

Dynamic Data Grid Replication Strategy based on Internet Hierarchy*

Sang-Min Park¹, Jai-Hoon Kim¹, Young-Bae Ko², and Won-Sik Yoon²

¹Graduate School of Information and Communication

Ajou University, South Korea

{smpark, jaikim}@ajou.ac.kr

²College of Information Technology

Ajou University, South Korea

{youngko, wsyoon}@ajou.ac.kr

Abstract. In data grid, large quantity of data files is produced and data replication is applied to reduce data access time. Efficiently utilizing grid resources becomes important research issue since available resources in grid are limited while large number of workloads and large size of data files are produced. Dynamic replication in data grid aims to reduce data access time and to utilize network and storage resources efficiently. This paper proposes a novel dynamic replication strategy, called BHR, which reduces data access time by avoiding network congestions in a data grid network. With BHR strategy, we can take benefits from '*network-level locality*' which represents that required file is located in the site which has broad bandwidth to the site of job execution. We evaluate BHR strategy by implementing it in an OptorSim, a data grid simulator initially developed by European Data Grid Projects. The simulation results show that BHR strategy can outperform other optimization techniques in terms of data access time when hierarchy of bandwidth appears in Internet. BHR extends current site-level replica optimization study to the network-level.

1. Introduction

A grid is large scale resource sharing and problem solving mechanism in virtual organizations [6]. Large number of computational and storage resources are linked globally to form a grid. In some scientific application areas such as high energy physics, bioinformatics, and earth observations, we encounter huge amounts of data. People expect the size of data to be terabyte or even petabyte scale in some applications [7]. Managing such huge amounts of data in a centralized manner is almost impossible due to extensively increased data access time. Data replication is a key

* This work is supported by a grant of the International Mobile Telecommunications 2000 R&D Project, Ministry of Information & Communication in South Korea, ITA Professorship for Visiting Faculty Positions in Korea (International Joint Research Project) by Ministry of Information & Communication in South Korea, and grant No. (R01-2003-000-10794-0) from the Basic Research Program of the Korea Science & Engineering Foundation.

technique to manage large data in a distributed manner. By its nature, we can achieve better performance (access time) by replicating data in geographically distributed data stores. In data grid, user's jobs require access to large number of files. If the required files are replicated in some site in which the job is executed, job is able to process data without any communication delay. However, if required files are not in the site, they should be fetched from other sites and it usually takes very long time because size of single replica may reach giga-byte scale in some applications and network bandwidth between sites is limited. As a result, job execution time becomes very long due to delay of fetching replicas over Internet. Dynamic replication is an optimization technique which aims to maximize chances of data locality. In other words, dynamic replica optimizer running in a site tries to locate files which are likely to be requested in the near future. As the number of file hit ratio increases, job execution time reduces significantly. Various dynamic replication strategies have been introduced so far [2, 3, 11].

In this paper, we propose novel dynamic replication strategy; called BHR (Bandwidth Hierarchy based Replication). The existing replication strategies try to maximize locality of file to reduce data access time. However, grid sites may be able to hold only small portion of overall amount of data since very large quantity of data is produced in data grid and the storage space in a site is limited. Therefore, effect from this locality is limited to a certain degree. BHR strategy takes benefit from other form of locality, called *network-level locality*. Although the required file is not in the same site performing job, there will be not long delay of fetching replica if the replica is located in the site having broad bandwidth to the site of job execution. We call this condition as *network-level locality*. In data grid, some sites may be located within a region where sites are linked closely. For instance, a country can be referred to as this network region. Network bandwidth between sites within a region will be broader than bandwidth between sites across regions. Thus, hierarchy of network bandwidth may appear in Internet. If the required file is located in the same region, less time will be consumed to fetch the file. In other words, benefit of *network-level locality* can be exploited. BHR strategy reduces data access time by maximizing this *network-level locality*.

2. Related Works

Dynamic replication is a long-term optimization technique which aims at reducing average job execution time in data grid. Since very large quantity of data files are deployed in data grid, there will be certain limitation of amount of files which can be stored at each site. If SE (Storage Element) at a grid site is already filled up with replicas, some of them should be deleted in order to store newly requested data.

Kavitha Ranganathan et al. present various traditional replication and caching strategies and evaluate them from the perspective of data grid in [11]. They measure access latency and bandwidth consumptions of each strategy with simulation tool and their simulation results show that Cascading and Fast Spread perform best among traditional strategies.

Economy based replication is proposed in [2, 3]. In economic approach, a kind of auction protocol is used to select the best replica for a job and to trigger long-term optimization (dynamic optimization) by using file access patterns. The authors show the improvement compared to traditional replication techniques by performing simulation with OptorSim. OptorSim is a data grid simulation tool developed as part of European Data Grid Project [8]. General data grid scenarios are modeled in OptorSim and one can evaluate various replication strategies implemented in it.

The existing replication techniques mentioned above are based on file access pattern at each site. If a grid site requests some files more frequently than others, it is better for the site to hold these files for near future usage. Even though this site-level locality can reduce data access time to some extent, there remains limitation. Performance gain from site-level locality can make sense when grid sites have enough space to store large portion of data and certain predictable file access patterns come out. However, we cannot assure in many cases that a single grid site will have enough space to store large portion of whole data and there will be predictable file access patterns. We find another key to the performance improvement by broadening our view of locality to the network level.

3. Dynamic Replication Strategy based on Bandwidth Hierarchy

In this section, we propose novel dynamic replication strategy, called BHR, which is based on bandwidth hierarchy of Internet. Our BHR strategy takes benefit from network-level locality of files. The idea of proposed strategy is motivated from the assumption that hierarchy of bandwidth appears in Internet. Figure 1 shows this assumption.

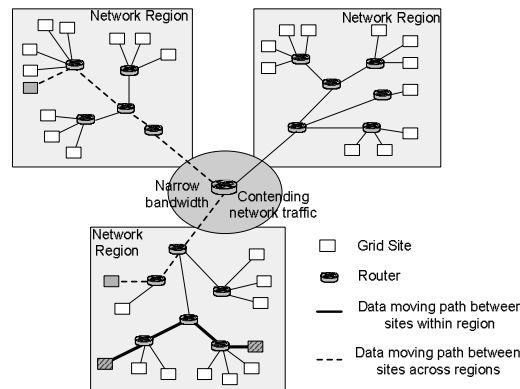


Figure 1. Bandwidth hierarchy in Internet

We assume that group of sites are located on the same network region. A network region is a network topological space where sites are located closely. This network region can be seen as an Internet topology within a country. It is generally known that lower bandwidth can be allocated for network link between sites across countries than link between sites within a country. In many cases, network region may be usually correspondent to geographical space like a country or a continent. If the

required replica is in the same region, the job is able to fetch the replica easily since broader bandwidth can be provided within a region. In contrast, if the required replica is located at the site in other region, much time will be consumed to fetch this replica via many links including highly congested one. Thus, a form of locality emerges which we call network-level locality. Main purpose of BHR strategy is to maximize this network-level locality within job execution model in data grid. BHR tries to replicate files which are likely to be used frequently within the region in near future. BHR optimizer runs both on a region and on a site cooperating with each other. Figure 2 describes detail of BHR algorithm.

```

BHR Optimizer (region):
    Keep track of names of stored files and access frequency of each file within the region;
BHR Optimizer (site):
1.  if (needed replica not in a site)
        Fetch replica from other site;
2.  Proceed to execute the job with the replica;
3.  if (free space in SE to store new replica)
        Store it;
    else {
4.      if (new replica is duplicated in other sites within region) {
            Terminate optimizer;      // avoid duplication
        else {
            Sort files in SE in the order of less frequently accessed;
            for (each file in SE) {
                if (file is duplicated in other sites within region)
                    delete it;
                if (enough free space to store new replica)
                    break;
            } }
5.      if (!enough free space) {
            Sort files in SE in the order of less frequently accessed;
            // it's based on the access history gathered by region-optimizer
            for (each file in sorted list) {
                if (access frequency of new replica > access frequency of the file)
                    delete file;
                if (enough free space)
                    break;
            } }
        if (enough free space)
            Store new replica;}

```

Figure 2. BHR replication algorithm

The access frequency gathered by region-optimizer means number of file requests made by jobs run on the sites within a region. It reflects regional popularity of files. If the job fetches a file from other sites and the SE is already filled up with replicas, we should determine whether storing newly received file is beneficial. If it turns out to be profitable, then we choose a file that should be deleted in order to store new replica. We apply 2-step decision process. First one is avoiding duplication. The procedure 4 in Figure 2 locates variety of replicas as many as possible in the region without duplication. Secondly, we take account of popularity of files as represented

by procedure 5. In data grid, there can be popularity of file accesses, that is, certain files will be requested more frequently than others by grid job. While the previous strategies consider popularity of files at the site level, we focus on access popularity at the region level. BHR replaces unpopular files from the regional point of view. By applying above two steps, chance of hitting network-level locality can be maximized.

4. Experiments

4.1 Simulation Tool

We evaluate the performance of BHR by implementing it in OptorSim, a data grid simulator developed to test dynamic replication strategies [8]. In OptorSim, general job execution scenario for data grid is modeled and various dynamic replica optimizers are implemented to test their effectiveness. After jobs are distributed to grid sites through Broker, they run on CE (Computing Element) at each site. Each job in CE has list of required replicas. For the first phase of replica optimization, Optimizer selects the best site to fetch the replica based on the available network bandwidth between sites. Then, Optimizer performs the second phase of optimization (dynamic optimization) by deciding whether storing (replicating) fetched file is beneficial or not.

4.2 Configuration

We perform simulation with assumed grid network topology and job execution scenarios. Figure 3 describes the network topology assumed in our simulations. We assume there are 4 regions and each region has 13 sites on the average. File transfer time is decided according to the narrowest bandwidth along the path to the destination. Broader bandwidth can be provided between sites within a region whereas bandwidth between sites across region is relatively narrow. Since many sites within a region try to fetch files from other region through single inter-region link, this inter-region link is highly congested with network traffic and it causes hierarchy of bandwidth.

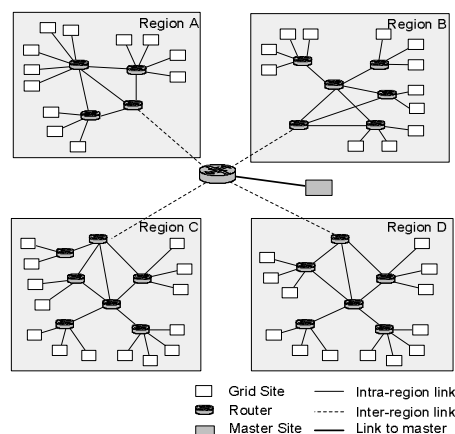


Figure 3. Grid topology in simulation

In data grid environment, various job execution scenarios will be present. We try to apply general job execution scenarios as presented by Table 1. In the simulation, 1000 jobs are distributed to grid sites through broker. According to the file set each job accesses, we classify jobs into 50 job types. Each file set consists of 15 files. We assume that certain preference of job types appears. This preference of job type makes popularity of files. Each job sequentially requests access to files in a file set. There is no overlap between file sets each job type accesses and the size of single file is 1 GB. Therefore, total size of data in this configuration is 750 GB (50 job types * 15 files in a file set * 1 GB for each file). We assume all files are initially held at master site. Replication takes place after jobs start to execute at each site.

Table 1. General configuration of parameters

| Parameters | Values |
|---------------------------------|--------|
| Number of jobs | 1000 |
| Number of job types | 50 |
| Number of file accessed per job | 15 |
| Size of single file | 1 GB |
| Total size of files | 750 GB |

4.3 Results

We compare the performance of BHR with site-level file replacement schemes, LRU Delete and Delete Oldest. In LRU Delete, the least recently accessed file is chosen for deletion whenever replacement takes place. Delete Oldest is another replacement-based scheme which deletes the oldest file in SE first when newly required replica is received and replacement is necessary. In order to easily interpret the result, we assume that all network links within region (intra-region) show same bandwidth. And also all inter-region link bandwidths are assumed to be the same in the scenario. Initially, we roughly set the bandwidth and storage space as shown in Table 2. We set the bandwidth between master site and its adjacent router as 2000 Mbps. It is much broader than other links to avoid effects from network traffic congestion at master site.

Table 2. Bandwidth and storage spaces

| Parameters | Values |
|-------------------------|-----------|
| Intra-region bandwidth | 1000 Mbps |
| Inter-region bandwidth | 1000 Mbps |
| Master-router bandwidth | 2000 Mbps |
| Storage space at site | 50 GB |

Figure 4 shows the achieved results with initial parameters. BHR takes the least total job execution time among strategies. It takes 33,174 seconds which is about 30 % less than other strategies. Since size of SE at each site, 50 GB, is not enough to hold large portion of overall data (750 GB), we cannot achieve much performance improvements with site-level replacement schemes. However, BHR strategy takes benefit from network-level locality by locating variety of files in a region as many as possible. Also, it locates files which are likely to be used in the region based on regional access history. In this simulation, Delete LRU and Delete Oldest show almost the same job execution time. The reason is that we do not assume any specific file access pattern in the data grid system.

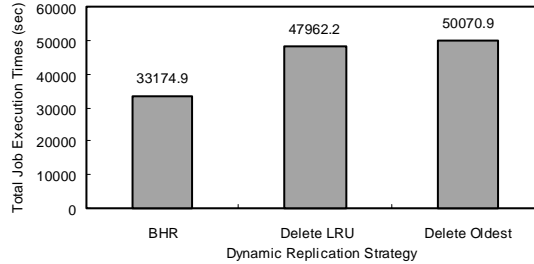


Figure 4. Total job times with parameters shown in Table 1 and Table 2

We continue performance evaluation with varying bandwidths and storage spaces. As we increase the inter-region bandwidth, file transfer via inter-region link does not take long although it is highly congested with data traffics. Thus, hierarchy of bandwidth becomes indistinct. Here we show results with varying bandwidths in Figure 5 (a). When we set narrow bandwidth on the inter-region link, BHR outperforms other strategy considerably. However the differences of job execution time become smaller as broader inter-region bandwidth is set. Finally, the difference becomes negligible when more than 1800 Mbps is provided for inter-region bandwidth. We can conclude that BHR strategy can be effectively utilized when hierarchy of bandwidth appears apparently. Size of SE in a grid site also affects the result significantly. As we mentioned, traditional replacement-based scheme can be effective when large storage space is provided in a grid site. In Figure 5 (b), as the size of storage space decreases in grid sites, BHR outperforms other strategies greatly. However, as the storage size increases, job execution time of two replacement-based schemes reduces sharply. The reason is that the file hit ratio in a site increases when large number of replicas can be stored in a site, and regional file hit ratio also increases though no region-based optimization strategy is applied. After all, efficiency of all three strategies become almost the same when large quantity of storage is provided at a site. Our BHR strategy can be more effective when grid sites have relatively smaller storage. One may argue that 100 GB is not an impractical storage size for a grid site. However, enough size of storage which makes the site-level replacement schemes effective is relative to the total size of data in a data grid system. In this simulation, only 750 GB of data is assumed to be in data grid while, in practice, terabyte or even petabyte scale of data is expected to be common in data grid.

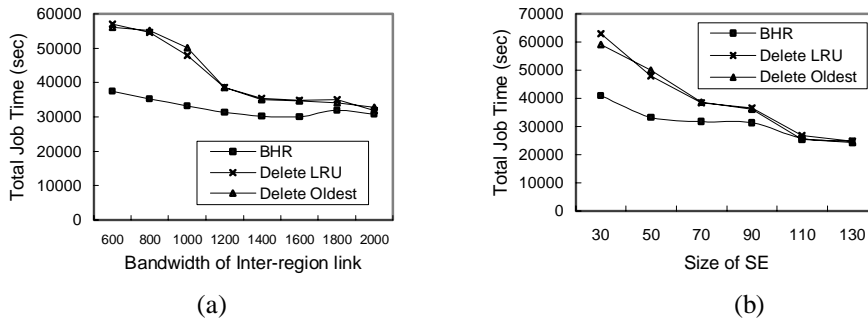


Figure 5. Total job time with varying bandwidth and storage size

5. Conclusion and Future Works

In this paper, we propose novel dynamic replica optimization strategy which is based on the network-level locality. BHR tries to replicate popular files as many as possible within a region, where broad bandwidth is provided between sites. The simulation results show that BHR takes less job execution time than other strategies especially when grid sites have relatively small size of storage and hierarchy of bandwidth clearly appears. BHR extends current site-level replica optimization study to more scalable way by exploiting network-level locality.

In our future work, we plan to survey actual Internet topology for data grid and study on how we group the grid sites as a region. Also, we will collect the experimental data such as data access patterns from real data grid applications and apply it to our BHR strategy to verify its performance in practical applications.

References

- [1] William H. Bell, David G. Cameron, Luigi Capozza, A. Paul Millar, Kurt Stockinger, and Floriano Zini.: Simulation of Dynamic Grid Replication Strategies in OptorSim. In Proc. of the 3rd Int'l. IEEE Workshop on Grid Computing (Grid'2002), Baltimore, USA, November 2002. Springer Verlag, Lecture Notes in Computer Science.
- [2] William H. Bell, David G. Cameron, Ruben Carvajal-Schiaffino, A. Paul Millar, Kurt Stockinger, and Floriano Zini.: Evaluation of an Economy-Based File Replication Strategy for a Data Grid. In International Workshop on Agent based Cluster and Grid Computing at CCGrid 2003, Tokyo, Japan, May 2003. IEEE Computer Society Press.
- [3] Mark Carman, Floriano Zini, Luciano Serafini, and Kurt Stockinger.: Towards an Economy-Based Optimisation of File Access and Replication on a Data Grid. In International Workshop on Agent based Cluster and Grid Computing at International Symposium on Cluster Computing and the Grid (CCGrid'2002), Berlin, Germany, May 2002. IEEE Computer Society Press.
- [4] Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury and Steven Tuecke.: The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets. *Journal of Network and Computer Applications*, 23:187-200, 2001.
- [5] EU Data Grid Project: <http://www.eu-datagrid.org>
- [6] I. Foster, C. Kesselman and S. Tuecke.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, 15(3), 2001.
- [7] Wolfgang Hoschek, Javier Jaen-Martinez, Asad Samar, Heinz Stockinger and Kurt Stockinger.: Data Management in an International Data Grid Project. 1st IEEE/ACM International Workshop on Grid Computing (Grid'2000), Bangalore, India, Dec 2000.
- [8] OptorSim – A Replica Optimizer Simulation: <http://edg-wp2.web.cern.ch/edg-wp2/optimization/optorsim.html>
- [9] Sang-Min Park and Jai-Hoon Kim.: Chameleon: A Resource Scheduler in a Data Grid Environment. 2003 IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'2003), Tokyo, Japan, May 2003. IEEE Computer Society Press.
- [10] Kavitha Ranganathan and Ian Foster.: Design and Evaluation of Dynamic Replication Strategies for a High Performance Data Grid. International Conference on Computing in High Energy and Nuclear Physics, Beijing, September 2001.
- [11] Kavitha Ranganathan and Ian Foster.: Identifying Dynamic Replication Strategies for a High Performance Data Grid. International Workshop on Grid Computing, Denver, November 2001.