PROCEEDINGS OF SCIENCE

Simulation of the cache hit rate for data readout at the Tokyo Tier-2 center

T.Kishimoto*, **J.Tanaka, T.Mashimo, M.Kaneda, N.Matsui**
International Center for Elementary Particle Physics, The University of Tokyo
E-mail: tomoe@icepp.s.u-tokyo.ac.jp

The Tokyo Tier-2 center, which is located in the International Center for Elementary Particle Physics at the University of Tokyo, provides computing resources to the ATLAS experiment in the Worldwide LHC Computing Grid. In order to improve the I/O performance and scalability of file servers in the future system, a possibility of introducing a cache system using fast devices such as SSD is under discussion. Therefore, a simulation has been performed to understand the cache behavior using past data access logs at the center. This paper reports a method of the simulation and gives a discussion about its results.

International Symposium on Grids & Clouds 2019, ISGC2019
31st March - 5th April, 2019
Academia Sinica, Taipei, Taiwan

*Speaker.

PoS(ISGC2019)030

1. Introduction

The Tokyo Tier-2 center, which is located in the International Center for Elementary Particle Physics (ICEPP) [1] at the University of Tokyo, has been providing computing resources to the ATLAS experiment [2] in the Worldwide LHC Computing Grid (WLCG) [3]. The official site operation in the WLCG was launched in 2007 after several years of development. The site has been achieving a stable and reliable operation since then.

Hardware devices in the center were upgraded in every three years in order to satisfy the requirement of the ATLAS experiment. In the current system, a high disk I/O utilization of file servers, which is due to a large amount of data accesses from worker nodes, has been observed. The number of concurrently running jobs in the center has increased by the system upgrade, and it will increase in the future system. Furthermore, a study about utilizing external computing resources such as commercial clouds and high performance computer (HPC) as a part of the center is in progress. These upgrade and extensions will produce additional data accesses to the file servers. In order to improve the I/O performance and scalability of the file servers in the future system, a possibility of introducing a cache area to the storage system using fast devices such as SSD is under discussion. However, an efficient caching system can not be designed without a knowledge of data access patterns. Therefore, a simulation has been performed to understand the cache behavior using past data access logs at the center. This paper reports a method of the simulation and gives a discussion about its results. The paper is organized as follows: Section 2 describes the configuration of the Tokyo Tier-2 center. Section 3 provides the status of the current system. Section 4 and 5 give details of the simulation and its results. Section 6 shows an examination of cache deployment using the XCache. Section 7 summarizes this paper.

2. Configuration of the Tokyo Tier-2 center

In the Tokyo Tier-2 center, almost all hardware devices are supplied by a lease and replaced in every three years. The current system (so-called 5th system) has been in operation since January 2019, and is stably running.

Table 1 summarizes the computing resources of the 5th system, which are reserved for the ATLAS experiment as the WLCG Tier-2 site. The table also shows the computing resources in the previous system (4th system). In the 5th system, each worker node has 32 CPU cores and 96 GB memory. The Hyper-Threading Technology is not enabled on the worker nodes. The effective capacity of local disks on the worker node is 1 TB, which is configured by RAID1. The ARC-CE [5] is deployed as the computing element in front of the HTCondor [6] batch job scheduler. The disk storage system consists of 24 file servers. Each file server has two disk arrays, which are connected by 2×16G fibre channels. The effective capacity of each disk array is 220 TB, which is configured by RAID6. The disk storage system is managed by the DPM [7].

Each worker node has a 10 Gbps Ethernet interface. A group of 16 worker nodes are connected to a core switch via an edge switch with a 40 Gbps bandwidth. Each file server is directly connected to the core switch with a 25 Gbps bandwidth. The core switch is connected to SINET, which is a Japanese national research and educational network, with 2×10 Gbps bandwidth. Figure 1 shows an overview of the local area network configuration in the 5th system.

	# of CPU cores	HEPSPEC06 [4] / core	Disk capacity
4th system	6144 (256 worker nodes)	18.11 (SL6)	6,336 TB (48 files servers + 48 disk arrays)
5th system	7680 (240 worker nodes)	18.97 (CentOS7)	10,560 TB (24 files servers + 48 disk arrays)

Table 1: The computing resources deployed as the WLCG Tier-2 site. Only the ATLAS experiment is supported.

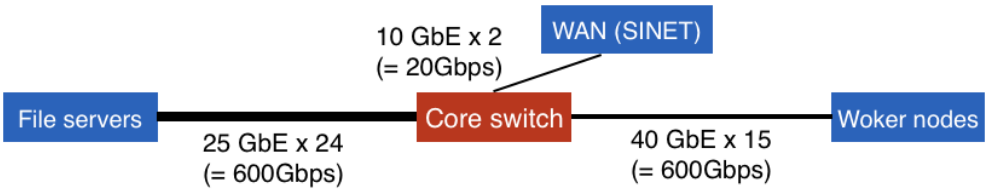


Figure 1: The configuration of local area network in the 5th system.

3. Status of the 5th system

Figure 2 shows a network traffic of the file servers and the worker nodes. A large network traffic from the file servers to the worker nodes, which reaches 100 Gbps, is observed. This is because typical Grid jobs of the ATLAS experiment at the Tokyo Tier-2 center copy data from the file servers to its worker nodes. The local network bandwidth from the file servers to the worker nodes is 600 Gbps as illustrated in Figure 1. The local network bandwidth of the system is still sufficient for these data transfers.

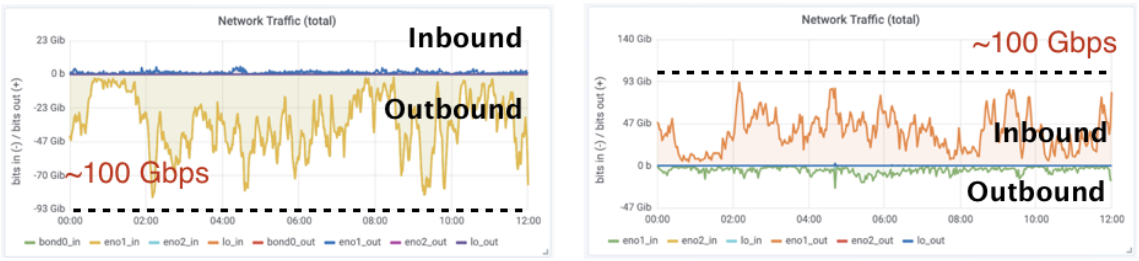


Figure 2: The network traffic of file servers (left) and worker nodes (right).

Figure 3 shows the disk I/O utilization of a file server and a worker node. The disk I/O utilization is defined as the percentage of elapsed time during which I/O requests were issued to the device. It can be confirmed that the disk I/O utilization reaches 75% for the both file server and the worker node. In particular, the disk I/O utilization on the worker node is often saturated due to data copies to the local disks. Several improvements of the disk I/O performance of the worker node are under consideration. The configuration change of the local disks from RAID1 to RAID0 can improve the write performance. The direct access to the file servers using XRootD protocol [8] can also avoid writing data to the local disks. It is important to improve the I/O performance and scalability of the file servers as well. The cache system using fast devices is a common approach

to improve the I/O performance. However, performances of the cache system depend on the data accesses pattern. Therefore, a simulation has been performed in order to understand the cache behavior in the center.

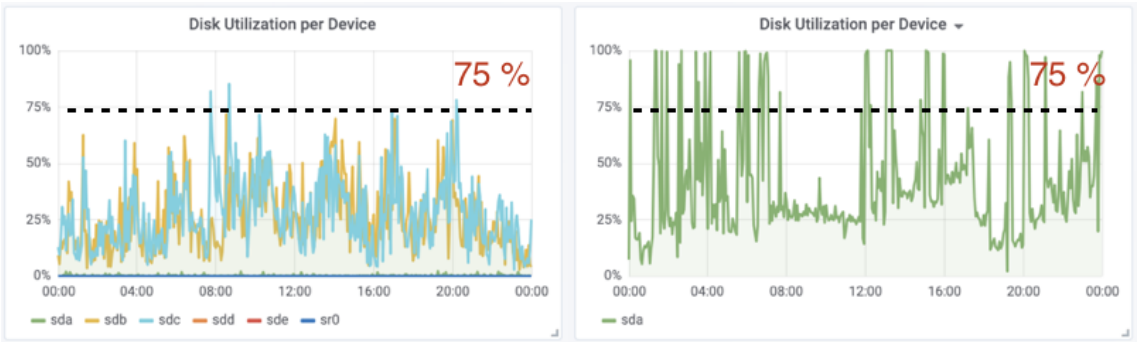


Figure 3: The disk I/O utilization of a server (left) and a worker node (right).

4. Simulation of the cache hit rate

The simulation of the cache hit rate is performed using past file transfer logs at the Tokyo Tier-2 center. The transfer logs for two years (2017–2018) in the 4th system are used. Figure 4 shows an overview of the simulation.

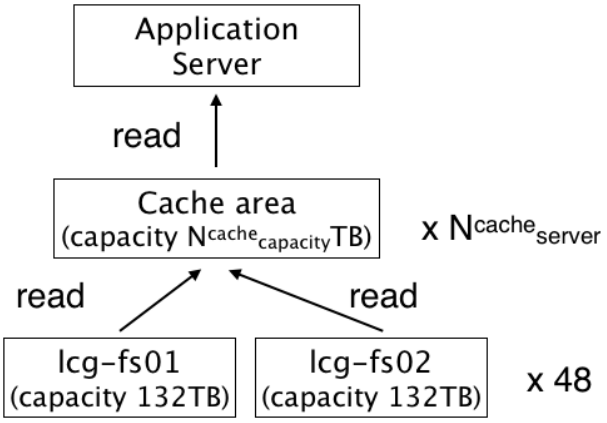


Figure 4: Overview of the cache simulation.

There were 48 file servers named lcg-fs01 to lcg-fs48 in the 4th system. Each file server had 132 TB of disk space. The simulation creates cache areas between the file servers and application servers such as worker nodes. The past transfer logs provide information about timestamp of transfer, file name, file size and file path including file server name. In the simulation, the transfer logs are analyzed according to the order of timestamps, and the file is cached to the cache areas. The files in the cache areas are managed based on cache algorithms. The following three parameters are changed in the simulation:

- $N^{cache}_{capacity}$: The storage capacity of each cache area.

- N^{cache}_{server} : The number of cache areas (servers).
  - If N^{cache}_{server} = 48, each file server has a cache area.
  - If N^{cache}_{server} = 1, one cache area covers all file servers.
- Cache algorithm: Three cache algorithms are examined.
  - Least Recently Used (LRU): LRU algorithm deletes the least recently used items first.
  - Least Frequently Used (LFU): LFU deletes the least often used items first.
  - Adaptive Replacement Cache (ARC): ARC algorithm [9] has two queues (T1 and T2) for recently and frequently referenced entries. T1 and T2 have ghost queues, which have only metadata for keeping track. The algorithm continuously revises how to invest in T1 and T2 according to the access pattern using the ghost queues.

Only the read data accesses is considered in the simulation. The following two efficiencies are defined to evaluate the data access pattern:

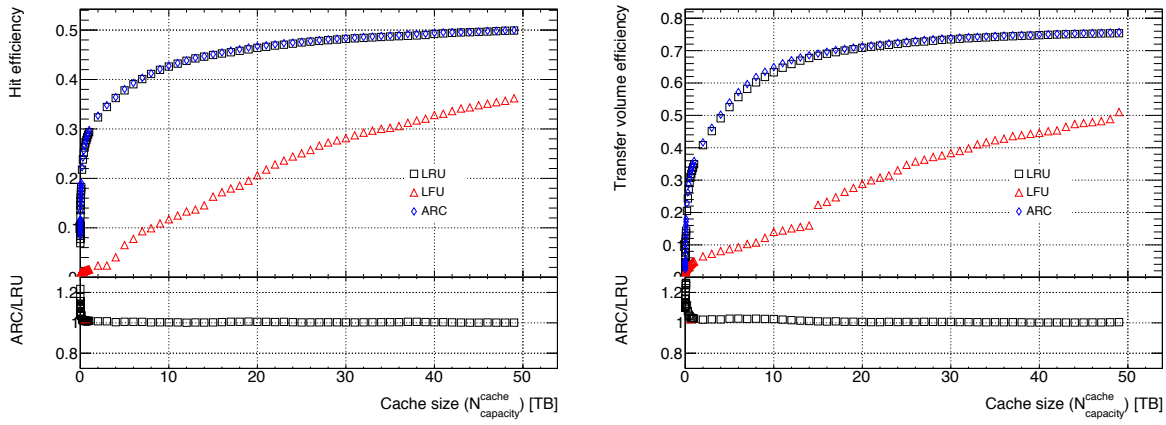
- Hit efficiency (hit rate) = (# of data accesses from cache areas) / (Total # of data accesses)
- Transfer efficiency = (Transfer volume from cache areas) / (Total transfer volume)

5. Simulation results

Figure 5 shows simulated results of the hit efficiency and transfer efficiency in terms of the $N^{cache}_{capacity}$. The efficiencies are shown for the different cache algorithms. In the figure, the N^{cache}_{server} is set to 48, and the efficiencies of the cache area, which is associated to lcg-fs01 are shown. It can be confirmed that the LRU and ARC algorithms show similar performances. The ARC algorithm shows +20% efficiencies compared to the LRU algorithm if the $N^{cache}_{capacity}$ is very small (< 100 GB). If the $N^{cache}_{capacity}$ is 1 TB, which is about 1% of the file server disk space, the cache hit rate will be about 30%. If the cache area is constructed with $N^{cache}_{capacity}$ = 10 TB, which is about 10% of the file server disk space, almost maximum efficiencies will be achieved. These efficiency curves provide key inputs to determine the capacity of cache area in the system.

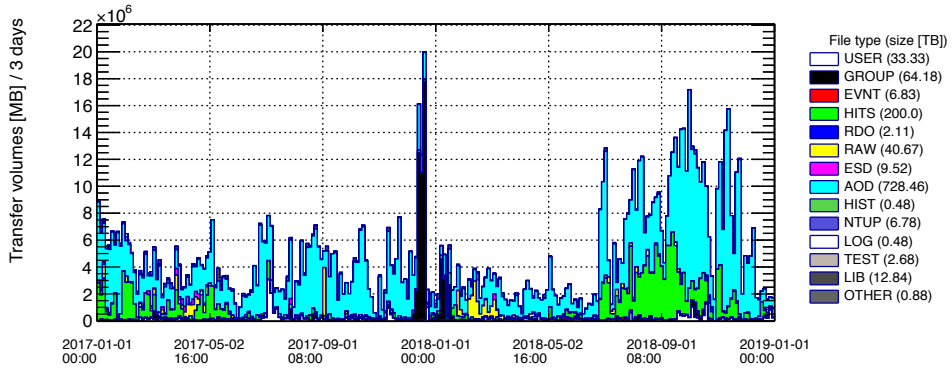
Figure 6 shows the total transfer volumes and the transfer volumes from cache area for different types of files. It can be seen that the HITS (green histogram) and AOD (light blue histogram) files dominate the transfer volumes. The HITS are files produced from detector simulations in the ATLAS experiment. The AOD files are also produced from reconstruction, which are mainly used for user physics analysis. It can be confirmed that the cache works effectively for the HITS files. This is because common HITS files are frequently accessed by the pile-up simulation in the ATLAS experiment. The transfer efficiency for the HITS file is 0.66 in this simulation condition. On the other hand, the transfer efficiency for the AOD file is low, which is 0.22, because a wide variety of AOD files are accessed.

Figure 7 shows the hit efficiency and transfer efficiency in terms of the N^{cache}_{server}. The total $N^{cache}_{capacity}$ is set to 48 TB. In other words, there are 48 cache areas, which has 1 TB storage capacity, if the N^{cache}_{server} = 48, and there is one cache area, which has 48 TB storage capacity, if the N^{cache}_{server} =

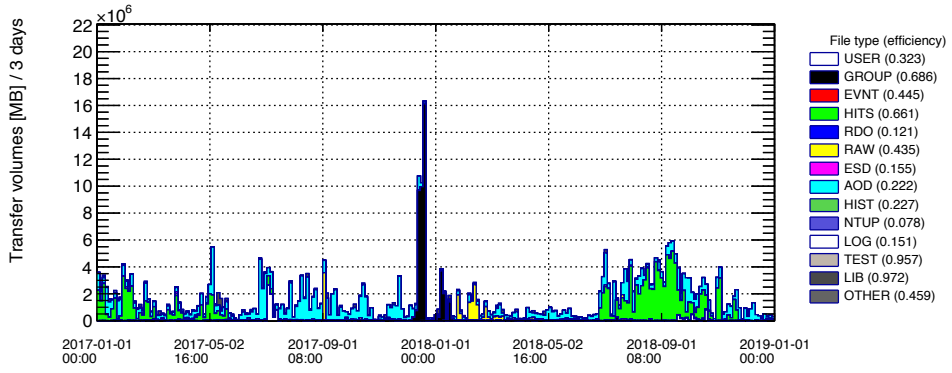


(a) (b)

Figure 5: The hit efficiency (a) and transfer efficiency (b) in terms of the $N^{cache}_{capacity}$ for the different cache algorithms. N^{cache}_{server} is 48. The efficiencies of the cache area, which is associated to lcg-fs01 are shown.



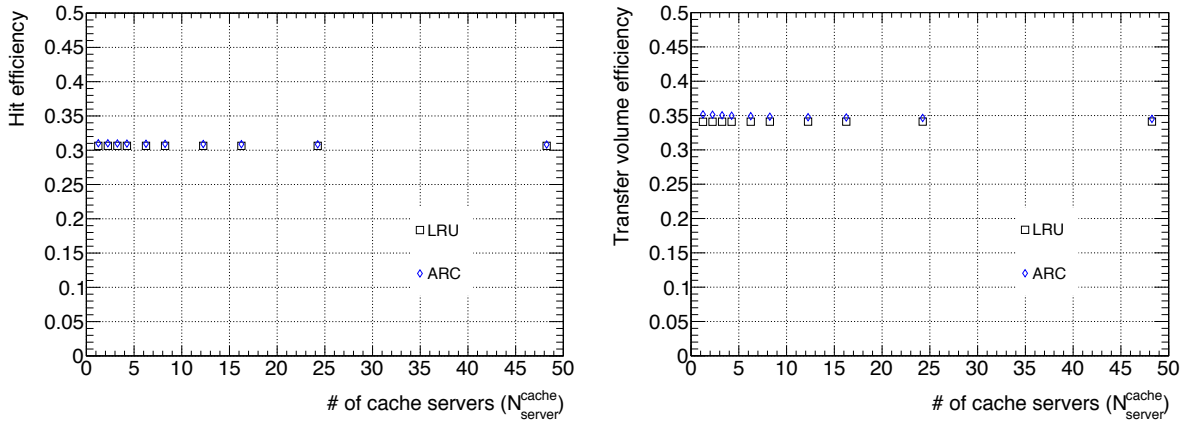
(a)



(b)

Figure 6: The total transfer volumes (a) and the transfer volumes from cache area (b) for different types of files. $N^{cache}_{capacity}$ is 1 TB and N^{cache}_{server} is 48. The transfer volume from the cache area, which is associated to lcg-fs01 is shown. The cache algorithm is LRU.

1. The variation in the efficiencies for the N^{cache}_{server} is less than 1%. In terms of securing network bandwidth and I/O performances, it is required to prepare multiple cache areas (servers). The simulation confirms that the efficiencies are stable when the multiple cache areas are deployed.



(a) (b)

Figure 7: The hit efficiency (a) and transfer efficiency (b) in terms of the N^{cache}_{server}. The total $N^{cache}_{capacity}$ is set to 48 TB.

6. Cache deployment using XCache

The XCache is a software to provide a cache system, which primarily uses the XRootD protocol. The XCache was deployed, and the cache hit rate was also measured using the real ATLAS Grid jobs. Worker nodes and a XCache server were deployed on Google Cloud Platform (GCP) as illustrated in Figure 8.

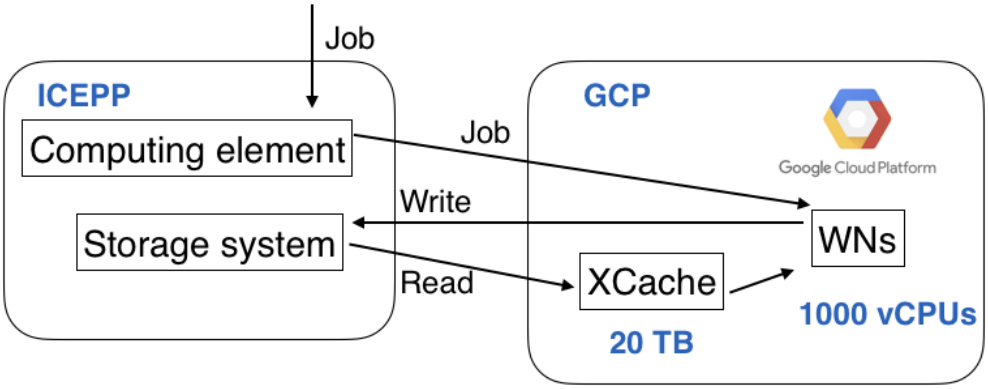


Figure 8: Overview of the configuration of the XCache deployment.

The computing element and the storage system are deployed on the Tokyo Tier-2 at the ICEPP. The cloud resources were used in this test because it is easy to purchase and scale the resources for quick tests. The XCache server was operated for 7 days, and it had 20 TB storage capacity as a cache area. Figure 9 shows the network traffic of the XCache server during the test. The cache was

working well in the test because the sent network traffic was larger than the received network traffic. Table 2 summarizes the observed hit efficiency and transfer efficiency using the XCache server in comparison with results of the simulation. Reasonable agreements with the simulation have been observed by the XCache test. Therefore, the simulation can provide reasonable information to consider the cache construction in the future. The small discrepancies in the comparison are due to a different access pattern between the simulation and the XCache test.

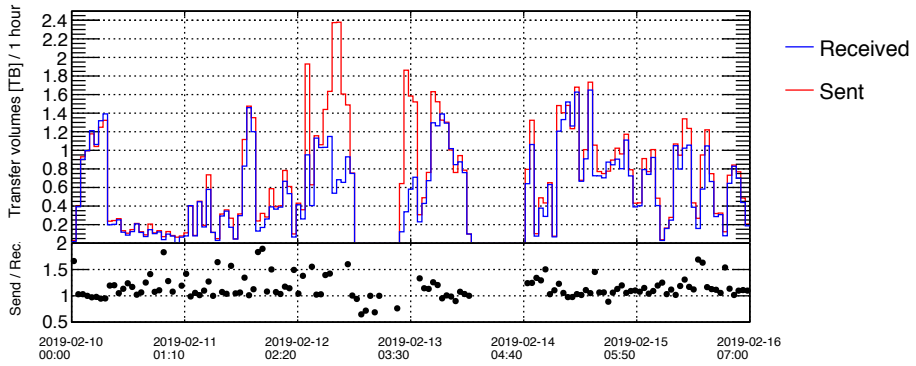


Figure 9: The network traffic of the XCache server. The network traffic is sampled using iftop command in Linux.

	HIt efficiency	Transfer efficiency
Observation (XCache)	0.20	0.19
Simulation	0.24	0.26

Table 2: The observed hit efficiency and transfer efficiency using the XCache. The simulation is performed with the $N^{cache}_{capacity}$ = 20 TB and the N^{cache}_{server} = 1. The cache algorithm is LRU.

7. Summary

The Tokyo Tier-2 center has been providing computing resources to the ATLAS experiment in the WLCG. The simulation of the cache hit rate has been performed using the past transfer logs at the Tokyo Tier-2 in order to consider a possibility to construct the cache system in the future. The simulation confirmed that:

- The LRU algorithm shows similar performances compared to the ARC algorithm. The cache hit rate is about 30% if the cache capacity is 1% of the total storage capacity.
- The cache hit rate depends on file types. The cache works effectively especially for the HITS file for the pile-up simulation in the ATLAS experiment.
- Variations in the cache efficiencies for the N^{cache}_{server} are small, which are less than 1%. It indicates that multiple cache areas can be deployed to secure the network bandwidth and I/O performance.

- The simulation shows reasonable agreements with the XCache test using the real jobs.

These results conformed that the simulation can provide useful inputs to consider the cache construction in the future. We are planning to investigate and purchase cache hardwares based on the simulation, and measure its I/O performances.

References

[1] ICEPP web page, https://www.icepp.s.u-tokyo.ac.jp/en/index.html

[2] ATLAS Collaboration, 2008 JINST 3 S08003.

[3] WLCG web page, http://wlcg.web.cern.ch/

[4] HEP-SPEC06 web page, http://w3.hepix.org/benchmarks/doku.php

[5] ARC-CE web page, http://www.nordugrid.org/arc/ce/

[6] HTCondor web page, https://research.cs.wisc.edu/htcondor/

[7] DPM web page, http://lcgdm.web.cern.ch/dpm

[8] XRootD web page, http://xrootd.org/

[9] Nimrod Megiddo and Dharmendra Modha, In Proceedings of the 2003 Conference on File and Storage Technologies (FAST) (2003)