ISBN 978-89-6211-501-7

타키온 클러스터
I/O 워크로드 분석 보고서

부　서: KISTI 슈퍼컴퓨팅본부
작성자: 조혜영, 김성호

한국과학기술정보연구원

Korea Institute of Science and Technology Information

목 차

제 1장 서 론 3

제 2장 연구 동향 3

제 3장 워크로드 분석 환경 4

3.1 시스템 환경 4

3.2 사용자 환경 6

제 4장 워크로드 분석 7

4.1 파일 종류별 분포 7

4.2 파일 크기 8

4.3 디텍토리당 파일 수 10

4.4 디렉토리 DEPTH 10

4.5 전체 데이터량 변화 12

제 5 장 워크로드 분석 결과 및 향후 계획 14

제 6 장 워크로드 분석 상세 데이터 15

참고문헌 29

제 1 장 서 론

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. 이러한 워크로드 분석은 시스템의 운영뿐 아니라 새로운 클러스터 시스템을 설계하는데도 요구되는 요소이다. 이에 본 연구에서는 188 노드, 3000 CPU 이상 규모의 클러스터 시스템에서 병렬/분산 파일 시스템에 대한 워크로드를 분석하였다.

본 연구의 구성은 다음과 같다. 2 장에서는 기존의 I/O 워크로드 관련 연구를 살펴보고, 3 장에서는 워크로드 분석 대상인 타키온 클러스터 시스템의 시스템 환경과 사용자 환경을 설명하였다. 4 장에서는 분석한 워크로드 분석 결과를 설명하고, 5 장에서는 결론 및 향후 연구 계획에 대하여 기술한다. 6 장에서는 실제 워크로드 분석 상세 데이터 및 분석 과정에서 주의점들을 기술한다.

제 2 장 연구 동향

I/O 워크로드에 관한 선행된 연구들을 살펴보면, 대부분은 I/O 집중적인 응용과학 프로그램에 대해서 접근 패턴(access pattern)과 I/O request 크기와 파일 크기 등의 워크로드에 대해 분석하였다. 그리고 일부 연구들은 윈도우 NT 와 UNIX 시스템에서 파일 시스템의 이벤트를 추적(trace)하여 워크로드를 분석하였다. 그러나 대부분의 관련 연구들이 1990 년대에 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.

제 3 장 워크로드 분석 환경

3.1 시스템 환경

본 연구에서는 클러스터 시스템의 파일 시스템에서 요구되는 워크로드 정보를 확보하기 위하여 현재 운영중인 188 노드 규모의 클러스터 시스템(TACHYON)[1]을 대상으로 파일 시스템의 특성을 분석하였다. 클러스터 시스템은 SUN 의 Blade 6048 시스템을 기반으로 구성되었으며, 이론최고성능(Rpeak) 24TFlops 인 테라플롭스급 클러스터 시스템이다. 타키온 시스템은 2008 년 6 월 32 회 Top500 리스트에서 130 위를 차지한 고성능 시스템이다. 그림 1 은 타키온 시스템의 로긴 통계를 보여준다. 타키온 시스템은 한 달 평균 로긴 건수가 5,266 건으로, 응용 과학기술 분야를 연구 자원으로 사용되고 있는 대규모 시스템이다.

표 1. Tachyon 시스템 사양

구 분	내 용
모델명	SUN Blade 6048
블레이드 노드	188 개 24TFlops(Rpeak)
CPU	AMD Opteron 2.0GHz 16 개(노드당) 3,008(전체)
메모리	32GB(노드)
스토리지	207TB(디스크) 422TB(테이프)

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

	노드 간 네트워크	Infiniband 4X DDR

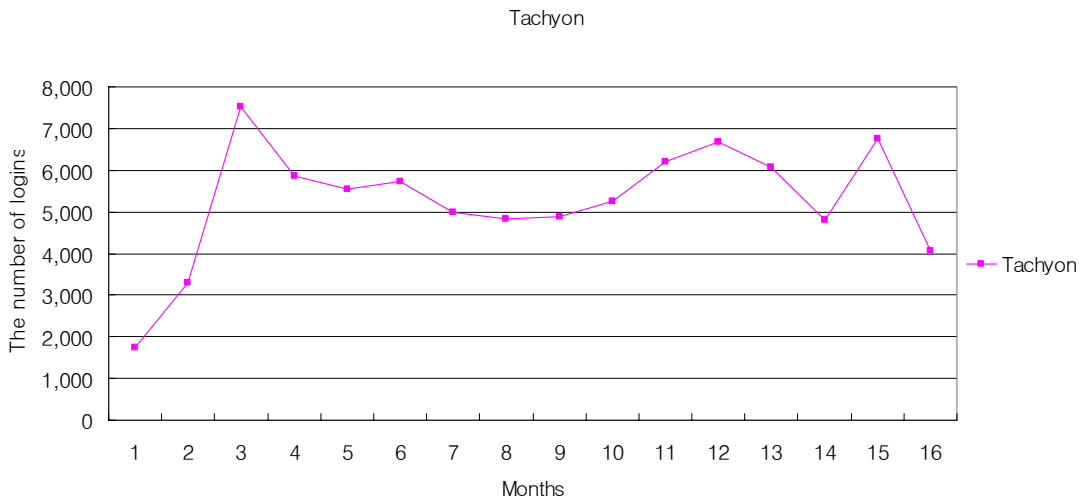

그림 1. 시스템의 로긴 통계

타키온의 스토리지 시스템을 살펴보면, 스크래치와 홈 파일 시스템을 위해 19 대의 SUN X4500 시스템을 연동하여 스토리지 시스템을 구성하였다. 이를 통해 11TB 용량의 홈 디렉토리(/home01)와 두 개의 글로벌 스크래치 디랙토리(54TB 의 /work01, 40TB 의 /work02 를 제공하고 있다. 본 연구에서는 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. 스크래치 및 홈 디렉토리는 Lustre 병렬 분산 파일시스템을 통해 계산 노드를 비롯한 모든 노드에 서비스되고 있다.

3.2 사용자 환경

본 연구의 분석 대상인 타키온 클러스터 시스템은 주로 대학, 정부출연연구소, 정부기관, 산업체 등의 기관을 대상으로 대기, 환경, 물리, 화학, 생명 등의 다양한 과학기술 분야의 연구 주제에 대하여 서비스를 실시하고 있다. 그림 2 에 2009 년 8 월에서 2009 년 7 월까지 타키온 시스템의 CPU 사용률 분포를 나타내었다. 물리가 전체 사용률이 32.5%로 가장 높았으며, 화학 26.1%, 기계 21.2%, 전기/전자 8.3%, 대기/환경 8,3%순으로 높은 사용률을 보였다.

워크로드 분석 대상은 사용자의 기본 환경 설정 파일들과 사용자 소스 데이터들을 포함하는 사용자 홈 디렉토리(/home)와 사용자들이 대용량의 I/O 작업을 실행할 때 주로 사용하는 임시 작업 디렉토리(/ytmp)인 스크래치(scratch)를 대상으로 조사, 분석하였다.

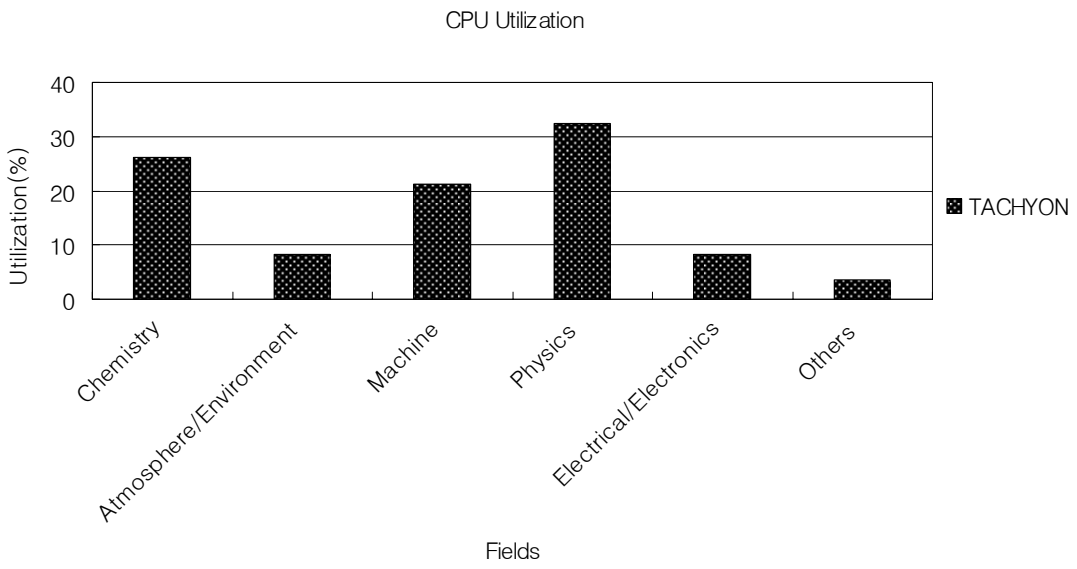

그림 2. 시스템 CPU 사용률

제 4 장 워크로드 분석

4.1 파일 종류별 분포

본 절에는 파일 종류에 대한 분석 결과를 정리한다. 그림 3 은 사용자 홈(home)과 스크래치(scarch) 디렉토리에 대해서, 전체 파일을 디렉토리, 실행파일, ".bash"로 시작하는 설정 파일, 숨김 파일, 심볼릭 링크 파일 등 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.

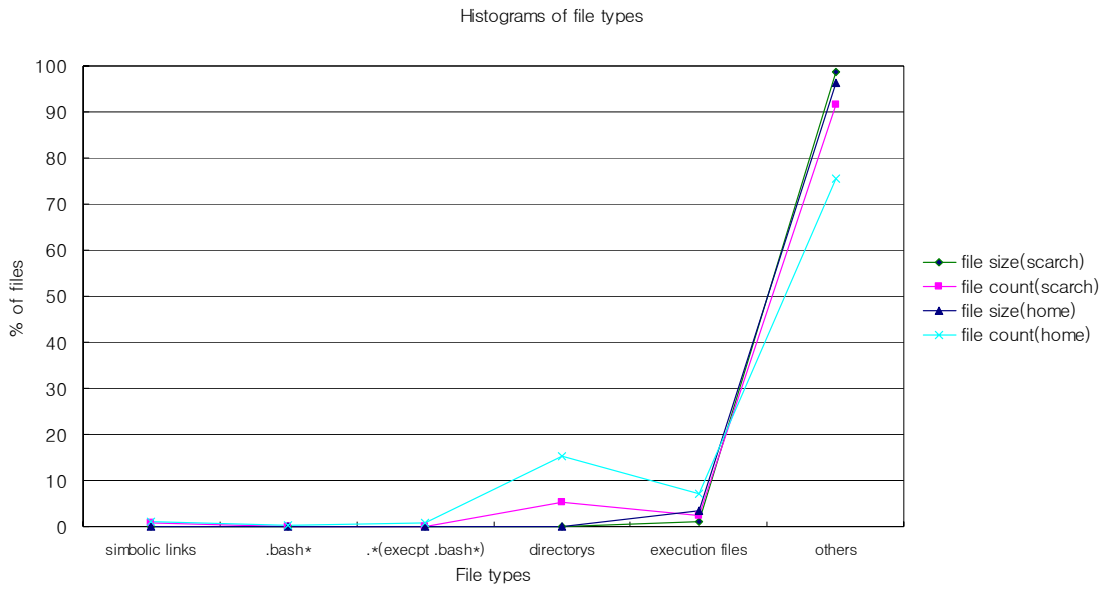

그림 3. 파일 종류별 분포

4.2 파일 크기

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.

타키온 클러스터 시스템의 스크래치(scratch) 공간에 대해 파일 크기에 따른 분포를 파일 수와 디스크 스페이스를 기준으로 분석한 결과를 그림 4,5 에 나타내었다. 파일 수 기준으로 볼 때, 1MB 이하 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.

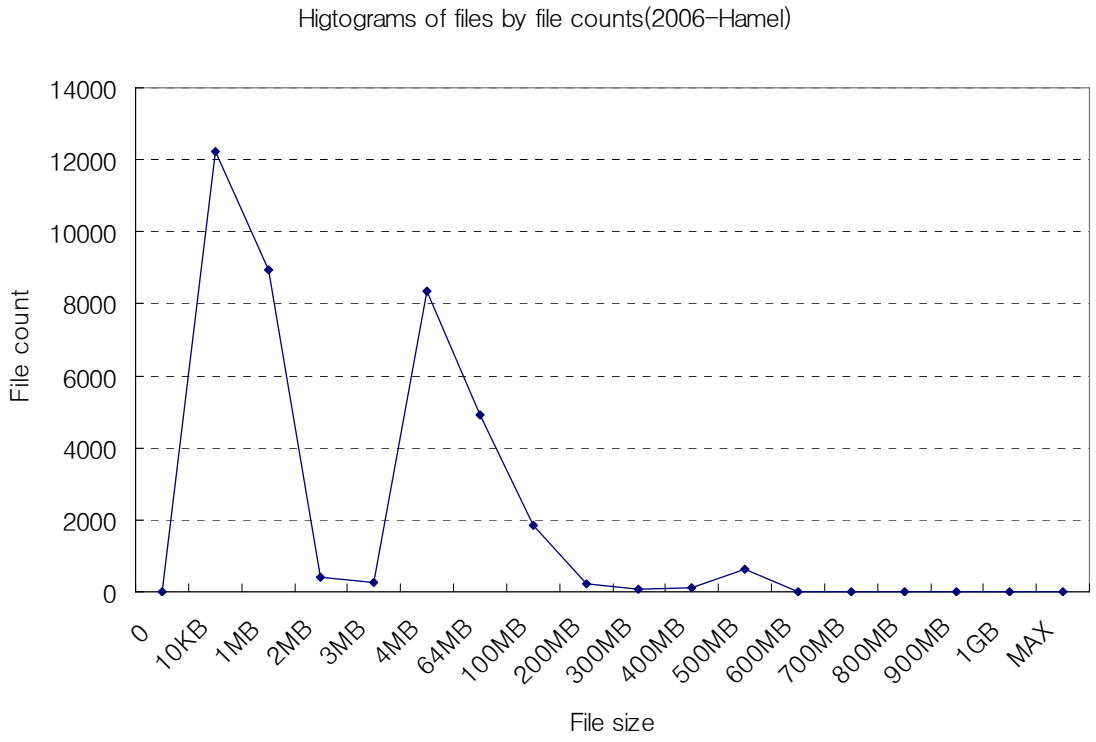

그림 4. 파일 크기별 분포(파일 수 기준)

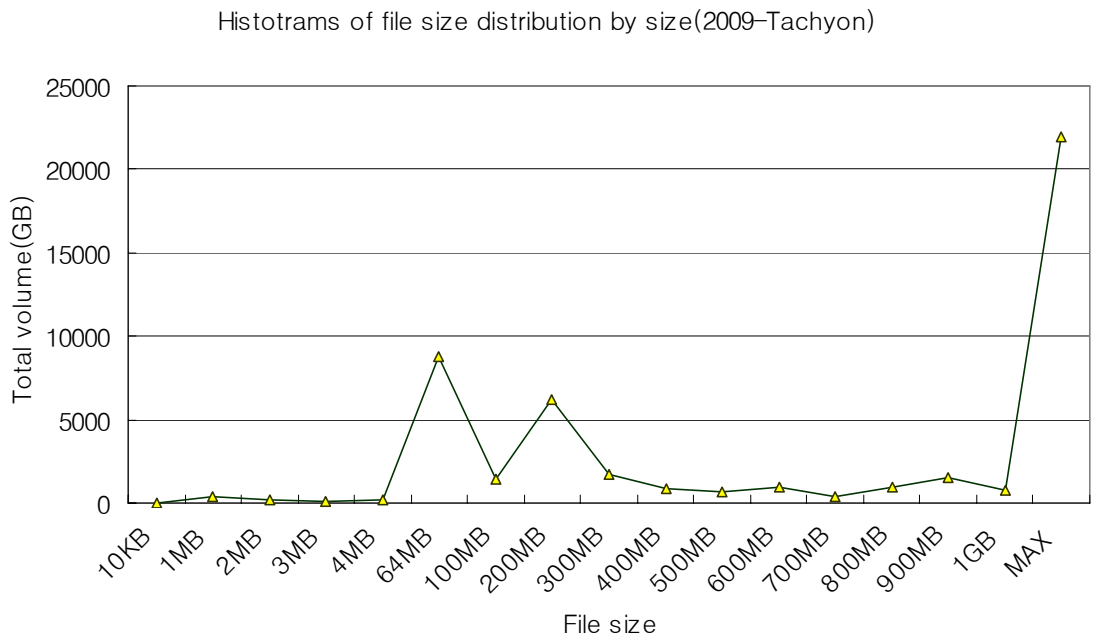

그림 5. 파일 크기별 분포(파일 크기 기준)

4.3 디렉토리당 파일 수

디렉토리당 소유하고 있는 파일의 전체 분포를 분석하기 위해서 디렉토리당 파일 개수의 누적 분포를 그림 6 에 나타내었다. 10 이하의 파일을 가지고 있는 디렉토리의 수가 스크래치 73.35%, 홈 디렉토리 77.53%를 차지하였다. 평균 디렉토리당 파일 수는 홈 16.85, 스크래치 46.15 였으며, 최대 파일 수는 홈 20,748, 스크래치 153,497 이었다.

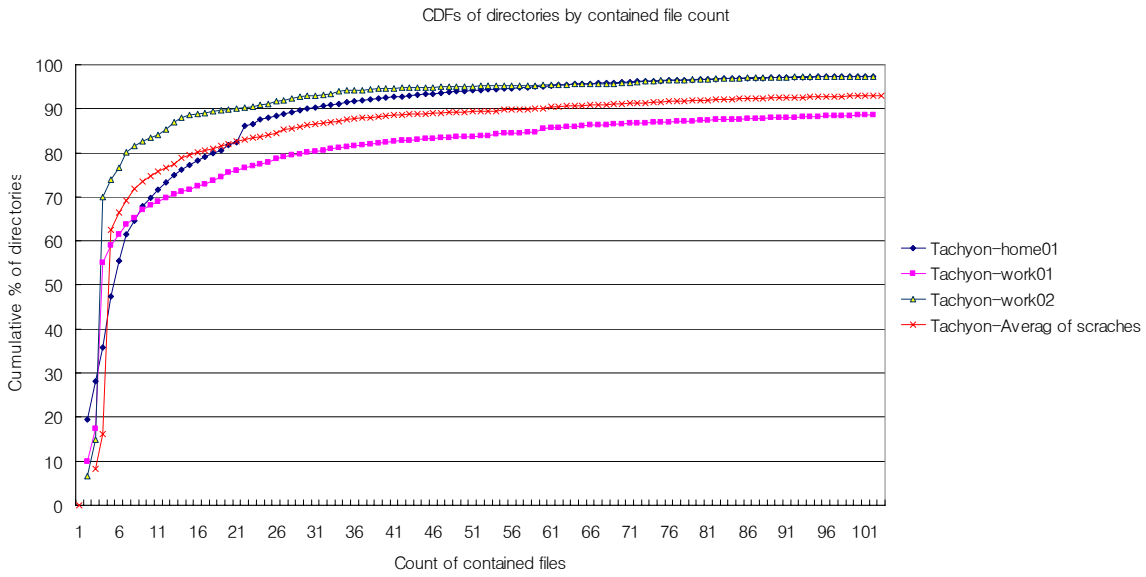

그림 6. 디렉토리당 파일 수 누적 분포

4.4 디렉토리 Depth

디렉토리 Depth 란 네임스페이스 트리(namespace tree)의 root(/)에서부터의 깊이를 나타낸다. 예를 들면, 홈 스토리지의 경우 /home 디렉토리의 디렉토리 Depth 는 0 이다. /home/hycho 디렉토리의 경우, 디렉토리 Depth 는 1 이다. 그림 7 와 8 에 디렉토리 Depth 에 대한 통계를 나타내었다. 분석한 결과, 90%의 디렉토리가 9 이하(/home01),

11 이하(/work01), 7 이하(/work02)의 디렉토리 Depth 를 가졌다. 대부분의 디렉토리 Depth 가 10 이하였으며, 평균 디렉토리 Depth 는 6.71(/home01), 7.41(/work01), 6.49(/work02)였다.

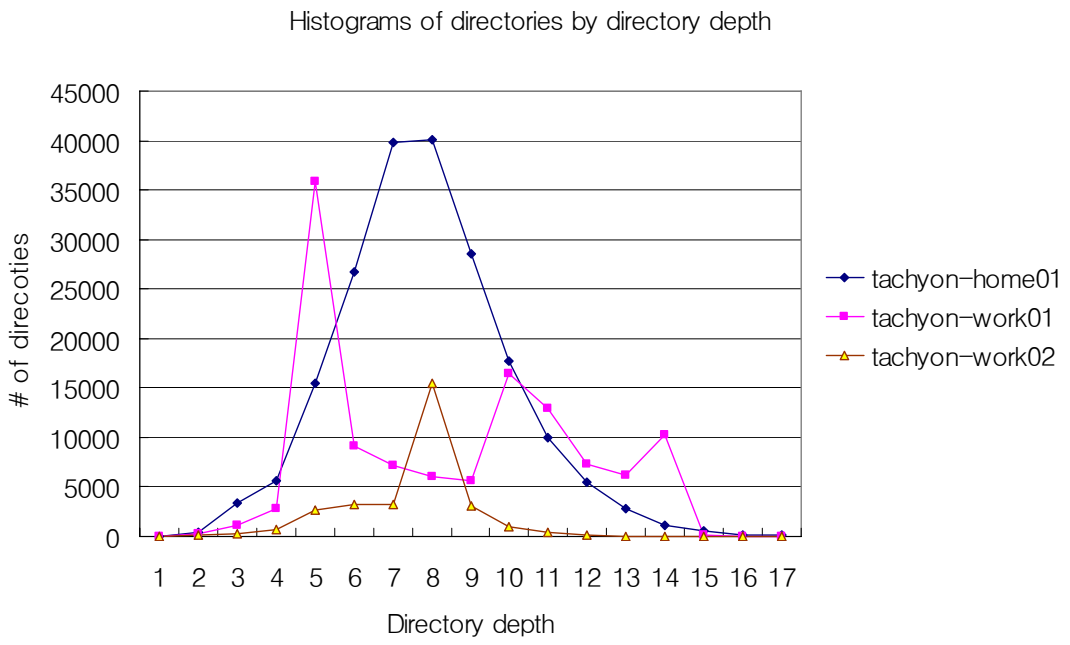

그림 7. 디렉토리 Depth 분포

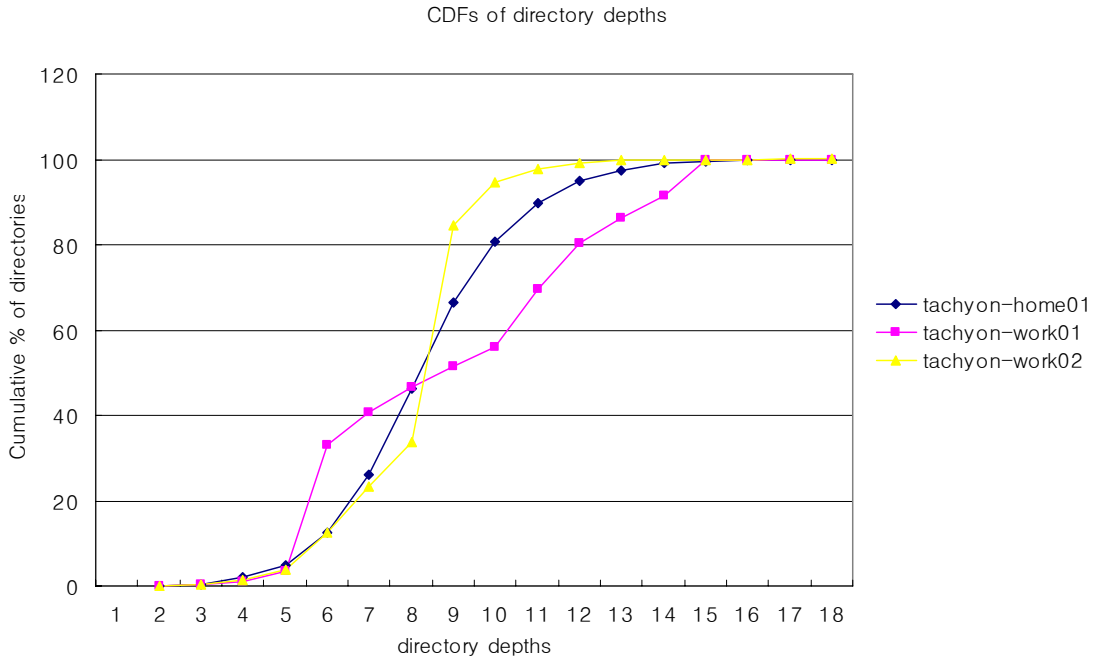

그림 8. 디렉토리 Depth 누적 분포

4.5 전체 데이터량 변화

클러스터 시스템의 데이터량의 변화를 예측하는 것은 클러스터를 구축하고 관리하는데 있어서 중요하다. 그림 9 에 2009 년 1 월부터 2009 년 6 월 15 일까지 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.

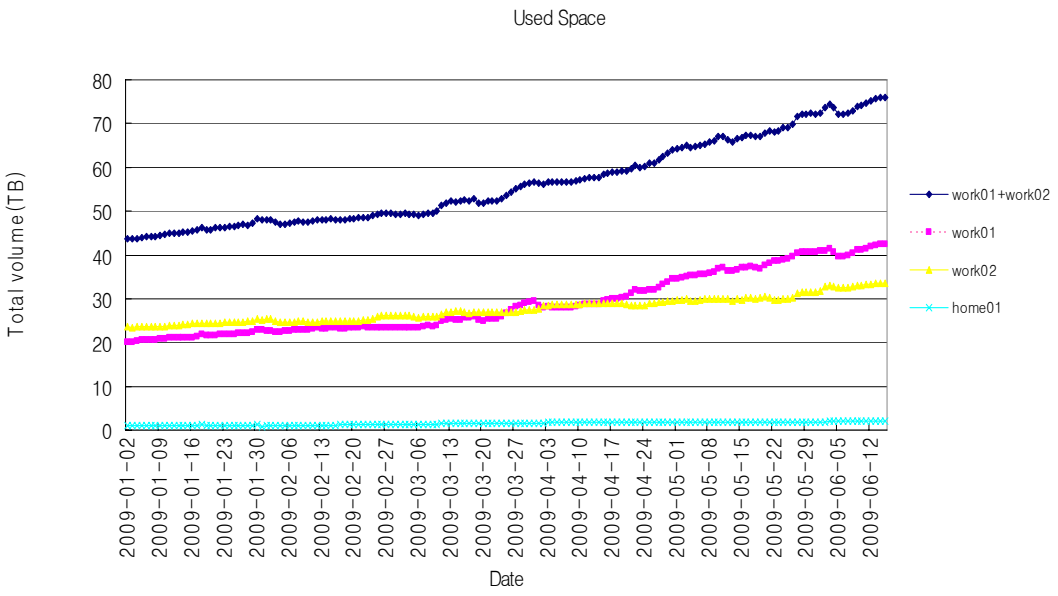
Used Space
Total volume(TB)
80
70
60
50
40
30
20
10
0
2009-01-02
2009-01-09
2009-01-16
2009-01-23
2009-01-30
2009-02-06
2009-02-13
2009-02-20
2009-02-27
2009-03-06
2009-03-13
2009-03-20
2009-03-27
2009-04-03
2009-04-10
2009-04-17
2009-04-24
2009-05-01
2009-05-08
2009-05-15
2009-05-22
2009-05-29
2009-06-05
2009-06-12
Date
work01+work02
work01
work02
home01

그림 9. 용량 변화 추이

제 5 장 워크로드 분석 결과 및 향후 계획

클러스터 시스템에서 파일 시스템에 대한 워크로드 분석은 새로운 시스템의 설계 및 운영에 있어 중요한 자료이다. 이에 본 연구는 가동중인 188 노드/ 3008 코어 규모의 대규모 클러스터 파일 시스템에서 사용자 홈 디렉토리와 I/O 집중적인 프로그램을 수행할 때 주고 사용하는 임시 디렉토리인 스크래치를 대상으로 파일의 종류별 분포, 파일 크기, 디렉토리당 파일 수, 디렉토리 Depth, 전체 데이터 변화량 등을 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. 주로 많이 사용하고 있는 파일 크기는 파일 수 기준으로 분석할 때, 1MB 이하가 전체의 83%를 차지하였고, 1-64MB 크기의 파일도 6%를 차지하였다. 용량 기준으로 분석할 경우, 4-64MB 가 19%, 100-200MB 가 13%를 차지하였다. 디렉토리당 파일 수는 70%이상의 디렉토리가 10 이하의 파일을 포함하고 있었으며, 평균 디렉토리 Depth 는 6-7 정도였다. 전체 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 연구의 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 연구의 워크로드 분석은 전체 데이터의 변화량 분석으로부터 일부 동적인 데이터의 변화를 하루 단위로 보였으나, 대부분은 정적(static) 정보들을 분석한 결과여서 그 한계가 있다. 이에 향후 Access 패턴이나 빈도 등에 관한 동적(dynamic) 워크로드 분석에 관한 연구가 필요할 것으로 보인다.

제 6 장 워크로드 분석 상세 데이터

1. Data file size & Data scale

A. 개요

현재 사용자가 사용하고 있는 파일 크기에 대한 통계

◆ 분석 대상

- /home2 : user home

- /work01 : scratch 로 사용되는 temporary working directory

- /work02: scratch 로 사용되는 temporary working directory

B. 현재 Tachyon 시스템 File System 상태

```
1032% [root@tachyon192 20090630]# date
Fri Jul   3 15:26:15 KST 2009
1031% [root@tachyon192 20090630]# df -h
Filesystem              Size   Used Avail Use% Mounted on
/dev/hda2               7.3G   5.4G   1.5G   79% /
/dev/hda1               122M    13M   103M   12% /boot
none                     16G      0    16G    0% /dev/shm
none                    198M      0   198M    0% /tmp
none                     10M      0    10M    0% /var/tmp
none                     10M   244K   9.8M    3% /var/log
/dev/ram0                16M    15M      0 100% /swap
mds01-ib0@o2ib:mds02-ib0@o2ib:/applic
                        3.5T   192G   3.1T    6% /applic
mds01-ib0@o2ib:mds02-ib0@o2ib:/lustre1
                         63T    27T    33T   45% /work01
mds01-ib0@o2ib:mds02-ib0@o2ib:/lustre2
                         54T    14T    38T   28% /work02
mds03-ib0@o2ib:mds04-ib0@o2ib:/home01
                         11T   2.3T   8.0T   23% /home01
```

타키온 시스템은 /home01, /work01, /work02, 이 네트워크 파일 시스템으로 Lustre 파일 시스템을 사용하고 있다. 2009 년 7 월 13 일 시점에, /home01 은 11TB 중 2.3TB 를 사용 23%를 사용하고 있었고, /work01 은 63TB 중 27TB(45%)를, /work02 는 54TB 중 14TB(28%)를 사용하고 있었다.

C. File Type & File Size

```
=================================================================
  1. The analysis of Directory ( /home01 )      Unit: # of files
=================================================================
# of file                    :          3737672(      100.00000 %)
# of current/parents dir     :           380260(       10.17371 %)
# of Symbolic Link           :            43055(        1.15192 %)
# of .bash*                  :             6593(        0.17639 %)
# of .* (except .bash*)      :            28181(        0.75397 %)
# of directory               :           190129(        5.08683 %)
# of execute file            :           269618(        7.21353 %)
# of real sample data        :          2819836(       75.44365 %)
=================================================================

=================================================================
  2. The analysis of Directory ( /home01 )      Unit: the size of files(byte)
=================================================================
# of file                    :     2487734647287(      100.00000 %)
# of current/parents dir     :        1720713216(        0.06917 %)
# of Symbolic Link           :           1535631(        0.00006 %)
# of .bash*                  :          14018706(        0.00056 %)
# of .* (except .bash*)      :         270776040(        0.01088 %)
# of directory               :         828903424(        0.03332 %)
# of execute file            :       85688868891(        3.44445 %)
# of real sample data        :     2399209831379(       96.44155 %)
=================================================================
```

```
=====================================================================
  3.Sample Data : real sample data            Unit: # of files
=====================================================================
 0    -   10K :                  1897944(          67.30689 %)
10K   -   1M  :                   843698(          29.92011 %)
 1M   -   2M  :                    21271(           0.75433 %)
 2M   -   3M  :                     5771(           0.20466 %)
 3M   -   4M  :                     3339(           0.11841 %)
 4M   -   5M  :                     1886(           0.06688 %)
 5M   -   6M  :                     1703(           0.06039 %)
 6M   -   7M  :                     1759(           0.06238 %)
 7M   -   8M  :                     1367(           0.04848 %)
 8M   -   9M  :                     3390(           0.12022 %)
 9M   - 10M   :                     1198(           0.04248 %)
10M   - 11M   :                     1686(           0.05979 %)
11M   - 12M   :                     1926(           0.06830 %)
12M   - 13M   :                     1329(           0.04713 %)
13M   - 14M   :                     1944(           0.06894 %)
14M   - 15M   :                     1542(           0.05468 %)
15M   - 16M   :                     2112(           0.07490 %)
16M   - 17M   :                     2091(           0.07415 %)
17M   - 18M   :                     1695(           0.06011 %)
18M   - 19M   :                     1037(           0.03678 %)
19M   - 20M   :                     1151(           0.04082 %)
20M   - 21M   :                     6651(           0.23586 %)
21M   - 22M   :                      691(           0.02450 %)
22M   - 23M   :                      937(           0.03323 %)
23M   - 24M   :                      678(           0.02404 %)
24M   - 25M   :                      447(           0.01585 %)
25M   - 26M   :                     1160(           0.04114 %)
26M   - 27M   :                     1080(           0.03830 %)
27M   - 28M   :                      906(           0.03213 %)
28M   - 29M   :                      414(           0.01468 %)
29M   - 30M   :                      633(           0.02245 %)
30M   - 31M   :                      359(           0.01273 %)
```

31M - 32M :	221(	0.00784 %)
32M - 33M :	211(	0.00748 %)
33M - 34M :	254(	0.00901 %)
34M - 35M :	189(	0.00670 %)
35M - 36M :	505(	0.01791 %)
36M - 37M :	111(	0.00394 %)
37M - 38M :	700(	0.02482 %)
38M - 39M :	101(	0.00358 %)
39M - 40M :	44(	0.00156 %)
40M - 41M :	54(	0.00192 %)
41M - 42M :	50(	0.00177 %)
42M - 43M :	46(	0.00163 %)
43M - 44M :	16(	0.00057 %)
44M - 45M :	61(	0.00216 %)
45M - 46M :	50(	0.00177 %)
46M - 47M :	44(	0.00156 %)
47M - 48M :	13(	0.00046 %)
48M - 49M :	232(	0.00823 %)
49M - 50M :	41(	0.00145 %)
50M - 51M :	7(	0.00025 %)
51M - 52M :	20(	0.00071 %)
52M - 53M :	20(	0.00071 %)
53M - 54M :	19(	0.00067 %)
54M - 55M :	32(	0.00113 %)
55M - 56M :	28(	0.00099 %)
56M - 57M :	49(	0.00174 %)
57M - 58M :	11(	0.00039 %)
58M - 59M :	162(	0.00575 %)
59M - 60M :	14(	0.00050 %)
60M - 61M :	13(	0.00046 %)
61M - 62M :	30(	0.00106 %)
62M - 63M :	2(	0.00007 %)
63M - 64M :	9(	0.00032 %)
64M - 1G :	2521(	0.08940 %)
1G - MAX :	161(	0.00571 %)

```
=====================================================================
NUM_FILE      :                2819836(        100.00000 %)
=====================================================================


=====================================================================
  4.Sample Data : real sample data          Unit: the size of files(byte)
=====================================================================
 0   -   10K  :            5180799128(          0.21594 %)
10K -    1M  :           76513680526(          3.18912 %)
 1M   -   2M  :           31234052624(          1.30185 %)
 2M   -   3M  :           15753860057(          0.65663 %)
 3M   -   4M  :           11979057545(          0.49929 %)
 4M   -   5M  :            9098333680(          0.37922 %)
 5M   -   6M  :           10042954266(          0.41859 %)
 6M   -   7M  :           11993116543(          0.49988 %)
 7M   -   8M  :           10419009162(          0.43427 %)
 8M   -   9M  :           29838810691(          1.24369 %)
 9M   - 10M   :           11961768866(          0.49857 %)
10M   - 11M   :           18616813625(          0.77596 %)
11M   - 12M   :           23351747251(          0.97331 %)
12M   - 13M   :           17448039592(          0.72724 %)
13M   - 14M   :           27269994687(          1.13662 %)
14M   - 15M   :           23452327224(          0.97750 %)
15M   - 16M   :           34291118053(          1.42927 %)
16M   - 17M   :           36371590645(          1.51598 %)
17M   - 18M   :           31309955037(          1.30501 %)
18M   - 19M   :           20012406938(          0.83412 %)
19M   - 20M   :           23515934571(          0.98015 %)
20M   - 21M   :          143720237903(          5.99032 %)
21M   - 22M   :           15622332081(          0.65114 %)
22M   - 23M   :           21987450521(          0.91645 %)
23M   - 24M   :           16775250870(          0.69920 %)
24M   - 25M   :           11437992016(          0.47674 %)
25M   - 26M   :           30779447302(          1.28290 %)
26M   - 27M   :           29991704066(          1.25007 %)
```

27M	- 28M :	26248574096(	1.09405 %)
28M	- 29M :	12368183838(	0.51551 %)
29M	- 30M :	19566971093(	0.81556 %)
30M	- 31M :	11466310244(	0.47792 %)
31M	- 32M :	7329107615(	0.30548 %)
32M	- 33M :	7192094339(	0.29977 %)
33M	- 34M :	8926641950(	0.37207 %)
34M	- 35M :	6830340491(	0.28469 %)
35M	- 36M :	18896906907(	0.78763 %)
36M	- 37M :	4241142146(	0.17677 %)
37M	- 38M :	27759367833(	1.15702 %)
38M	- 39M :	4087057611(	0.17035 %)
39M	- 40M :	1821237920(	0.07591 %)
40M	- 41M :	2293509273(	0.09559 %)
41M	- 42M :	2173119865(	0.09058 %)
42M	- 43M :	2057562377(	0.08576 %)
43M	- 44M :	730106852(	0.03043 %)
44M	- 45M :	2836164636(	0.11821 %)
45M	- 46M :	2381277778(	0.09925 %)
46M	- 47M :	2146656744(	0.08947 %)
47M	- 48M :	648459935(	0.02703 %)
48M	- 49M :	11800142829(	0.49183 %)
49M	- 50M :	2118216858(	0.08829 %)
50M	- 51M :	371595878(	0.01549 %)
51M	- 52M :	1078271308(	0.04494 %)
52M	- 53M :	1101024477(	0.04589 %)
53M	- 54M :	1064939314(	0.04439 %)
54M	- 55M :	1827072618(	0.07615 %)
55M	- 56M :	1630836247(	0.06797 %)
56M	- 57M :	2898939903(	0.12083 %)
57M	- 58M :	663126949(	0.02764 %)
58M	- 59M :	9901994921(	0.41272 %)
59M	- 60M :	874221836(	0.03644 %)
60M	- 61M :	824174077(	0.03435 %)
61M	- 62M :	1929731016(	0.08043 %)

```
62M  - 63M  :              131068428(          0.00546 %)
63M  - 64M  :              599308884(          0.02498 %)
64M  -  1G  :           532134858830(         22.17959 %)
1G  -  MAX  :           906289727993(         37.77451 %)
=====================================================================
NUM_FILE     :                2819836(        100.00000 %)
TOTAL_SIZE   :          2399209831379(     byte      )
MAX_SIZE     :            98221178880(     byte      )
=====================================================================


=====================================================================
  5.Sample Data : 0 -        1024 (Kbyte), interval :      102.40( Kbyte )
    (byte)  - (byte)             # of file
=====================================================================
            0 ~        104857 :              2552326(      93.09480 %)
       104858 ~        209715 :                81394(       2.96880 %)
       209716 ~        314573 :                50410(       1.83868 %)
       314574 ~        419431 :                23363(       0.85215 %)
       419432 ~        524289 :                12149(       0.44313 %)
       524290 ~        629147 :                 7292(       0.26597 %)
       629148 ~        734005 :                 5088(       0.18558 %)
       734006 ~        838863 :                 4339(       0.15826 %)
       838864 ~        943721 :                 2971(       0.10837 %)
       943722 ~       1048579 :                 2310(       0.08426 %)
         NUM_TOTAL            :              2741642(     100.00000 %)
=====================================================================


=====================================================================
  6.Sample Data : 1048576 - 1073741824 (byte), interval :
107269325.00( byte )
  6.Sample Data : 1024 - 1048576
(Kbyte), interval : 104755.20( Kbyte )
  6.Sample Data :          1 -        1024 (Mbyte), interval :      102.30( Mbyte )
    (byte)  - (byte)             the size of file(byte)
=====================================================================
```

```
          0 ~      107269324 :          908456269527(       64.37979 %)
  107269325 ~      214538649 :          236249844515(       16.74237 %)
  214538650 ~      321807974 :           66960011720(        4.74527 %)
  321807975 ~      429077299 :           17407962755(        1.23365 %)
  429077300 ~      536346624 :          125454005827(        8.89058 %)
  536346625 ~      643615949 :            7588674800(        0.53779 %)
  643615950 ~      750885274 :           18060142104(        1.27987 %)
  750885275 ~      858154599 :            7174248798(        0.50842 %)
  858154600 ~      965423924 :            9185151846(        0.65093 %)
  965423925 ~     1072693249 :           14552996960(        1.03133 %)
        NUM_TOTAL             :         1411089308852(      100.00000 %)
=====================================================================
=====================================================================
  5.Sample Data : 0 - MAX,            # of file(byte)
=====================================================================
 000M  -  100M  :                  2817516(          99.91773 %)
 100M  -  200M  :                     1518(           0.05383 %)
 200M  -  300M  :                      280(           0.00993 %)
 300M  -  400M  :                       51(           0.00181 %)
 400M  -  500M  :                       31(           0.00110 %)
 500M  -  600M  :                      220(           0.00780 %)
 600M  -  700M  :                       21(           0.00074 %)
 700M  -  800M  :                       13(           0.00046 %)
 800M  -  900M  :                        8(           0.00028 %)
  900M  -  1G   :                      17(           0.00060 %)
    1G  -  MAX  :                     161(           0.00571 %)
    NUM_TOTAL   :                  2819836(         100.00000 %)
=====================================================================
  6.Sample Data : 0 - MAX,     the size of file(byte)
=====================================================================
 000M  -  100M  :            989651612115(          41.24906 %)
 100M  -  200M  :            236245347943(           9.84680 %)
 200M  -  300M  :             66326820559(           2.76453 %)
 300M  -  400M  :             18681102434(           0.77864 %)
 400M  -  500M  :             14689067587(           0.61225 %)
```

```
 500M  -  600M  :          118353613040(          4.93302 %)
 600M  -  700M  :           14353971020(          0.59828 %)
 700M  -  800M  :           10031274650(          0.41811 %)
 800M  -  900M  :            7173657502(          0.29900 %)
  900M  -  1G   :           17413636536(          0.72581 %)
    1G  -  MAX  :          906289727993(         37.77451 %)
    SIZE_TOTAL  :         2399209831379(        100.00000 %)
=================================================================
```

위의 결과에서 1.The analysis of files 는 전체 파일을 디렉토리, 실행파일, .bash* 파일, 숨김 파일 심볼릭 링크 파일 등의 종류에 따라 통계를 낸 것이다. 2 이후는 The analysis of files 의 결과에서 마지막 real sample data 를 대상으로 파일 크기별로 분포를 파일 수, 파일 사이즈 기준으로 통계 구간은 넓은 범위에서 좁은 범위로 좁혀가면서 조사한 것이다.

D. Directory Depth

```
=================================================================
  1. The analysis of Files ( 20090710-scanDir-home01.result )
=================================================================
# of Data Set(directory)      :                193939(        100.00000 %)
TOTAL_FILE                    :                 3442069
MAX(# of File(each directory)) :                    9952
MIN(# of File(each directory)) :                       0
Average(# of File(each directory)):            17.74820
=================================================================


=================================================================
  2.Sample Data : the depth of directory
 Depth      :              # of directory
=================================================================
     0      :                19552(        10.08152)
     1      :                17588(         9.06883)
     2      :                16303(         8.40625)
     3      :                26564(        13.69709)
```

```
     4        :                    18117(          9.34160)
     5        :                    16137(          8.32066)
     6        :                     7196(          3.71045)
     7        :                     9528(          4.91288)
     8        :                     4548(          2.34507)
     9        :                     4314(          2.22441)
    10        :                     4219(          2.17543)
    11        :                     3534(          1.82222)
    12        :                     2789(          1.43808)
    13        :                     2410(          1.24266)
    14        :                     2249(          1.15964)
    15        :                     2088(          1.07663)
=====================================================================
 Total Depth   = 676789,         # of directory = 193939
 avg of Depth = 3.489700
=====================================================================
  3.Sample Data : 0 -          100(개 미만 ), interval :          10(개)
       # of file      :                  # of directory
=====================================================================
            0 ~            9 :                139847(      74.24059 %)
           10 ~           19 :                 24055(      12.77008 %)
           20 ~           29 :                  9614(       5.10379 %)
           30 ~           39 :                  5105(       2.71009 %)
           40 ~           49 :                  2955(       1.56872 %)
           50 ~           59 :                  2287(       1.21410 %)
           60 ~           69 :                  1799(       0.95504 %)
           70 ~           79 :                  1265(       0.67155 %)
           80 ~           89 :                   838(       0.44487 %)
           90 ~           99 :                   605(       0.32118 %)
        NUM_TOTAL            :                188370(     100.00000 %)
=====================================================================
```

E. Directory Set Size

```
[hycho@tachyon189 test3]$ dataSetsize ./20090713-scanDir-work01.result
File open  : ./20090713-scanDir-work01.result
=====================================================================
  1. The analysis of Files ( ./20090713-scanDir-work01.result )
=====================================================================
# of Data Set(directory)      :                5374280(        100.00000 %)
TOTAL_FILE                    :                12376656
MAX(# of File(each directory)) :                47397
MIN(# of File(each directory)) :                0
Average(# of File(each directory)):             2.30294
=====================================================================


=====================================================================
  2.Sample Data : # of Data set(directory)
# of file            :            # of directory
=====================================================================
0       -  100  :                5360209(        99.73818 %)
100     -  500  :                9363(           0.17422 %)
500     - 1000  :                3941(           0.07333 %)
1000    -  MAX  :                767(            0.01427 %)
=====================================================================


=====================================================================
  3.Sample Data : 0 -        100(개 미만 ), interval :        10(개)
      # of file     :                # of directory
=====================================================================
          0 ~          9 :          5337116(      99.56918 %)
         10 ~         19 :             8118(       0.15145 %)
         20 ~         29 :             5331(       0.09946 %)
```

```
        30 ~        39 :        2614(      0.04877 %)
        40 ~        49 :        1287(      0.02401 %)
        50 ~        59 :        2295(      0.04282 %)
        60 ~        69 :        1190(      0.02220 %)
        70 ~        79 :         899(      0.01677 %)
        80 ~        89 :         707(      0.01319 %)
        90 ~        99 :         652(      0.01216 %)
      NUM_TOTAL          :     5360209(     100.00000 %)
=================================================================
```

F. Server 별 Operation 모니터링

Lustre 파일 시스템의 경우 /proc 데이터를 통해 operation 수준으로 모니터링 기능을 제공한다. 그러나 현재 티키온 시스템은 모니터링 옵션이 가동되어 있지 않다. 각 클라이언트 별까지는 아니더라도 MDS 나 OST 서버에서 모니터링하여 데이터를 모니터링하고 분석하는 것은 슈퍼컴퓨터 본부의 사용자 I/O 특성을 분석하는데 의미 있다고 여기어진다. 향후 5 호기의 파일시스템, 사용자 I/O 관련 백데이터를 축척하기 위해서도 사용자 I/O 워크로드 분석에 관한 연구에 지속적인 지원이 필요하다.

```
[hycho@tachyon192 home01-OST0000-osc-0000010470ed6400]$ llstat -i 1 /proc/fs/lustre/osc/lustre1-OST0001-osc-0000010470c92c00/
sh: lctl: command not found
/usr/bin/llstat: STATS on 07/15/09 /proc/fs/lustre/osc/lustre1-OST0001-osc-0000010470c92c00//stats on
snapshot_time          1247649281.916203
req_waittime           1532750
req_active             1532750
ost_read               362
ost_write              7712
ost_destroy            146178
```

```
ost_connect          1
ost_punch            7949
ost_statfs           637
ldlm_cancel          93917
obd_ping             87446
/proc/fs/lustre/osc/lustre1-OST0001-osc-0000010470c92c00//stats @ 1247649282.918648
Name                 Cur.Count  Cur.Rate   #Events   Unit      last      min       avg       max     stddev
req_waittime         0          0          1532750   [usec]    0         51        1778.58   23396145  37165.65
req_active           0          0          1532750   [reqs]    0         1         1.05      10        0.36
ost_read             0          0          362       [usec]    0         4096      108272.44 356352    37744.73
ost_write            0          0          7712      [usec]    0         6         141270.97 1048576   312513.87
ost_destroy          0          0          146178    [usec]    0         0         0.00      0         0.00
ost_connect          0          0          1         [usec]    0         0         0.00      0         0.00
ost_punch            0          0          7949      [usec]    0         0         0.00      0         0.00
ost_statfs           0          0          637       [usec]    0         0         0.00      0         0.00
ldlm_cancel          0          0          93917     [usec]    0         0         0.00      0         0.00
obd_ping             0          0          87446     [usec]    0         0         0.00      0         0.00
/proc/fs/lustre/osc/lustre1-OST0001-osc-0000010470c92c00//stats @ 1247649283.921623
Name                 Cur.Count  Cur.Rate   #Events   Unit      last      min       avg       max     stddev
req_waittime         0          0          1532750   [usec]    0         51        1778.58   23396145  37165.65
req_active           0          0          1532750   [reqs]    0         1         1.05      10        0.36
ost_read             0          0          362       [usec]    0         4096      108272.44 356352    37744.73
ost_write            0          0          7712      [usec]    0         6         141270.97 1048576   312513.87
ost_destroy          0          0          146178    [usec]    0         0         0.00      0         0.00
ost_connect          0          0          1         [usec]    0         0         0.00      0         0.00
ost_punch            0          0          7949      [usec]    0         0         0.00      0         0.00
ost_statfs           0          0          637       [usec]    0         0         0.00      0         0.00
ldlm_cancel          0          0          93917     [usec]    0         0         0.00      0         0.00
obd_ping             0          0          87446     [usec]    0         0         0.00      0         0.00
/proc/fs/lustre/osc/lustre1-OST0001-osc-0000010470c92c00//stats @ 1247649284.924601
Name                 Cur.Count  Cur.Rate   #Events   Unit      last      min       avg       max     stddev
req_waittime         0          0          1532750   [usec]    0         51        1778.58   23396145  37165.65
req_active           0          0          1532750   [reqs]    0         1         1.05      10        0.36
ost_read             0          0          362       [usec]    0         4096      108272.44 356352    37744.73
ost_write            0          0          7712      [usec]    0         6         141270.97 1048576   312513.87
```

```
ost_destroy          0    0    146178    [usec]    0    0    0.00    0    0.00
ost_connect          0    0    1         [usec]    0    0    0.00    0    0.00
ost_punch            0    0    7949      [usec]    0    0    0.00    0    0.00
ost_statfs           0    0    637       [usec]    0    0    0.00    0    0.00
ldlm_cancel          0    0    93917     [usec]    0    0    0.00    0    0.00
obd_ping             0    0    87446     [usec]    0    0    0.00    0    0.00
/proc/fs/lustre/osc/lustre1-OST0001-osc-0000010470c92c00//stats @ 1247649285.927600
Name                 Cur.Count  Cur.Rate  #Events  Unit  last  min  avg  max  stddev
req_waittime         0    0    1532750   [usec]    0    51    1778.58    23396145    37165.65
req_active           0    0    1532750   [reqs]    0    1    1.05    10    0.36
ost_read             0    0    362       [usec]    0    4096    108272.44    356352    37744.73
ost_write            0    0    7712      [usec]    0    6    141270.97    1048576    312513.87
ost_destroy          0    0    146178    [usec]    0    0    0.00    0    0.00
ost_connect          0    0    1         [usec]    0    0    0.00    0    0.00
ost_punch            0    0    7949      [usec]    0    0    0.00    0    0.00
ost_statfs           0    0    637       [usec]    0    0    0.00    0    0.00
ldlm_cancel          0    0    93917     [usec]    0    0    0.00    0    0.00
obd_ping             0    0    87446     [usec]    0    0    0.00    0    0.00
```

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," Proc. of the 10th Symposium on Operating 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.

타키온 클러스터 I/O 워크로드 분석 보고서

2009 년 12 월 인쇄

2009 년 12 월 발행

발행처: 한국과학기술정보연구원 Korea Institute of Science and Technology Information

대전시 유성구 과학로 335 번지

Tel: (042) 869-1004(代)

ISBN 978-89-6211-501-7 보급가 : 비매품