타키온 클러스터 시스템의 병렬 분산 파일 시스템 워크로드 분석

조혜영°, 김성호, 이식
한국과학기술정보연구원 슈퍼컴퓨팅본부
e-mail:{chohy°, sungho, siklee}@kisti.re.kr

Analysis of Parallel and Distributed File System Workloads on Tachyon Cluster System

Hyeyoung Cho°, Sungho Kim, Sik Lee
Supercomputing Center
Korea Institute of Science and Technology Information

요 약

클러스터 시스템의 응용 분야가 다양화되고 복잡해짐에 따라, 대규모 클러스터 시스템을 보다 효율적으로 사용하기 위해서 실제 사용자의 이용 패턴을 예측할 수 있는 워크로드 분석의 필요성이 높아지고 있다. 이에 본 논문에서는 현재 가동중인 188개의 계산 노드, 3008개 CPU 자원을 보유한 대규모 클러스터 시스템에서 병렬 분산 파일 시스템에 대한 워크로드를 분석하였다.

1. 서론

클러스터 시스템의 응용 분야가 다양화되고 복잡해짐에 따라, 대규모 클러스터 시스템을 보다 효율적으로 사용하기 위해서 실제 사용자의 이용 패턴을 예측할 수 있는 워크로드 분석의 필요성이 높아지고 있다. 특히 파일 시스템은 사용자의 사용 패턴에 따라 최적화함으로써 많은 성능 향상을 가져올 수 있다[1]. 이러한 워크로드 분석은 시스템의 운영뿐 아니라 새로운 클러스터 시스템을 설계하는데도 요구되는 요소이다. 이에 본 논문에서는 188노드, 3000 CPU 이상 규모의 클러스터 시스템에서 병렬/분산 파일 시스템에 대한 워크로드를 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 I/O 워크로드 관련 연구를 살펴보고, 3장에서는 워크로드 분석 대상인 타키온 클러스터 시스템의 시스템 환경과 사용자 환경을 설명하였다. 4장에서는 분석한 워크로드 분석 결과를 설명하고, 마지막으로 5장에서는 결론에 대하여 기술한다.

2. 관련연구

I/O 워크로드에 관한 선행된 연구들을 살펴보면, [2,3,4,5,6]은 I/O 집중적인 응용과학 프로그램에 대해서 접근 패턴(access pattern)과 I/O request 크기와 파일 크기 등의 워크로드에 대해 분석하였다. [1, 7]은 윈도우 NT와 UNIX 시스템에서 파일 시스템의 이벤트를 추적(trace)하여 워크로드를 분석하였다. 그러나 대부분의 관련 연구들이 1990년대에 이루어진 연구들이다. 따라서 최근의 고성능 컴퓨팅 환경이나 고성능 네트워크 환경을 반영하지 못하고 있다. 또한 연구의 대부분이 로컬 파일 시스템에 대한 연구이거나, 분산 파일 시스템에 대한 연구라도 분석 대상이 특정 응용 프로그램에 제한되어 있다.

3. 워크로드 분석 환경

본 논문에서는 클러스터 시스템의 파일 시스템에서 요구되는 워크로드 정보를 확보하기 위하여 현재 운영중인 188노드 규모의 클러스터 시스템(TACHYON)[1]을 대상으로 파일 시스템의 특성을 분석하였다. 클러스터 시스템은 SUN의 Blade 6048 시스템을 기반으로 구성되었으며, 이론최고성능(Rpeak) 24TFlops인 테라플롭스급 클러스터 시스템이다. 타키온 시스템은 2008년 6월 32회 Top500리스트[8]에서 130위를 차지한 고성능 시스템이다.

타키온의 스토리지 시스템을 살펴보면, 스크래치와 홈 파일 시스템을 위해 19대의 SUN X4500시스템을 연동하여 스토리지 시스템을 구성하였다. 이를 통해 11TB 용량의 홈 디렉토리(/home01)와 두 개의 글로벌 스크래치 디렉토리(54TB의 /work01, 40TB의 /work02를 제공하고 있고, Lustre[9] 병렬 분산 파일시스템을 사용하고 있다. 본 논문에서는 워크로드 분석 대상은 사용자의 기본 환경 설정 파일들과 사용자 소스 데이터들를 포함하는 사용자 홈 디렉토리(/home)와 사용자들이 대용량의 I/O 작업을 실행할 때 주로 사용하는 임시 작업 디렉토리(/ytmp)인 스크래치(scratch)를 대상으로 하였다.

4. 워크로드 분석

4.1 파일 종류별 분포

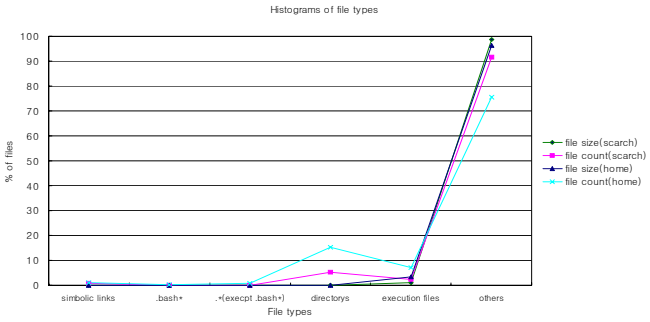

그림 3. 파일 종류별 분포

본 절에는 파일 종류에 대한 분석 결과를 정리한다. 그림 3은 파일 종류에 따라 파일 수와 디스크 스페이스 기준으로 통계를 낸 것이다. 파일 수 기준으로 볼 때, 홈 디렉토리가 스크래치 디렉토리에 비해서 10.06% 더 많은 비율을 보였다. 즉 사용자들이 임시적으로 대용량의 I/O 작업을 실행할 때 주로 사용하는 스크래치 스토리지에 비해, 홈 스토리지를 디렉토리로 분류하여 체계적으로 관리하고 있다는 사실을 확인할 수 있다.

4.2 파일 크기

1) 클러스터 시스템(TACHYON)은 한국과학기술연구원(KISTI) 슈퍼컴퓨팅센터 슈퍼컴퓨터 4호기의 일부 시스템으로 2008년에 구축되어, 운영되고 있음.

본 절에는 파일 크기에 대한 분석 결과를 정리한다. 파일 크기는 파일 시스템에서 블록(block) 크기와 같은 설정을 결정하는데 있어서 중요한 특성이 된다. 예를 들어 파일 시스템의 블록 크기를 크게 설정할 경우, 시스템에서 주로 사용하는 파일의 크기가 크다면 성능에 도움이 되지만 시스템에서 주로 사용하는 파일이 크기가 작은 파일이라면 스페이스 낭비가 발생할 수 있다[10].

타키온 클러스터 시스템의 스크래치(scratch) 공간에 대해 파일 크기에 따른 분포를 파일 수와 디스크 스페이스를 기준으로 분석한 결과를 그림 4에 나타내었다. 파일 수 기준으로 볼 때, 1MB이하 크기의 파일이 전체의 83% 정도를 차지하였고, 1-64MB 크기의 파일이 6% 로 높은 비율을 차지하였다. 파일 크기 기준으로 볼 때, 4-64MB 크기의 파일 18%로 가장 많았고, 100-200MB 크기의 파일이 13%를 차지하였다. 또한 1GB 이상의 파일이 용량 면에서 많은 비중을 차지하였는데, 이는 시스템 관리자가 큰 사이즈의 백업 파일들을 저장하고 있거나 사용자들이 스크래치 디렉토리를 데이터 백업용 이용하고 있기 때문이었다.

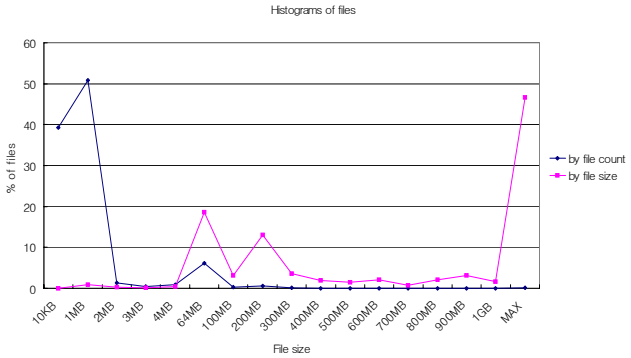

그림 4. 파일 크기별 분포

4.3 디렉토리당 파일 수

디렉토리당 소유하고 있는 파일의 전체 분포를 분석하기 위해서 디렉토리당 파일 개수의 누적 분포를 그림 5에 나타내었다. 10이하의 파일을 가지고 있는 디렉토리의 수가 스크래치 73.35%, 홈 디렉토리 77.53%를 차지하였다. 평균 디렉토리당 파일 수는 홈 16.85, 스크래치 46.15였으며, 최대 파일 수는 홈 20,748, 스크래치 153,497이었다.

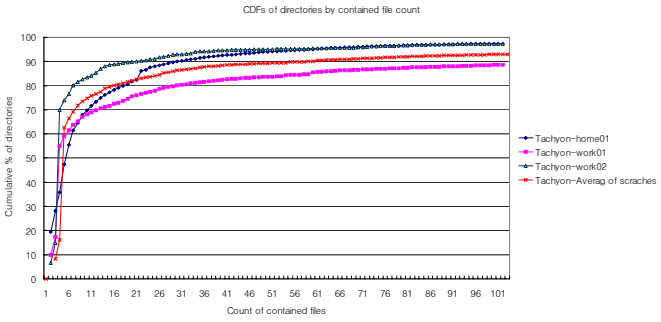

그림 5. 디렉토리당 파일 수 누적 분포

4.5 전체 데이터량 변화

클러스터 시스템의 데이터량의 변화를 예측하는 것은 클러스터를 구축하고 관리하는데 있어서 중요하다. 그림 7에 2009년 1월부터 2009년 6월 15일까지 기간 동안 타키온 시스템의 데이터량 변화 추이를 나타내었다. 본 워크로드의 분석 대상 중 홈 디렉토리는 사용자의 컴파일 환경과 소스들을 포함하고 있기 때문에 용량의 변화가 많지 않은 편이었다. 반면 사용자들이 주로 I/O 작업이 많은 프로그램을 실행할 때 사용하는 임시 작업 디렉토리인 스크래치는 홈 디렉토리에 비해 데이터량의 변화폭도 컸다. 전체 데이터량의 변화는 일부 동적인 데이터의 변화를 하루 단위로 보여준다. 분석 결과, 하루에 평균적으로 1,484GB(/home01), 28,782GB(/work01), 23,438GB(/work02)의 용량 증가를 보였다.

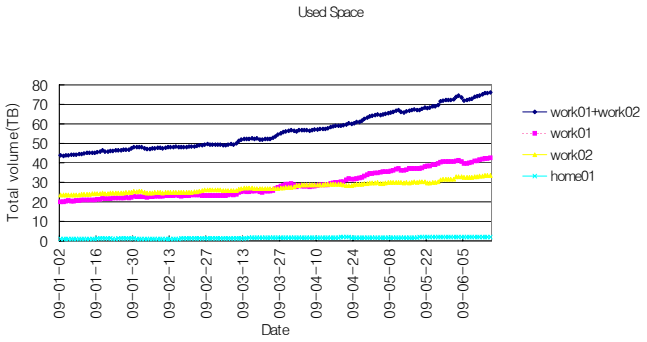

그림 7. 용량 변화 추이

5. 결론

클러스터 시스템에서 파일 시스템에 대한 워크로드 분석은 새로운 시스템의 설계 및 운영에 있어 중요한 자료이다. 이에 본 논문은 가동중인 188노드/ 3008코어 규모의 대규모 클러스터 파일 시스템에서 사용자 홈 디렉토리와 I/O 집중적인 프로그램을 수행할 때 주고 사용하는 임시 디렉토리인 스크래치를 대상으로 I/O 워크로드를 분석하였다.

파일 종류별 분석 결과, I/O 집중적인 프로그램을 수행할 때 임시로 사용하는 스크래치에 비해서 홈 디렉토리를 디렉토리로 보다 잘 분류하여 사용하고 있는 것으로 나타났다. 파일 크기는 파일 수 기준으로 1MB이하가 전체의 83%를 차지하였다. 디렉토리당 파일 수는 70%이상의 디렉토리가 10이하의 파일을 포함하고 있었전체 데이터량의 변화는 홈 스토리지에 비해 스크래치 스토리지가 변동량이 많았으며, 하루에 평균적으로 1,484GB(/home01), 28,782GB(/work01), 23,438GB (/work02)의 용량 증가를 보였다.

본 논문의 워크로드 분석 결과는 클러스터 시스템 도입이나 시스템의 관리 및 운영에 자료로 사용될 수 있다. 또한 클러스터 시스템을 위한 새로운 파일 시스템의 설계에도 반영될 수 있다. 그러나 본 논문의 워크로드 분석은 전체 데이터의 변화량 분석으로부터 일부 동적인 데이터의 변화를 하루 단위로 보였으나, 대부분은 정적(static) 정보들을 분석한 결과이여서 그 한계가 있다. 이에 향후 Access 패턴이나 빈도 등에 관한 동적(dynamic) 워크로드 분석에 관한 연구가 필요할 것으로 보인다.

참고문헌

[1] Drew Roselli, Jacob R. Lorch, and Thomas E. Anderson, "A comparison of file system workloads," Proc. of 2000 USENIX Annual Technical Conference, 2000.

[2] E. Smirni and D. A. Reed, "Workload characterization of input /output intensive parallel applications," Proc. of the Conference on Modeling Techniques and Tools for Computer Performance Evaluation, LNCS, Vol. 1245, Springer-Verlag, 1997.

[3] Phyllis E. Crandall and RuthA. Aydt and Andrew A. Chien and Daniel A. Reed, "Input/Output Characteristics of Scalable Parallel Applications," Proc. of Supercomputing 1995.

[4] Nils Nieuwejaar, David Kotz, Apratim Purakayastha, Carla Schlatter Ellis, and Michael L. Best, "File-Access Characteristics of Parallel Scientific Workloads," IEEE Transactions on Parallel and Distributed Systems, Vol.7, No. 10, pp. 1075~1089, 1996.

[5] Apratim Purakayastha, Carla Schlatter Ellis, David Kotz, Nils Nieuwejaar, and Michael Best, "Characterizing Parallel File-Access Patterns on a Large-Scale Multiprocessor," Proc. of the 9th IEEE International Parallel Processing Symposium", pp. 165~172, 1995.

[6] Feng Wang, Qin Xin, Bo Hong, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, and Tyce T. McLarty, "File System Workload Analysis For Large Scale Scientific Computing Applications," Proc. of the 21st IEEE / 12th NASA Goddard Conference on Mass Storage Systems and Technologies, 2004.

[7] J. Ousterhout, Da Costa, H. Harrison, J Kunze,M. Kupfer, and J. Thompson, "A Trace-Driven Analysis of the Unix 4.2 BSD File System," InProceedings of the 10th Symposium on Operat-ing Systems Principles, pages 15 - 24, December1985.

[8] Top500 Supercomputing sites, http://www .top500. org

[9] Lustre wep sites, http://www.lustre.org

[10] Dominic Giampaolo, "Practical File System Design with the Be File System," Morgan Kaufmann, 1999.