최근의 디스크 드라이브의 내부 동작 모델링을 통한 컴퓨터 I/O 시스템의 성능 향상 모색*

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

Operational Modeling of Modern Disk Drives for Improving Disk I/O Performance

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

요 약

우리는 이 논문에서 현대 마그네틱 디스크 드라이브를 대상으로 하여 내부의 세부적인 동작 메커니즘을 상위 레벨의 검침 프로그램을 이용하여 정확하게 측정하고 이를 통해서 디스크 드라이브를 사용하는 다양한 응용 프로그램 및 운영 체제 시스템의 디스크 입/출력 성능을 개선하고자 한다.

1. 서론

컴퓨터 시스템 및 관련 기술의 끊임없는 발전은 다양한 응용 프로그램들이 크고 복잡한 데이터를 사용할 수 있도록 하였다. 이러한 엄청난 데이터를 감당할 만한 저장소로서 마그네틱 하드 디스크는 모든 정보 처리 시스템 분야에서 극도로 유용한 역할을 담당하고 있다 [1,2]. 그 크기에서 하드 디스크는 꾸준히 무어의 법칙 (Moore's law)을 따라 발전하고 있으며, 데이터 전송률이나 펌웨어 기술 또한 마찬가지의 양상을 보이고 있다 [3]. 그럼에도 불구하고 컴퓨터에서 프로세싱 요소 와 저장소 요소와의 증가하는 성능 차는 실시간 시스템 및 멀티미디어 시스템, 그리고 데이터베이스 시스템과 같이 대용량 저장소를 필요로 하는 컴퓨터 시스템 및 응용 프로그램의 성능의 향상을 저해하는 부분이다.

자연스럽게 많은 연구들이 보다 효율적인 디스크 드라이브의 사용에 초점을 맞추었으며, 최근의 많은 연구에서는 디스크 드라이브에서의 데이터 접근 방식의 문제점을 지적하고 이를 해결하고자 하였다 [4,5]. 그 중 대표적인 것이 디스크 헤드의 이동 및 원판의 회전 지연 시간이다. 이 기계적인 방식의 접근은 그 지연시간에 있어서 전기적인 신호에 의해 동작하는 메모리에 비해 무려 1,000~10,000 배 이상의 차이를 보인다.

이 논문에서 우리는 현대 디스크 드라이브의 내부 동작 메커니즘을 상위 레벨 (운영 체제 커널 및 사용자 레벨)에서 실험적으로 정확하게 추출하기 위한 기법을 제안한다. 우리가 관심을 갖는 요소로는 디스크의 헤드 이동 시간 (seek time), 회전 시간 (rotational delay), 데이터 전송 시간, 및 기타 지연 시간을 포함한다. 우리는 실험적인 방법*을 통해서 이러한 요소들을 정확하게 파악하는 기법을 제안하고 각 요소들이 전체 디스크 입출력 지연 시간에 어떻게 관련이 되는지에 대해서도 파악한다. 우리의 측정 기법은 디스크의 다양한 인터페이스 타입 - SCSI, ATA, SATA, SAS - 에 무관하게 적용가능 하다는 특징을 갖는다. 이러한 분석 및 측정 결과를 이용하여 우리는 디스크의 내부 동작에 관한 모델을 세울 수 있다.

하드 디스크에 대한 실제 입출력을 통한 실험 결과 우리의 모델은 디스크의 입출력 지연 시간과 약 1% 내외의 오차를 보이면서 디스크의 내부 동작을 잘 반영하는 것을 보였다. 우리의 디스크 모델링 결과는 디스크 블록 할당 기법 및 입출력 스케쥴링 기법등에 이용될 수 있다 [9,10,11]. 우리는 이러한 디스크 모델이 디스크 입출력의 효율성을 증대하는데 잘 적용될 수 있으리라 기대한다.

논문의 구성은 다음과 같다. 다음 절에서는 현대 디스크 드라이브의 동작의 모델링하기 위한 디자인을 설명한다. 3절에서는 디스크 동작 모델의 인자 추출 기법 및 결과를 설명하고 4절에서는 모델의 검증 결과를 설명하며, 5절에서 결론을 맺는다.

2. 현대 디스크 드라이브의 동작 모델 디자인

현대 디스크 드라이브의 동작을 모델링 하기 위해서 우리는 하나의 입출력 요청을 처리하기 위해 디스크 드라이브의 내부에서 수행되는 일련의 과정에 대해 조사하였다.

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

이와 관련된 과거의 연구들 [6,7,8]을 조사한 결과, 우리는 디스크의 동작 모델에 대한 가설을 세울 수 있었다. 우리는 실험을 통해서 이러한 가설을 검증하고 과거의 디스크 모델과 현대의 모델간의 어떠한 차이가 있는지를 통해서 새로운 디스크 동작 모델을 정립할 수 있다. 가설로부터, 하나의 입출력 요청을 처리하기 위해 소요되는 전체 수행 시간은 크게 *디스크 컨트롤러 지연 시간, 헤드 이동 시간,* 그리고 *데이터 전송 시간*으로 나뉜다. 좀 더 세부적으로 전체 수행 시간 T_{req} 는 다음과 같은 다섯 가지 요소로 구분 된다 (그림 1 참조).

$$T_{req} = T_{pre} + T_{pos} + T_{rot} + T_{trans} + T_{post} \qquad (1)$$

여기서 T_{pre} 와 $Tpost$ 는 각각 입출력 명령을 해독하고 논리적인 디스크 주소 (Logical Block Address)를 물리적인 주소로 변경하며 (T_{pre}), 처리한 요청의 결과 및 상태를 전송하는데 걸리는 (T_{post}) 컨트롤러 지연 시간을 의미한다. 그리고, T_{pos} 와 T_{rot}는 해당하는 디스크의 위치 (실린더 및 트랙) 로 head를 이동시키고 (T_{pos}), 원하는 sector 위치로 platter를 회전시키는 동안 대기하는 시간 (T_{rot})을 가리킨다. T_{trans}는 디스크의 기록 표면에 데이터를 기록하거나 읽는 시간이다.

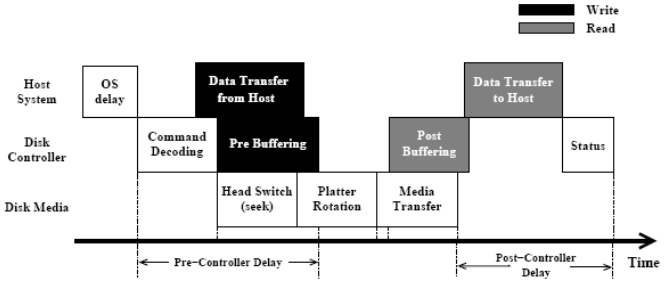
일반적으로 최근의 디스크 드라이브는 내부적으로 버퍼링과 캐싱을 위해 사용되는 비교적 적은 양의 메모리를 가지고 있으며, 이를 빠른 상위 레벨과의 속도 차를 경감시키기 위한 속도 정합 버퍼 (speed matching buffer)로 사용한다 [6]. 디스크 입출력 요청의 경우, 이러한 버퍼의 영향으로 별도의 버퍼링 지연 시간을 포함하게 되는데, 쓰기 요청의 경우, 쓰려는 데이터를 실제 기록하기 전에 버퍼링 하므로 선행 버퍼링 지연 (pre-buffering delay) 이라 정의하고, 읽기의 경우 미디어로부터 읽은 데이터를 전송하기 위한 후행 버퍼링 지연 (post-buffering delay) 이라고 정의한다. 따라서 우리는 이러한 버퍼링 지연을 컨트롤러 지연 시간에 포함시킨다.

다른 고려사항으로는 T_{pos} 와 T_{rot}과 같은 디스크의 기계적인 지연 시간이 여러 가지 외부적인 요인에 의해서 (물리적인 온도, 외부적인 진동, 디스크 결함 또는 트랙 및 실린더 휨 (track/cylinder skew)) 적은 오차를 가질 수 있다는 것이다. 문제는 이러한 작은 오차가 모델의 정확도를 크게 떨어뜨릴 수 있다는 데에 있다. 예를 들어, $Tpos$ 의 작은 오차는 디스크의 한 바퀴를 회전 시키는 만큼 - 7200 rpm disk에 대하여 약 8ms - 의 오차를 유발할 수 있다. 따라서 우리는 이러한 기계적인 지연 시간의 허용 가능한 오차 범위를 반복적인 실험 및 다양한 외부 조건 하에서 측정하고 이를 모델에 반영하여 모델의 정확도를 개선해야 한다.

3. 디스크 동작 모델의 인자 추출 기법

3.1 컨트롤러 지연 시간 (Controller Overhead)

우리는 디스크 내부 컨트롤러의 세부적인 동작을 정확하게 알아내기는 힘들지만, 실험을 통해서 하나의 입출력 요청에 대해 컨트롤러가 어떤 영향을 끼치는 지에 대해서는 파악할 수 있다. 컨트롤러의 지연시간을 측정하기 위해서



<그림 1>. 디스크의 동작 모델 가설.

우리는 연속된 입출력 요청간의 최소 시간 (Mean Time Between Request Completions : MTBRC)을 이용한다[]. MTBRC 로부터 우리는 요청간 지연 시간 (Inter Request Delay) T_{inter}를 다음과 같이 정의할 수 있다. 여기서 R_1과 R_2는 각각 하나의 입출력 요청을 의미한다.

$$T_{inter}(R_1, R_2) = T_{post}^{R_1} + T_{pre}^{R_2} + T_{pos}^{R_2} + T_{rot}^{R_2} \qquad (2)$$

식 2로부터 우리는 디스크의 기계적인 지연 시간을 제거한다면, 하나의 입출력 요청에 소요되는 컨트롤러 지연 시간 T_{pre} + T_{post}를 측정할 수 있다. 이를 위해서 우리는 같은 트랙 상에 존재하는 인접한 섹터를 접근하는 두 개의 요청을 생성하여 수행함으로서 디스크 헤드의 이동과 회전 지연 시간을 제거할 수 있었다. 또한 쓰기와 읽기의 요청을 조합한 연속된 요청을 수행함으로서 요청의 타입에 따른 컨트롤러 지연 시간의 변화도 알아 낼 수 있었다.

우리는 표1에 나타난 디스크들에 대해서 위와 같은 과정을 수행했고, 그 중에서 T7K250 모델에 대한 컨트롤러 지연 시간을 다음과 같이 측정하였다.

$$\begin{aligned}
&T_{con}(type_1, size_1, type_2, size_2) = \\
&T_{pre}(type_1, size_1) + T_{post}(type_2, size_2) \\
&T_{con}(write, x, write, y) = 160.69y + 228.78(usec) \\
&T_{con}(read, x, write, y) = 165.59x + 141.31(usec) \\
&T_{con}(write, x, read, y) = 1306(usec) \\
&T_{con}(read, x, write, y) = 165.59x + 160.69y + 1263.72(usec)
\end{aligned}$$

실험 결과로서 우리는 요청 타입 및 요청 데이터 사이즈에 따라 컨트롤러의 지연 시간이 가변적인 것을 알 수 있었다. 이는 앞서 언급한 버퍼링 지연의 영향임을 확인할 수 있었다. 즉, 쓰기 요청의 경우 선행 버퍼링의 영향으로 요청 데이터의 사이즈에 비례하여 컨트롤러 지연시간이 증가하는 것을 알 수 있었다. 마찬가지로 읽기 요청의 경우 후행 버퍼링의 영향으로 컨트롤러의 지연시간이 증가하는 것을 알 수 있었다. 또한 어떠한 버퍼링 지연도 나타

Model	WD Caviar SE WD2500JB	Seagate Barracuda 7200.10 ST3250820A	Hitachi Deskstar T7K250
Capacity	250GB	250GB	250GB
RPM	7200	7200	7200
Interface	IDE	IDE	IDE
Year	2005	2006	2005
Buffer Cache	8MB	8MB	8MB

<표 1>. 실험 디스크 명세

나지 않는 경우에 대해서 컨트롤러 지연시간은 요청 사이즈와 상관없이 상수로 나타남을 확인하였다.

3.2 디스크 헤드 이동 시간 (Disk Head Switch Time)

디스크의 요청 지연 시간 중에서 가장 큰 지연시간을 차지하면서 가장 큰 오차를 가지는 요소가 바로 디스크 헤드 이동 시간이다. 일반적으로 탐색 시간 (seek time) 이라고 불리는 헤드 이동시간은 디스크 암의 끝에 붙어 있는 읽기/쓰기 헤드가 접근하려는 트랙 위로 이동할 때까지 모터를 회전시켜 암을 이동하는데 걸리는 시간과 활성화된 헤드를 전환시켜 읽거나 쓰려는 디스크 원판을 변경하는 경우를 모두 포함하는 개념이다. 헤드 이동 시간을 측정하기 위해서 우리는 앞서 언급한 식 1을 이용한다. 이 경우에는 T_{pos}에 해당하는 요소만을 추출하기 위해서 동일한 트랙에 요청되는 2개의 연속된 입출력 요청과 서로 다른 트랙에 요청되는 2개의 연속된 요청의 응답 시간을 비교한다. 이러한 방법을 통해서 실험적으로 헤드 이동 시간만을 추출할 수 있으며, 이러한 과정을 반복적으로 다양한 조건하에서 - 주로 디스크의 물리적인 온도를 변화시키면서 - 측정해 보았다. 그 결과 표 1의 ST3250820A 모델에 대해서 그림 2와 같은 실린더 거리에 따른 헤드 이동 시간을 구할 수 있었다.

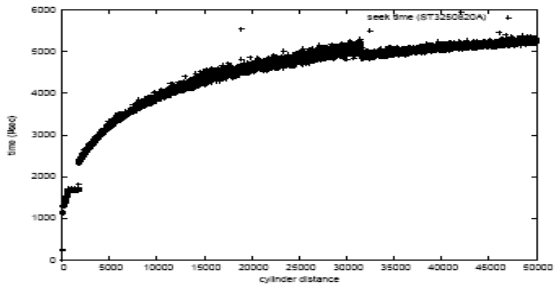
또한 반복적인 실험을 통해서 동일한 거리에 따른 헤드 이동 시간의 에러의 분포를 알아낼 수 있었으며 이를 이용하여 표 2와 같은 디스크 헤드 이동 시간에 대한 모델을 세울 수 있었다.

seek distance	fitting curve (μsec)	R^2(%)	margin error (μsec)
$1 \le x \le 15$	$y = 260$	n/a	35
$16 \le x \le 95$	$y = 845$	n/a	98
$96 \le x \le 267$	$y = 1000$	n/a	120
$268 \le x \le 665$	$y = 0.89\,x + 1091.52$	97.98	71
$666 \le x \le 1775$	1700	n/a	160
$1776 \le x \le 15000$	$y = 23.30\,\sqrt{x} + 1328.384$	98.6	400
$x \ge 15001$	$y = 0.04\,x + 3480.279$	92.69	850

표 2. 디스크 헤드 이동 시간 모델링 결과 (ST3250820A)

3.3 디스크 회전 지연 시간 (Disk Rotational Latency)

일반적으로 많은 연구에서 하나의 디스크 요청에 대한 회전 지연 시간을 평균 회전 지연 시간으로 단순화하여 모델링 하였다. 그러나 이러한 단순화는 모델의 오차를 허



<그림 2.> 거리에 따른 디스크 헤드 이동 시간 분포

용할 수 없을 정도로 크게 만들 수 있다. 우리의 모델링의 목적은 디스크의 내부 동작을 이용하여 디스크 성능을 향상시키고자 함이므로 좀 더 세부적인 모델을 구하는 것은 중요한 선행 과정이 된다. 디스크가 한 바퀴를 도는 데 걸리는 시간은 7,200 RPM의 디스크에 대해서 8 ms 정도가 걸린다. 그러나 하나의 요청이 처리될 때, 좀 더 세부적으로 디스크 헤드가 원하는 위치에 올라온 순간에 읽기나 쓰기를 시작하려는 시작 섹터의 위치까지 회전하는데 걸리는 시간을 정확히 측정하는 것은 다음과 같은 요소들을 고려해야만 한다.

- 전체 회전 지연 시간 T_R.
- 트랙 또는 실린더 간의 휨 요소 (skew factor)
- 컨트롤러 지연 시간
- 이전에 처리된 요청의 위치와 현재 위치와의 실린더 거리 및 이전 요청이 종료된 시점과의 시간차.

이러한 요소들 역시 실험적으로 측정이 가능하며 공간의 제약으로 인해 정확한 측정 방법 및 결과는 생략한다.

4. 디스크 모델의 정확도 검증 (Model Verification)

우리는 우리의 디스크 동작 모델의 정확도를 실제 디스크 드라이브의 요청 응답 시간과 우리의 모델을 통한 예상 응답 시간과의 비교를 통해서 검증한다. 검증을 위해서 우리는 가상의 작업 부하(synthetic workload)을 생성하고 표 1에 나온 디스크들에 대해서 읽기/쓰기 요청을 모방하였다. 작업 부하의 각 요청들은 인자 λ의 도달율 (arrival rate)을 갖는 poisson 분포를 보인다.

우리의 모델의 정확도를 정량화 하기 위해서 우리는 demerit figure of model [6]을 사용하였다. 우리는 우리의 동작 모델을 다음과 같은 부분적이고 평균적인 모델과 비교한다.

- 헤드 이동 모델 : 가장 단순한 모델이며 오직 디스크의 평균 헤드 이동 시간과 컨트롤러 지연시간의 고정된 [1] 부분만을 고려한다.
- 헤드 이동 및 회전 모델 : 헤드 이동 모델에 평균 회전 지연 시간 - 전체 회전 지연 시간의 절반 - 을 추가하여 고려한다.
- 헤드 이동 및 회전 그리고 버퍼링 모델 : 헤드 이동 및 회전 모델에 선행/후행 버퍼링 지연 시간을 추가적으

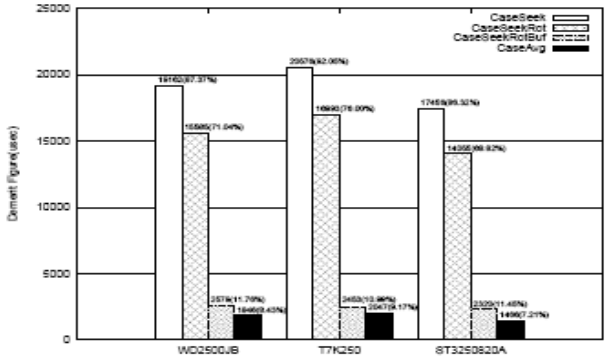


그림 3. 디스크 동작 모델 검증 결과

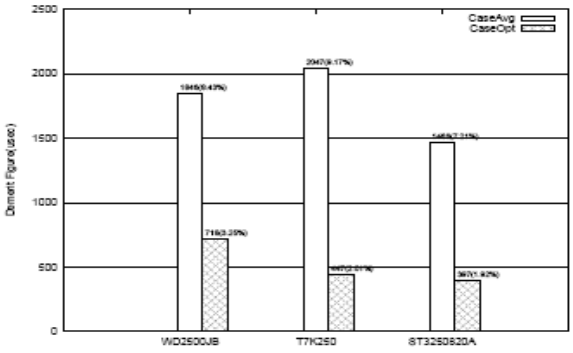


그림 6. 개선된 디스크 동작 모델 검증 결과

로 고려한다.

그림 3은 표 1의 디스크 드라이브들에 대해 demerit figure of model로서 모델의 정확도를 묘사한 그래프이다. 전체적으로 우리의 모델이 다른 모델에 비해 높은 정확도를 나타내며, 실제 디스크 드라이브와 약 7~9% 정도의 차이를 보인다. 또한 그래프를 통해서 각 지연 시간 요소에 대해서 보면 버퍼링 지연 시간의 고려가 모델의 정확도에 가장 중요한 역할을 한다. 그러나 우리의 모델의 정확도는 더 높은 정확도를 보일 거라는 우리의 기대와는 달리 생각보다 많은 오차를 보였다. 우리는 오차의 분포를 분석해 보았고, 그 결과 몇 개의 입출력 요청에서 거의 전체 회전 지연 시간 만큼의 오차가 전체 모델링의 정확도를 떨어뜨림을 확인할 수 있었다. 좀 더 세부적인 분석을 통해서 이러한 오차는 앞서 언급한 디스크 헤드 이동 시간의 오차 범위보다 예측한 회전 거리가 더 짧은 경우 발생한 다는 것을 확인하였다. 즉, 디스크 헤드의 탐색 시간의 오차 때문에 한 바퀴 만큼의 회전 지연 시간의 오차가 발생한다는 것이다.

우리는 이러한 오차를 고려하여 오차의 범위보다 예측한 회전 거리가 짧은 경우 오차에 의해 여분의 전체 회전 지연 시간만큼이 더 걸릴 수 있다는 분석 결과를 모델에 반영하여 이러한 경우 예상 지연 시간이 하나의 값이 아닌 두 가지 값으로 나오도록 모델을 수정하였다. 이를 토대로 모델을 재검증한 결과 그림 4와 같이 더 정확한 모델링 결과를 얻을 수 있었다. 그림에서와 같이 개선된 모델은 약 1~3%의 모델링 오차를 포함하며, 이는 평균적으로 약 500 usec 내외의 지연 시간의 차이를 의미한다. 이러한 실험 결과는 우리의 모델이 정확하게 실제 디스크 드라이브의 내부 동작을 반영한다는 것을 의미한다.

5. 결론

우리는 실험적인 기법을 사용하여 상위 레벨에서 숨겨진 디스크의 내부 동작을 추출하고 이를 토대로 정확한 디스크 동작 모델을 정의하고 검증하였다. 검증 결과 우리의 모델은 상당히 높은 정확도를 가지고 실제 디스크의 내부 동작을 모사함을 확인할 수 있었다. 이러한 정확한 모델을 토대로 우리의 연구는 디스크 입출력의 효율성을 향상시키고자 하는 다양한 연구를 위한 기본적인 토대가 될 수 있으리라 기대한다.

참고문헌

[1]. M. Beynon and T. Kurc and U. Catalyurek and A. Sussman and J. Saltz. Efficient manipulation of large datasets on heterogeneous storage systems. In Proceedings of the 11th Heterogeneous Computing Workshop (HCW2002), 2002.

[2] Meikel Poess and Raghunath K. Othayoth. Large scale data warehouses on grid: Oracle database 10g and HP proliant servers Proceedings of the 31st international conference on Very large data bases (VLDB 2005), pages 1055 - -1066, 2005

[3] D.A. Thompson and J.S. Best. The future of magnetic data storage technology, IBM Journal of Ressearch and Development, Vol. 44, Number 3, 2000.

[4] Jiri Schindler, John Linwood Griffin, Christopher R. Lumb, and Gregory R. Ganger. Track-aligned extents: matching access patterns to disk drive characteristics. Conference on File and Storage Technologies (Monterey, CA, 28 - -30 January 2002.

[5] D. Jacobson, J. Wilkes, Disk Scheduling Algorithms Based on Rotational Position, Hewlett-Packard Technical Report, HPL-CSP-91-7, Feb. 26, 1991.

[6] C. Ruemmler and J. Wilkes. An introduction to disk drive modeling. IEEE Computer, 27(3):17 - -28, 1994.

[7] B. L. Worthington, G. R. Ganger, Y. N. Patt, and J. Wilkes. On-line extraction of SCSI disk drive parameters. In SIGMETRICS, pages 146 - -156, May 1995. 14

[8] Z. Dimitrijevic, R. Rangaswami, and E. Chang. Diskbench: User-level disk feature extraction tool. Technical report, UCSB, November 2001

[9] Christopher R. Lumb, Jiri Schindler, and Gregory R. Ganger. Freeblock scheduling outside of disk firmware. In Proceedings of the 1st USENIX Conference on File And Storage Technologies (FAST'02), Monterey, CA, pages 275 - -288 USENIX, Janurary 2002.

[10] Robert M. English and Alexander A. Stepanov. Loge: a self-organizing storage device. USENIX Winter 1992 Technical Conference Proceedings, pages 237 - -51, January, 1992.

[11] S. W. Schlosser, J. Schindler, S. Papadomanolakis, M. Shao, A. Ailamaki, C. Faloutsos, and G. R. Ganger. On multidimensional data and modern disks. In Proceedings of the 4th USENIX Conference on File and Storage Technology (FAST ï05).