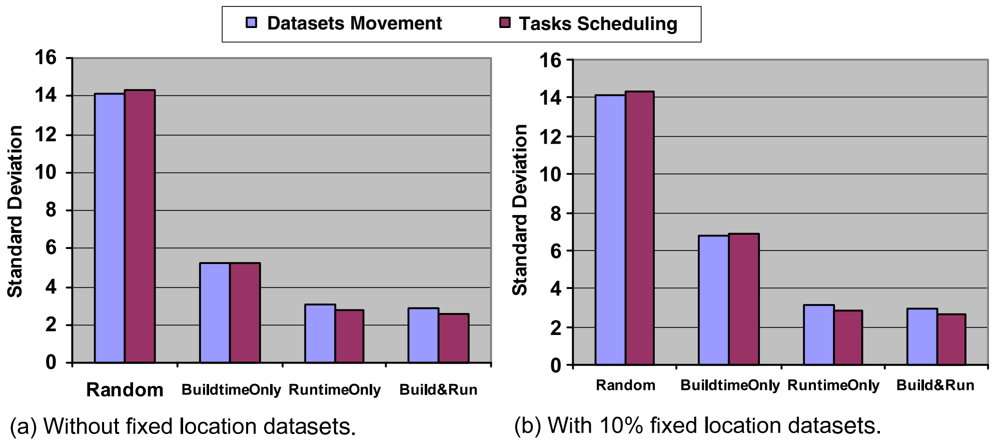


Fig. 10. Standard deviation of workload among data centres.

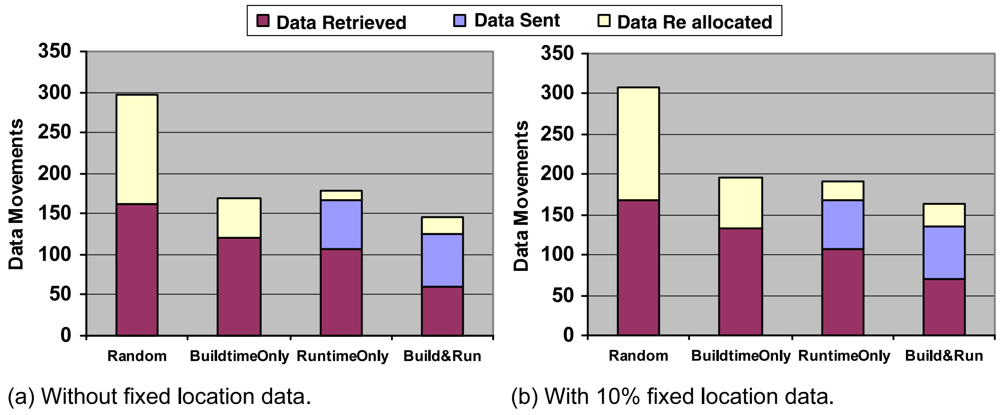


Fig. 11. Proportions of 3 types of data movements.

build-time and runtime algorithms together. In Fig. 11(a), using both algorithms caused 146.505 movements of datasets on average. Comparing this to the random simulation, 297.807 datasets movements on average, our algorithms reduce the data movement by 50.8%. On the other hand, the build-time algorithm and runtime algorithm cause movements of 170.26 and 178.662 datasets on average. Compared to the random situation, they reduce the data movements by 42.8% and 40.0% respectively. In Fig. 11(b), with 10% fixed location datasets in the system, our algorithms (Build&Run) can reduce the data movement by 47.4% compared to the Random simulation.

To better evaluate the performance of our algorithms, we give every data centre a runtime storage limit and run the same simulation workflows as Fig. 8. We get the final results of data movement which are shown in Fig. 12.

From Fig. 12 we can see that as the number of data centres and datasets increases, the performance of the build-time algorithm decreases. This is because without the runtime algorithm the datasets and tasks are gathering on the one data centre. This triggers the adjustment process more frequently, which costs extra data movements. Furthermore, we ran the same simulation as Fig. 12 under the condition that the system has fixed location datasets. Fig. 13 shows the data movements when we set the percentage of fixed location datasets to 10%. We can see our algorithms can still reduce the data movements significantly. Furthermore, with higher percentages of fixed location datasets in the system, our algorithms still work, and we will demonstrate this in the next simulation.

Fig. 13 has consistent results with Fig. 9, in that the fixed location datasets have a negative impact on the algorithms' performance. In the algorithms, we try to place the datasets on data centres based on dependencies, however, the fixed location datasets have to be stored in particular data centres. This will decrease the performance, as fixed location datasets will prevent the algorithms from placing datasets with their dependencies. However, given the existence of fixed location datasets, our algorithms can still reduce the data movement by placing the flexible location datasets with dependencies. To demonstrate the impact of fixed location datasets on the algorithms, we conducted another batch of simulations. We ran 1000 test workflows on 15 data centres each having 80 existing datasets and 80 tasks, but with different percentages of fixed location datasets. As the number of fixed location datasets increases, we can see their impact on the data movement in Fig. 14.

From Fig. 14(a) we can see that as the percentage of fixed location datasets goes up, the data movements of the Build-time only and Build & Run simulations go up accordingly; however the Random and Runtime only simulations keep steady. This means the fixed location datasets primarily have an impact on the build-time algorithm. This is because all the fixed location datasets are existing data, which are placed by the build-time stage algorithm. When the percentage reaches 60%, the data movements of Build & Run simulation even exceeds the Random simulation. This is because the pre-allocation of datasets in the runtime algorithm causes more data movements, as the build-time algorithm gets worse. In Fig. 14(b) it may seem slightly confusing that the data movements of all simulations go up and then drop, as the percentage of fixed location datasets goes up. This is because when we set the runtime storage limit, many data movements are caused by data reallocation. However, the fixed location datasets

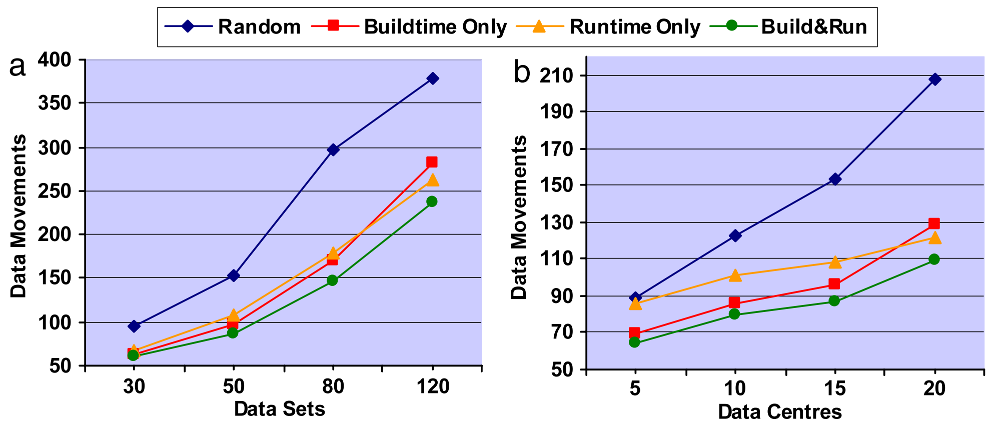


Fig. 12. Data movements with runtime storage limit.

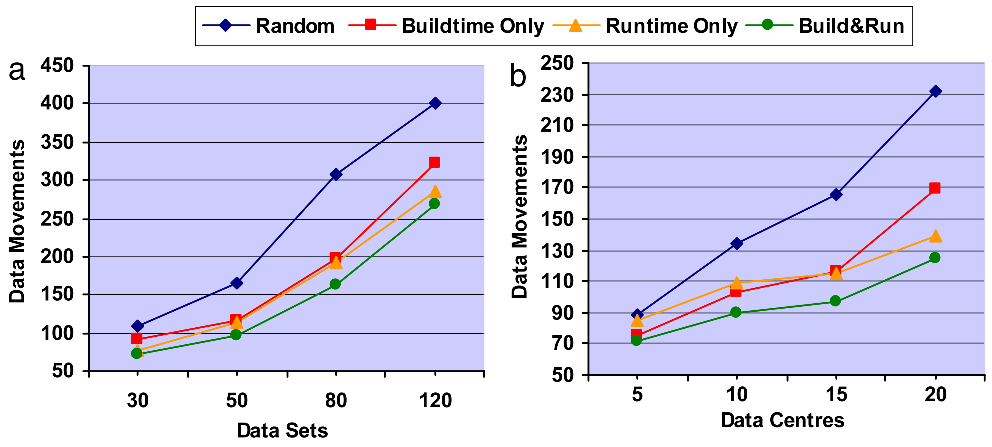


Fig. 13. Data movements with runtime storage limit and with 10% fixed location datasets.

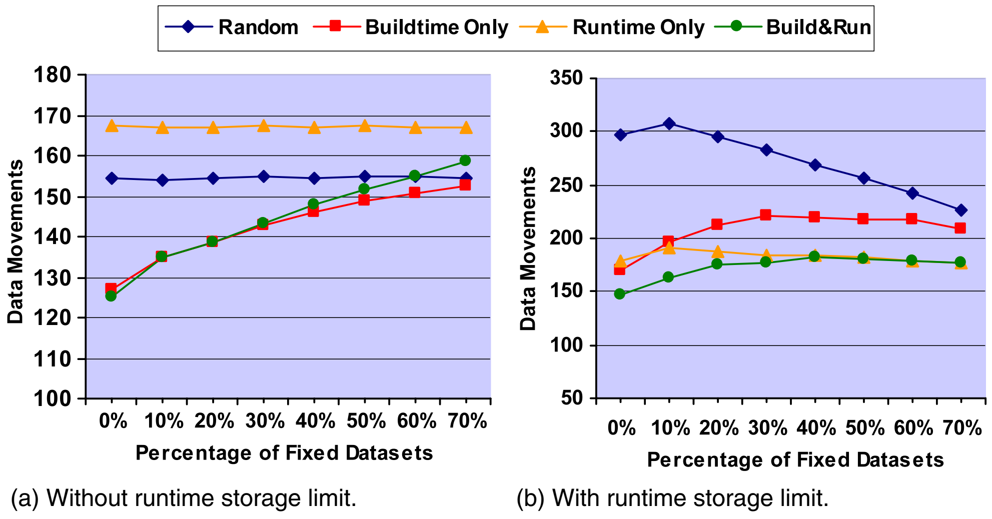


Fig. 14. Data movements with different percentage of fixed location datasets.

are not involved in the overload adjustment process. Hence, the data movement decreases. In this figure we can also see that the fixed location datasets may have a negative impact on the build-time algorithm.

7. Conclusions and future work

In this paper, we examined the unique features of scientific cloud workflows and proposed a clustering data placement strategy that can automatically allocate application data among data centres based on the dependencies. Simulations in our cloud workflow system SwinDeW-C indicated that our data placement strategy can effectively reduce the data movement during the workflow's execution. The build-time algorithm reduces the amount of data retrieved and the run time algorithm guarantees a balanced distribution of data and can reduce the data movement incurred by data reallocation, even when fixed location data exist in the system.

In our current work, to guarantee the data reliability, we used Hadoop's replication mechanism within a data centre, and among data centres we did not use any replication strategies. The data used in scientific workflow applications are usually very large and as such it is not efficient to replicate all the application data in the system. However, replication of frequently used data could also reduce the data movement. In the future work, we will develop some efficient replication strategies for the data placement algorithm, which could balance the data movement and storage usage. Furthermore, in our current simulation we measure the reduction of the datasets' movements to evaluate our strategy. In the future, we will meter the execution time of the workflow as well, which can better demonstrate the effectiveness of our strategy. To be more comprehensive, we will also incorporate the size of datasets to calculate the data dependency, and adapt some popular cloud service providers' pricing models to our simulation, which will show the cost effectiveness of our strategy.

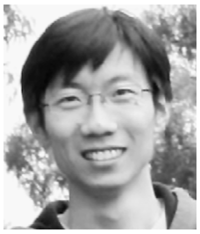
Acknowledgements

The research work reported in this paper is partly supported by Australian Research Council under Linkage Project LP0990393. We are grateful to Bryce Gibson and Michael Jensen for the accomplishment of the simulation work, as well as the careful English proofreading.

References

[1] E. Deelman, A. Chervenak, Data management challenges of data-intensive scientific workflows, in: IEEE International Symposium on Cluster Computing and the Grid, 2008, pp. 687–692.
[2] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, S. Patil, M.-H. Su, K. Vahi, M. Livny, Pegasus: Mapping scientific workflows onto the grid, in: European Across Grids Conference, 2004, pp. 11–20.
[3] B. Ludascher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E.A. Lee, Scientific workflow management and the Kepler system, Concurrency and Computation: Practice and Experience (2005) 1039–1065.
[4] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M.R. Pocock, A. Wipat, P. Li, Taverna: A tool for the composition and enactment of bioinformatics workflows, Bioinformatics 20 (2004) 3045–3054.
[5] A. Weiss, Computing in the Cloud, vol. 11, ACM Networker, 2007, 18–25.
[6] M. Brantner, D. Florescuy, D. Graf, D. Kossmann, T. Kraska, Building a Database on S3, in: SIGMOD, Vancouver, BC, Canada, 2008, pp. 251–263.
[7] R. Grossman, Y. Gu, Data Mining Using High Performance Data Clouds: Experimental Studies Using Sector and Sphere, in: SIGKDD, 2008, pp. 920–927.
[8] R. Buyya, C.S. Yeo, S. Venugopal, Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities, in: 10th IEEE International Conference on High Performance Computing and Communications, HPCC-08, Los Alamitos, CA, USA, 2008.
[9] C. Moretti, J. Bulosan, D. Thain, P.J. Flynn, All-Pairs: An abstraction for data-intensive cloud computing, in: IEEE International Parallel & Distributed Processing Symposium, IPDPS'08, 2008, pp. 1–11.
[10] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems 25 (2009) 599–616.
[11] I. Foster, Z. Yong, I. Raicu, S. Lu, Cloud computing and grid computing 360-degree compared, in: Grid Computing Environments Workshop, GCE'08, 2008, pp. 1–10.
[12] K. Keahey, R. Figueiredo, J. Fortes, T. Freeman, M. Tsugawa, Science clouds: Early experiences in cloud computing for scientific applications, in: First Workshop on Cloud Computing and its Applications, CCA'08, 2008, pp. 1–6.
[13] L. Wang, J. Tao, M. Kunze, A.C. Castellanos, D. Kramer, W. Karl, Scientific cloud computing: Early definition and experience, in: 10th IEEE International Conference on High Performance Computing and Communications, HPCC'08, 2008, pp. 825–830.
[14] E. Deelman, G. Singh, M. Livny, B. Berriman, J. Good, The cost of doing science on the cloud: The montage example, in: ACM/IEEE Conference on Supercomputing, Austin, Texas, 2008, pp. 1–12.
[15] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, J. Good, On the use of cloud computing for scientific workflows, in: 4th IEEE International Conference on e-Science, 2008, pp. 640–645.
[16] R. Barga, D. Gannon, Scientific versus business workflows, in: I.J. Taylor, E. Deelman, D.B. Gannon, M. Shields (Eds.), Workflows for e-Science, Springer, London, UK, 2007, pp. 9–16.
[17] W.T. McCormick, P.J. Sehweitzer, T.W. White, Problem decomposition and data reorganization by a clustering technique, Operations Research 20 (1972) 993–1009.
[18] T. Xie, SEA: A striping-based energy-aware strategy for data placement in RAID-structured storage systems, IEEE Transactions on Computers 57 (2008) 748–761.
[19] T. Kosar, M. Livny, Stork: Making data placement a first class citizen in the grid, in: Proceedings of 24th International Conference on Distributed Computing Systems, ICDCS 2004, 2004, pp. 342–349.
[20] J.M. Cope, N. Trebon, H.M. Tufo, P. Beckman, Robust data placement in urgent computing environments, in: IEEE International Symposium on Parallel & Distributed Processing, IPDPS 2009, 2009, pp. 1–13.
[21] N. Hardavellas, M. Ferdman, B. Falsafi, A. Ailamaki, Reactive NUCA: Near-optimal block placement and replication in distributed caches, in: Proceedings of the 36th Annual International Symposium on Computer Architecture, ISCA '09, Austin, TX, USA, 2009, PP. 184–195.
[22] S. Ghemawat, H. Gobioff, S.-T. Leung, The Google File System, vol. 37, SIGOPS Oper. Syst. Rev., 2003, 29–43.
[23] Hadoop, http://hadoop.apache.org/ (accessed 25.11.09).
[24] R. Prodan, T. Fahringer, Overhead analysis of scientific workflows in grid environments, IEEE Transactions on Parallel and Distributed Systems 19 (2008) 378–393.
[25] M. Wieczorek, R. Prodan, T. Fahringer, Scheduling of scientific workflows in the ASKALON grid environment, SIGMOD Record 34 (2005) 56–62.
[26] C. Baru, R. Moore, A. Rajasekar, M. Wan, The SDSC storage resource broker, in: IBM Centre for Advanced Studies Conference, Toronto, Canada, 1998, pp. 1–12.
[27] D. Churches, G. Gombas, A. Harrison, J. Maassen, C. Robinson, M. Shields, I. Taylor, I. Wang, Programming scientific and distributed workflow with Triana services, in: Concurrency and Computation: Practice and Experience, vol. 18, 2006, pp. 1021–1037.
[28] A. Chervenak, E. Deelman, I. Foster, L. Guy, W. Hoschek, A. Iamnitchi, C. Kesselman, P. Kunszt, M. Ripeanu, B. Schwartzkopf, H. Stockinger, K. Stockinger, B. Tierney, Giggle: A framework for constructing scalable replica location services, in: ACM/IEEE Conference on Supercomputing, Baltimore, Maryland, 2002, pp. 1–17.
[29] R. Buyya, S. Venugopal, The gridbus toolkit for service oriented grid and utility computing: An overview and status report, in: IEEE International Workshop on Grid Economics and Business Models, Seoul, 2004, pp. 19–66.
[30] S. Venugopal, R. Buyya, L. Winton, A grid service broker for scheduling distributed data-oriented applications on global grids, in: 2nd Workshop on Middleware in Grid Computing, Toronto, Canada, 2004, pp. 75–80.
[31] S. Venugopal, R. Buyya, K. Ramamohanarao, A taxonomy of data grids for distributed data sharing, management, and processing, ACM Computing Survey 38 (2006) 1–53.
[32] S. Doraimani, A. Iamnitchi, File grouping for scientific data management: Lessons from experimenting with real traces, in: Proceedings of the 17th International Symposium on High Performance Distributed Computing, ACM, Boston, MA, USA, 2008, pp. 153–164.
[33] G. Fedak, H. He, F. Cappello, BitDew: A programmable environment for large-scale data management and distribution, in: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, Austin, Texas, 2008, pp. 1–12.
[34] A. Chervenak, E. Deelman, M. Livny, M.-H. Su, R. Schuler, S. Bharathi, G. Mehta, K. Vahi, Data placement for scientific applications in distributed environments, in: 8th Grid Computing Conference, 2007, pp. 267–274.
[35] G. Singh, K. Vahi, A. Ramakrishnan, G. Mehta, E. Deelman, H. Zhao, R. Sakellariou, K. Blackburn, D. Brown, S. Fairhurst, D. Meyers, G.B. Berriman, J. Good, D.S. Katz, Optimizing workflow data footprint, Scientific Programming 15 (2007) 249–268.
[36] T. Kosar, M. Livny, A framework for reliable and efficient data placement in distributed computing systems, Journal of Parallel and Distributed Computing 65 (2005) 1146–1157.
[37] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Computing Survey 31 (1999) 264–323.
[38] S. Guha, A. Meyerson, N. Mishra, R. Motwani, L. O'Callaghan, Clustering data streams: Theory and practice, IEEE Transactions on Knowledge and Data Engineering 15 (2003) 515–528.
[39] ATNF Parkes Swinburne Recorder, http://astronomy.swin.edu.au/pulsar/?topic=apsr (accessed 25.11.09).
[40] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above the clouds: A Berkeley view of cloud computing, University of California at Berkeley, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf, Technical Report UCB/EECS-2009-28 (accessed 25.11.09).
[41] R. Grossman, Y. Gu, M. Sabala, W. Zhang, Compute and storage clouds using wide area high performance networks, Future Generation Computer Systems (2008) 179–183.
[42] Amazon elastic computing cloud, http://aws.amazon.com/ec2/ (accessed 25.11.09).
[43] H. Liu, D. Orban, GridBatch: Cloud computing for large-scale data-intensive batch applications, in: Eighth IEEE International Symposium on Cluster Computing and the Grid, 2008, pp. 295–305.
[44] J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large clusters, Communications of the ACM 51 (2008) 107–113.
[45] A. Matsunaga, M. Tsugawa, J. Fortes, CloudBLAST: Combining MapReduce and virtualization on distributed resources for bioinformatics applications, in: 4th IEEE International Conference on e-Science, 2008, pp. 222–229.
[46] M.T. Ozsu, P. Valduriez, Principles of Distributed Database Systems, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1991.

[47] E. Deelman, D. Gannon, M. Shields, I. Taylor, Workflows and e-Science: An overview of workflow system features and capabilities, Future Generation Computer Systems 25 (2009) 528–540.
[48] S. Venugopal, R. Buyya, An SCP-based heuristic approach for scheduling distributed data-intensive applications on global grids, Journal of Parallel and Distributed Computing 68 (2008) 471–481.
[49] Y. Yang, K. Liu, J. Chen, X. Liu, D. Yuan, H. Jin, An algorithm in SwinDeW-C for scheduling transaction-intensive cost-constrained cloud workflows, in: 4th IEEE International Conference on e-Science, 2008, pp. 374–375.
[50] J. Yan, Y. Yang, G.K. Raikundalia, SwinDeW — A P2P-based decentralized workflow management system, IEEE Transactions on Systems, Man and Cybernetics 36 (Part A) (2006) 922–935.
[51] Y. Yang, K. Liu, J. Chen, J. Lignier, H. Jin, Peer-to-peer based grid workflow runtime environment of SwinDeW-G, in: IEEE International Conference on e-Science and Grid Computing, 2007, pp. 51–58.
[52] VMware, http://www.vmware.com/ (accessed 25.11.09).



Dong Yuan was born in Jinan, China. He received the B.Eng. degree in 2005 and M.Eng. degree in 2008 both from Shandong University, Jinan, China, all in Computer Science.

He is currently a Ph.D. student in the Faculty of Information and Communication Technologies at Swinburne University of Technology, Melbourne, Vic., Australia. His research interests include data management in workflow systems, scheduling and resource management, grid and cloud computing.

Yun Yang was born in Shanghai, China. He received the B.S. degree from Anhui University, Hefei, China, in 1984, the M.Eng. degree from the University of Science and Technology of China, Hefei, China, in 1987, and the Ph.D. degree from the University of Queensland, Brisbane, Australia, in 1992, all in Computer Science.

He is currently a Full Professor in the Faculty of Information and Communication Technologies at Swinburne University of Technology, Melbourne, Vic., Australia. Prior to joining Swinburne as an Associate Professor, he was a Lecturer and Senior Lecturer at Deakin University during 1996–1999. Before that, he was a (Senior) Research Scientist at DSTC Cooperative Research Centre for Distributed Systems Technology during 1993–1996. He also worked at the Beijing University of Aeronautics and Astronautics during 1987–1988. He has co-edited two books and published more than 170 papers on journals and refereed conferences. His current research interests include software technologies, p2p/grid/cloud workflow systems, service-oriented computing, cloud computing, and e-learning.

Xiao Liu received his master degree in Management Science and Engineering from Hefei University of Technology, Hefei, China, 2007. He is currently a Ph.D. student in Centre for Complex Software Systems and Services in the Faculty of Information and Communication Technologies at Swinburne University of Technology, Melbourne, Australia. His research interests include workflow management systems, scientific workflow, business process management and data mining.

Jinjun Chen received his Ph.D. degree from Swinburne University of Technology, Australia. His thesis was granted Research Thesis Execellence Award. He received Swinburne Vice Chancellor's research award 2008. He is a core executive member of IEEE Technicial Committee of Scalable Computing and the coordinator of IEEE TCSC technicial area of Workflow Management in Scalable Computing Environments. He is the Editor-in-Chief of Springer book series on Advances in Business Process and Workflow Management (http://www.swinflow.org/books/springer/SpringerBook.htm) and Editor-in-Chief of Nova book series on Process and Workflow Management and Applications (http://www.swinflow.org/books/nova/NovaBook.htm). He has guest edited or is editing several special issues in quality journals such as in IEEE Transactions on Automation Science and Engineering. He has been involved in the organization of many conferences and awarded IEEE Computer Society Service Award (2007).

He has published more than 50 papers in journals and conferences such as ICSE2008 and ACM TAAS. His research interests include Scientific Workflow Management and Applications, Workflow Management and Applications in Web Service or SOC Environments, Workflow Management and Applications in Grid (Service)/Cloud Computing Environments, Software Verification and Validation in Workflow Systems, QoS and Resource Scheduling in Distributed Computing Systems such as Cloud Computing, Service Oriented Computing (SLA and Composition).