실험적인 방법으로 현대 디스크의 내부구조 추측

유영진, 신동인, 염헌영
서울대학교 전기컴퓨터 공학부
분산시스템 연구실
e-mail: {yjyu, dishin, yeom}@dcslab.snu.ac.kr

An Empirical Inspection of Modern Disk Drive Internals

Young-Jin Yu, Dong-In Shin, Heon-Young Yeom
Dept of Computer Science & Engineering, Seoul National University
Distributed Computing System Lab.

요 약

디스크는 내부의 정보를 최대한 숨기고 추상화하여 운영체제에 읽기와 쓰기같은 최소한의 인터페이스만을 제공한다. 결과적으로 상위 레이어의 소프트웨어는 디바이스에 대해 최소한의 가정만을 가지고 결정에 임할 수밖에 없으며 이는 여러가지 최적화에 걸림돌이 될 수 밖에 없다. 본 논문에서는 디스크가 제공하는 최소한의 인터페이스만을 가지고 내부 구조를 정확히 추측해 내는 기법을 소개한다. 기존에 SCSI 디스크에 대해 매핑 정보를 추출해내는 연구[1,2] 이미 존재했으나, 널리 사용되고 있는 ATA 디스크의 경우 이를 밝혀낸 논문은 알려진 바 없다. 이 논문에서는 ATA 뿐만 아니라 SCSI 디스크에서도 적용할 수 있는 더 빠르고 정확한 알고리즘을 제안하고, 실제 실험 결과를 제시하였다. 이러한 결과는 차후에 입출력 시스템을 최적화하는데 큰 도움을 줄 수 있을 것이라 여겨진다.

1. 서론

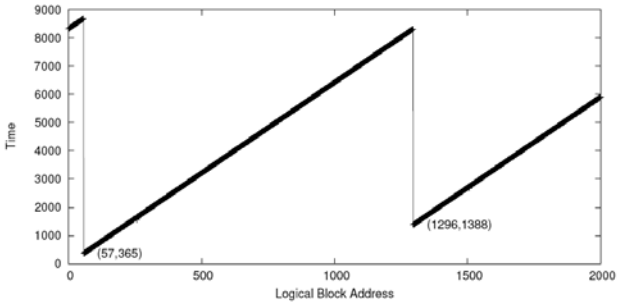
최근 디스크의 입출력 성능은 전체 시스템의 성능을 결정하는데 있어서 커다란 부분을 차지하게 되었다. 전자적인 신호로써 데이터가 입출력되는 메모리나 다른 플래시 기반의 저장 장치와는 다르게, 디스크는 헤드의 기계적인 움직임을 동반하게 되므로 접근 시간이 늦을 수 밖에 없다. 이러한 성능상의 단점을 극복하기 위해, 디스크의 동작 원리와 구조를 미리 알아내어 상위 소프트웨어에서 효율적으로 입출력 서비스를 제공하고자 하는 연구가 있어왔다.

하지만 디스크의 내부 정보는 외부에 숨겨져 있기 때문에 매우 제한적일 수 밖에 없으며, ATA 인터페이스를 가진 디스크의 경우에는 지원되는 커맨드셋이 훨씬 작아서 이마저도 쉽게 달성하기가 어렵다.

본 논문에서는 실험적인 방법으로 디스크의 내부 구조를 추측할 수 있는 방법을 소개한다. 우리가 아는 한, 이 작업은 ATA 디스크를 대상으로 하는 최초의 작업이다. 소개될 방법은 디스크의 Read/Write 인터페이스만을 사용하기 때문에 SCSI 디스크 등 ATA 이외의 디바이스에도 적용할 수 있다.

2. 측정 알고리즘

트랙 경계 측정 알고리즘: 디스크 스펙을 보면 벤더들이 제공하는 SPT(Sectors Per Track) 값들이 나와 있는데, 최근 디스크들은 zoning 기법[3] 으로 인하여 스핀들 바깥쪽의 트랙의 경우 더 많은 섹터들을 가지고 있게 된다. 따라서 디스크 블럭의 논리번호 0부터 시작하여 SPT가 어떻게 결정되는지, 즉 트랙의 경계를 추정해볼 필요가 있다.

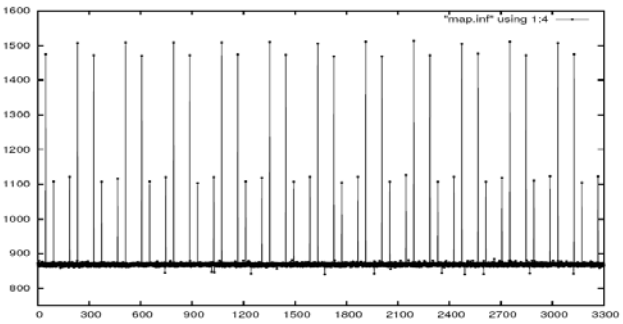


(그림 1) 리퀘스트간 딜레이 시간 분포

그림1은 Ganger의 논문[2]에서 언급했던 두 리퀘스트 간의 시간 분포에 대한 모습이다. 그 논문에서는 모든 섹터 거리에 대해 측정을 해야 하기 때문에 대단히 많은 시간이 걸리지만, 본 논문에서는 zone의 특징을 활용하여 단 몇번의 Read 리퀘스트 만으로 한 트랙의 SPT를 알아낼 수 있게 된다.

트랙 그룹간의 배치 추정: 이전에 언급한 SPT 를 모든 트랙에 대해 구하고 나면, LBA가 증가하는 순서대로 트랙들간에 번호를 매길 수가 있게 되는데 이를 여기서는 LTA(Logical Track Address)라고 한다. 이들 트랙이 서로간에 어떠한 순서로 배치되어 있는지를 알아내기 위해

서 이전 트랙과 현재 트랙 사이의 스위치 시간을 구할 필요가 있다. 그림2는 씨게이트의 ST3250820A 모델에서 추출한 인접 트랙 스위치 시간 분포 그래프를 나타낸 것이다. 그림에서 확인할 수 있듯이 이 그래프는 주기가 280인 함수 f(x) = f(x+280) 로 표현됨을 확인할 수 있는데, 이것은 곧 280개의 트랙이 하나의 배치 패턴을 이루고 이

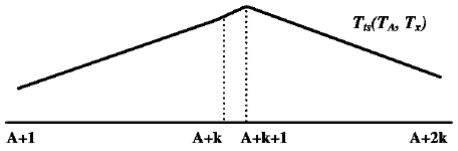


(그림 2) 인접 트랙 스위치 시간 분포 그래프

패턴이 연속적으로 배열되어 있음을 나타낸다. 다른 ATA 디스크 모델로 동일한 실험을 수행하여도 주기에만 차이가 있을뿐, 동일한 패턴이 연속적으로 나타남을 확인할 수 있었다. 이러한 하나의 패턴을 트랙 그룹이라고 정의를 하였다.

트랙 그룹 내의 트랙 배치 추정: 위에서 ST3250820A 모델이 280개의 트랙이 하나의 그룹으로 묶이게 된다는 것을 알게 되었다면, 다음 작업은 이 트랙들의 상하좌우 배치 순서를 추정하는 것이다. 디스크의 스핀들을 지나도록 수직으로 자르고 다시 절반을 자른 수직 단면에서는 모든 트랙들이 단 하나의 점으로 표현이 된다. 결국 트랙간의 배치를 파악하는 문제는 각 점에 LTA 를 할당하는 문제로 표현할 수 있고, 곧 행번호가 디스크 헤드 번호, 열번호가 실린더 번호인 행렬로 나타낼 수 있게 된다. 좀더 구체적으로 ST3250820A는 3개의 헤드를 가지고 있으므로 3 ×94 의 행렬에 280 개의 LTA를 할당함으로써 한 트랙 그룹내의 트랙 배치를 정의할 수 있게 되고, 전체적인 레이아웃은 앞의 트랙 그룹의 시퀀스로 정의할 수 있게 된다.

트랙 그룹내의 매핑 행렬을 구하기 위해 임의의 지점에 한 트랙을 고정해두고 다양한 구간에 대해 두 번째 트랙까지의 스위치 시간을 구해볼 필요가 있다. 결과로 얻어지는 그래프의 곡선은 크게 4가지 유형으로 분류할 수 있는데 1) 단조 증가/감소, 2) 일정한 구간, 3) 대칭적인 극대/극소점, 4) 비대칭적인 극대/극소점 이다. 그림3 의 경우 항상 A라는 트랙을 먼저 Seek했다가 곧바로 [A+1, A+2k] 구간의 트랙들을 한번씩 Seek 했을 때 걸리는 시간들의 분포를 나타낸 그래프이다. 이 경우에는 그래프 곡선의 3번째 유형인 대칭적인 극대점이 나타난 것이고, 이에 해당하는 매핑 행렬로는 그림4에 나타난 두가지를 생각해 볼 수 있다.



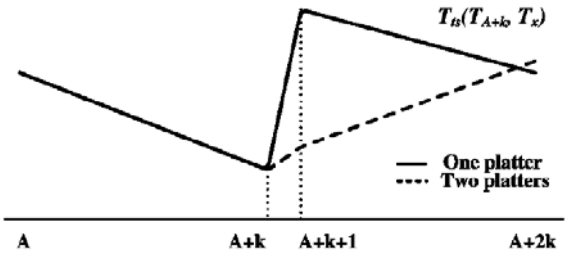
(그림 3) 대칭적인 극대점

$$\begin{bmatrix} A+k+1 & \dots & A+2k & A & A+1 & \dots & A+k \end{bmatrix}$$

$$\begin{bmatrix} A & A+1 & \dots & A+k-1 & A+k \\ A+2k & A+2k-1 & \dots & A+k+2 & A+k+1 \end{bmatrix}$$

(그림 4) 대칭적인 극대점일 경우 가능한 두가지의 매핑 행렬

둘 중 올바른 매핑을 찾기 위해 추가적인 실험을 수행해야 하는데, 이번에는 항상 A+k 라는 트랙을 먼저 Seek 한 뒤 곧바로 [A, A+2k] 에 속하는 트랙들을 한번씩 Seek 하는 경우의 그래프 결과를 살펴보아야 한다.



(그림 5) A+k 를 고정시켰을 때의 스위치 시간 분포

그림5는 그림4의 가능한 두 매핑 행렬에 대해 예상되는 그래프 결과이다. 만약 첫번째 행렬과 같이 2k+1 개의 트랙들이 동일 헤드상에, 즉 하나의 플래터위에 존재한다면 비대칭적 극소점이 나타날 것이고, 두번째 행렬과 같이 두개의 플래터위에 존재한다면 대칭적 극소점이 나타날 것이다.

이와 같이 두개의 트랙을 연속적으로 Seek을 하되, 첫번째 트랙은 항상 고정시키고 두번째 트랙을 변화시켜 가면서 스위치 시간을 측정하면 한 트랙 그룹내의 트랙들이 어떤 배치를 이루는 가를 알아낼 수 있고, 결과적으로 매핑 행렬을 구해낼 수 있다.

Model Name	WD Caviar SE (WD2500JB)	Seagate Barracuda 7200.10 (ST3250820A)
Capacity	250GB	250GB
RPM	7200	7200
Interface	IDE (ATA-100)	IDE (ATA-100)
Year	2004	2006
Buffer Cache	8MB	8MB

<표 1> 디스크 드라이브 기본 파라미터

3. 실험 결과

표1은 실험에서 사용했던 두 ATA 디스크들에 대해 벤더가 제공한 기본 파라미터를 표시한 것이다. 둘 모두 ATA 인터페이스를 따르기 때문에 매핑 정보를 직접 알아낼 수는 없고 위에서 언급되었던 알고리즘을 사용하여 실험적으로 추측해야만 한다. 결과는 다음과 같다.

$$\begin{bmatrix} A & A+1 & \dots & A+30 & A+31 \\ A+63 & A+62 & \dots & A+33 & A+32 \\ A+64 & A+65 & \dots & A+94 & A+95 \\ A+127 & A+126 & \dots & A+97 & A+96 \\ A+128 & A+129 & \dots & A+158 & A+159 \\ A+191 & A+190 & \dots & A+161 & A+160 \end{bmatrix}$$

```
Read/write head number :  6
Platter (recording surface) number:  3
(6)
Sectors per Track :  1116 ~ 675
Track Group Size :  192 (tracks)
```

(그림 6) WD Caviar SE 의 내부 정보

$$\begin{bmatrix} A & \dots & A+46 & A+233 & \dots & A+279 \\ A+47 & \dots & A+92 & A+187 & \dots & A+232 \\ A+93 & \dots & A+139 & A+140 & \dots & A+186 \end{bmatrix}$$

```
Read/write head number :  3
Platter (recording surface) number:  2
(3)
Sectors per Track :  1496 ~ 836
Track Group Size :  280 (tracks)
```

(그림 7) Seagate Barracuda 의 내부 정보

그림 6과 7의 결과를 보면 디스크 모델마다 매핑 행렬의 모양이 달라진다는 점을 확인할 수 있다. 즉 이런 정보를 상위 레이어(예를 들면 운영체제)에서 사용하기 위해서는 먼저 앞서 소개한 방법을 이용하여 내부 정보를 추출하는 시간이 필요하다. 결과가 올바른지를 확인해 보기 위해서, 위 행렬의 한 행에 대해서만 트랙 스위치 시간을 구해보았는데 그래프 결과는 seek curve 로 나타났고, 이것은 트랙들의 배치가 예측과 일치하다는 것을 암시한다고 볼 수 있을 것이다.

3. 결론

디스크의 물리적인 특징들을 추출해내기 위하여 실험적인 방법을 사용하였고, 그 결과 디스크 내부의 구조적인 모델을 정확하게 구성할 수 있었다. 실험 결과에 의하면 현대의 디스크 드라이브들은 벤더들에 의해 각각 고유의 복잡한 내부 구조를 가지게 되었음을 확인할 수 있었다. 이 결과들은 디스크 드라이브 모델링이나 I/O 성능을 최적화하는 다른 많은 연구들에 적용될 수 있을 것이다.

감사의 글

이 연구는 서울시 산학연 협력 사업에서 부분적으로 지원 받았으며, 서울대학교 컴퓨터 연구소는 이 연구의 시설을 제공하였습니다.

참고문헌

[1] Zoran Dimitrijevic, Raju Rangaswami, David Watson, and Anurag Acharya, "Diskbench: User-level Disk Feature Extraction Tool", UCSB Technical Report TR-2004-18, 2004.

[2] Bruce L. Worthington, Gregory R. Ganger, Yale N. Patt and John Wilkes, "On-line Extraction of SCSI Disk Drive Parameters", In Proceedings of the ACM SIGMETRICS, pp.146-156, 1995.

[3] Chris Ruemmeler and John Wilkes, "An introduction to disk drive modeling", IEEE Computer 27(3):17-29, March 1994.

[4] Jiri Schindler, John Linwood Griffin, Christopher R. Lumb and Gregory R. Ganger, "Track-aligned Extents: Matching Access Patterns to Disk Drive Characteristics", FAST, January 28-30, 2002.