3 차원 적층형 DRAM 의 성능 향상을 위한 데이터 배치 기법

이광준, 전민제, 정의영
연세대학교 전기전자공학부

초록

This paper presents an efficient way to place data to memory banks of 3D-stacked DRAM. The conventional DRAM has limitation on the number of banks due to several physical constraints such as limited pin/chip count. On the other hand, the 3D-stacked DRAM can have larger number of banks than the 2D one since the off-chip wiring can be now replaced by the inter-wafer vertical links, so called Through-Silicon-Vias. We take advantage of it by applying fine-grained bank-interleaving, such that the data which used to be placed in a row can be placed over multiple successive banks. The result shows that there is 1.51x speedup over the conventional data placement when of 8x8 memory banks are used, and the speedup is proportional to the burst length of the transaction.

1. 서론

컴퓨터 내부의 중앙 처리 장치(CPU)나 연산에 사용되는 코어들은 최근 몇 년 동안 비약적인 속도의 증가를 이루었다. 하지만 이와 반대로, DRAM 으로 대표되는 메인 메모리 모듈의 속도는 그 장치 구조의 근본적인 한계로 인하여 정체되어 있는 것이 사실이다. 날이 갈수록 점점 더 막대한 양의 메모리 접근을 필요로 하는 현재의 시스템에서, 메모리 모듈의 용량 역시 그에 발맞추어 증가하고 있지만, 저속의 접근 속도로 인하여 전체 시스템의 병목 현상을 유발하여 시스템 전체의 성능을 대폭적으로 저하시키는 결과를 보이고 있다. 본 논문에서는 최근 활발히 연구되고 있는 TSV(Through-Silicon-Via)를 사용한 3 차원 적층 기술을 통하여 비약적으로 증가한 메모리의 용량을 더 효율적으로 사용하고, 전송 지연 시간을 효과적으로 감소시킬 수 있는 DRAM 내부의 데이터 배치를 연구하였다. 제안된 데이터 배치를 사용하여 시뮬레이션 해본 결과, 메모리 뱅크의 개수를 기존의 8 배만큼 증가시켰을 때 데이터 전송 지연 시간의 감소로 인하여 동일 시간 내에서 처리할 수 있는 메모리 트랜잭션의 양이 최대 51%만큼 증가한 것을 확인할 수 있었다.

2. TSV 를 통한 3 차원 적층 기술

TSV 를 통한 3 차원 적층 기술은 하나의 칩 위에 수직 방향으로 다른 모듈을 적층하고, 칩 내부에 배치된 TSV 를 통하여 내부 모듈들의 연결 및 전원 공급을 수행하는 기술이다. 이를 통하여

1) 수평 방향으로 배치되어 있던 칩 내부의 모듈들이 수직 방향으로 이동함에 따라 전체 칩의 면적이 감소하는 효과를 얻을 수 있고[1],
2) 약 수십㎛ 정도의 짧은 수직 방향 전송 거리를 통하여 칩 내에서의 와이어를 통한 수평 방향의 전송보다 짧은 지연 시간을 기대할 수 있으며[2],
3) 서로 다른 공정을 가진 모듈을 수직 방향으로 연결할 수 있기 때문에, 중앙 처리 장치 위에 DRAM 메모리 모듈을 적층하여 off-chip 메모리 모듈로부터의 긴 전송 거리에 따른 막대한 와이어 지연 시간을 대폭 감소시킬 수 있을 뿐 아니라[1]
4) 2 차원상에서의 배치와 비교하였을 때 칩의 개수 및 그 연결을 위하여 사용할 수 있는 핀의 개수가 늘어남으로써[3] 더욱 다양한 구조를 제작 및 활용할 수 있다.

3. DRAM 데이터 전송 지연 시간 감소를 위한 기법

기존에 사용하던 DRAM 모듈은 CPU 와 비교하였을 때 훨씬 느린 속도를 보였고, 이에 따른 데이터 병목 현상 및 전송 지연 시간 증가 현상이 전체 시스템 속도 저하의 큰 부분을 차지하고 있었다. 이를 해결하기 위하여 여러 가지 기법들이 모색되었는데, 그 중 대표적인 것으로 활성화된 메모리 뱅크 내의 열려있는 행에 하나의 열 주소를 보내어, 그 주소로부터 연속된 여러 바이트의 주소에서 데이터를 전송 받는 버스트(Burst) 전송 기법을 들 수 있다. 기존의 DRAM 에서는 메모리 뱅크 내의 하나의 행을 한 개의 페이지로 설정하고, 연속된 주소를 가진 데이터를 각각의 행 내에 순서대로 배치하여 버스트 전송을 수행하도록 하였다. 이를 통하여 하나의 행을 활성화시킨 후 데이터를 전송할 때마다 매번 열 주소를 새로 보내고 그에 따른 데이터의 전송을 기다리는 시간을 획기적으로 절약할 수 있게 되었다. 하지만 연속된 데이터에 접근하는 속도는 메모리 모듈의 느린 동작 주파수에 비례하므로[그림 1], 더 많은 메모리의 접근을 요구하는 최근의 시스템들에서는 이에 따른 전송 속도의 한계를 극복할 새로운 방안이 필요하게 되었다.

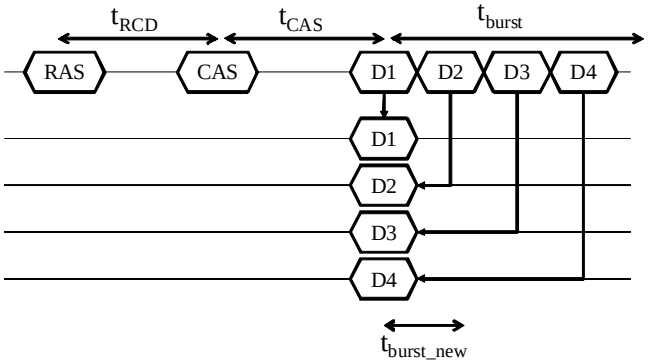
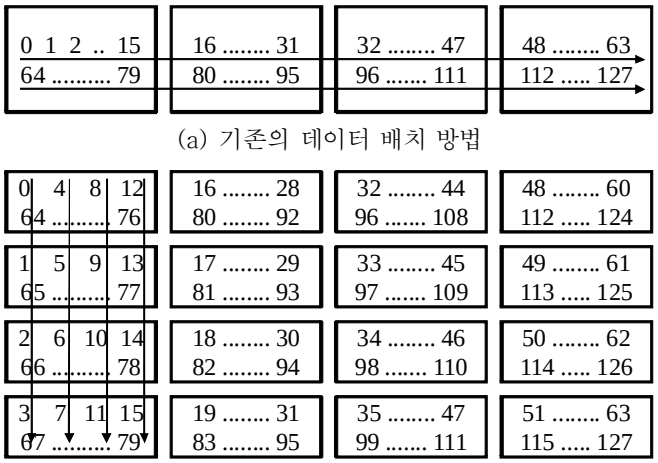


그림 1. 기존 버스트 모드의 전송 속도 및 개선 방안

본 연구에서는 DRAM 의 메모리 뱅크 내에서의 데이터 배치 방법을 조절하여, 더 작은 범위에서의 뱅크 인터리빙(Bank Interleaving)을 통한 빠른 버스트 전송을 수행할 수 있게 하였다. 기존의 2 차원 메모리 모듈에서는 패키지에서 외부와 연결되는 핀의 개수 및 첩의 숫자 등과 같은 여러 제약 조건이 있었으나, 3 차원으로 적층된 메모리 모듈은 수직 방향의 연결이 가능해짐으로 인하여 이런 조건에서부터 보다 자유로워지고[3], 이에 따라 기존의 제약에서 벗어나 주어진 메모리 모듈을 보다 많은 메모리 뱅크로 나누어 사용하는 것이 가능해졌다. 이와 같은 3 차원 적층

메모리 모듈의 특성을 이용하여, 하나의 행 안에 연속된 주소를 가진 데이터를 배치하는 기존의 메모리 뱅크 내에서의 데이터 배치[그림 2]와 달리, 기존의 메모리 뱅크를 몇 개의 뱅크들로 나눈 후([그림 2]에서는 4 개) 각각의 뱅크들을 한꺼번에 활성화시켜 수직 방향으로 연결된 TSV를 통하여 고속의 전송을 수행하게 함으로써 보다 많은 데이터를 동시에 전송할 수 있도록 한다. 이를 통하여 하나의 행에서 데이터를 읽어 들이는 것과 같은 동작을 하면서 데이터 전송 지연 시간은 함께 동작하는 뱅크의 개수만큼 짧아지는 효과를 얻을 수 있다. 또한, 기존의 버스트 전송의 길이를 그대로 사용할 수 있을 뿐만 아니라 늘어난 메모리 뱅크의 개수만큼 버스트 전송의 길이를 늘림으로써 연속된 주소에 위치하고 있는 데이터를 한 번에 더 많이 읽어오는 것이 가능하고, 이를 통하여 동일한 지연시간에 보다 많은 메모리 트랜잭션을 처리할 수 있다. 이를 토대로 동작 주파수가 같은 메모리 모듈을 사용하여 대역폭을 늘릴 수 있고, 이렇게 늘어난 대역폭을 통한 CPU 로의 데이터 전송은 메모리 모듈 및 TSV 에서의 I/O 버스의 고속화를 통하여 이루어질 수 있다.



(a) 기존의 데이터 배치 방법

(b) 3차원 적층 환경에서의 데이터 배치 방법

그림 2. 효율적 버스트 전송을 위한 데이터 배치 방법

4. 측정 결과

제안한 메모리 배치의 성능을 검증하기 위하여, C 로 설계한 cycle-accumulate 시뮬레이터를 사용하였다. DRAM 모듈의 사양은 [표 1]과 같다고 가정하고, SPEC CPU2000 벤치마크[4]의 하나인 176.gcc 의 메모리 주소 트레이스를 사용하여 정해진 시간 동안 시뮬레이션 작업을 수행하여 다음과 같은 결과[그림 3] 를 얻었다.

DRAM 종류	DDR2-800
뱅크 개수	8
RAS latency	5
CAS latency	5
Precharge policy	Open

표 1. 모델링된 DRAM 사양

기존의 메모리 뱅크 8 개를 각각 1/2, 1/4, 1/8 의 크기를 가진 16, 32, 64 개의 뱅크들로 쪼개어 데이터를 새로 배치하여 시뮬레이션 해 본 결과, 기존 값 대비 각각 24%, 41%, 51%의 성능 향상을 보이는 것을 확인할 수 있었다. 메모리 뱅크 개수가 증가할 때마다 데이터 전송량의 증가폭이 줄어드는 이유는, 버스트 데이터 전송 시간이 짧아졌기 때문에 뱅크 개수의 증가에도 불구하고 같은 뱅크 내의 다른 행의 활성화가 잦아져서 위와 같은 결과를 얻은 것으로 생각할 수 있다.

한 번에 메모리에서 읽어올 수 있는 버스트의 길이를 기존의 8 에서 16 으로 늘렸을 때의 결과를 기존의 데이터 배치를 사용하였을 때의 전송량과 비교해 본 결과, 각각 44%, 78%, 103%의 성능 향상을 보이는 것을 확인할 수 있었다. 이는 전체 전송에 걸리는 시간 중 t_{burst} 의 비율이 높아지면서 상대적으로 t_{RCD} 와 t_{CAS} 가 차지하는 비율이 낮아짐으로 인하여 나타나는 결과로 볼 수 있다. 따라서, 제안된 배치는 메모리 컨트롤러에서 버스트의 길이를 길게 정의하였을 경우에 좀 더 효율적으로 작동한다고 할 수 있다.

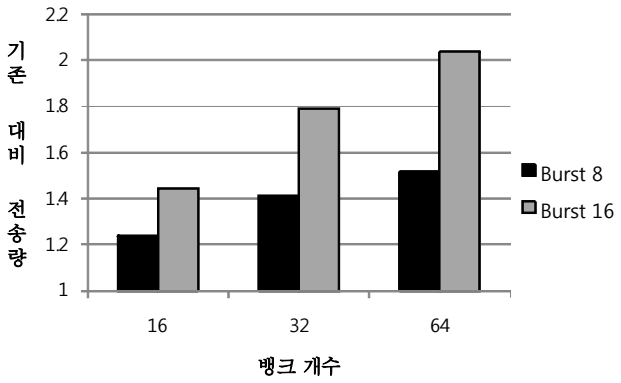


그림 3. 뱅크 분할 및 버스트 길이에 따른 전송량의 변화

5. 결론

TSV 를 이용한 3 차원 적층 기술은 메모리 장벽 문제를 효과적으로 해결할 수 있는 좋은 수단이 될 수 있지만, 이와 같은 변화된 구조를 효율적으로 이용하는 데에 대한 연구는 아직 활발하게 이루어지지 못하고 있다. 본 연구에서는 메모리 뱅크의 개수를 기존의 8 배만큼 증가시켰을 때, 메모리 모듈 내의 데이터의 배치 방법을 변화시킴으로써 최대 51%의 성능 향상을 보이는 것과, 한 번에 전송하는 버스트의 길이가 길수록 더 효율적으로 작동하는 것을 확인하였다. 향후 이를 토대로 3 차원 적층 환경에서의 메모리 모듈의 최적화된 데이터 배치를 연구하여 시스템 전체 성능의 더 큰 향상을 꾀할 수 있을 것이다.

참고문헌

[1] E. Beyne, “The rise of the 3rd dimension for system integration”, in *Interconnect Technology Conference, 2006 International*, 2006, pp. 1–5.

[2] I. Loi, F. Angiolini, L. Benini, “Supporting vertical links for 3D networks on chip: toward an automated design and analysis flow”, *Proc. Nano-Nets, 2007*.

[3] G. H. Loh, “3D – stacked Memory Architectures for Multi-core Processors”, *35th ACM International Symposium on Computer Architecture* (ISCA), pp. 453-464, June 21-25, 2008,

[4] SPEC CPU2000 Benchmark, [on-line] http://www.spec.org/cpu2000/