KU-The Future

GPU 메모리 접근 시간 부채널 특성 분석

정승호, 윤명국, 구건재

Outline

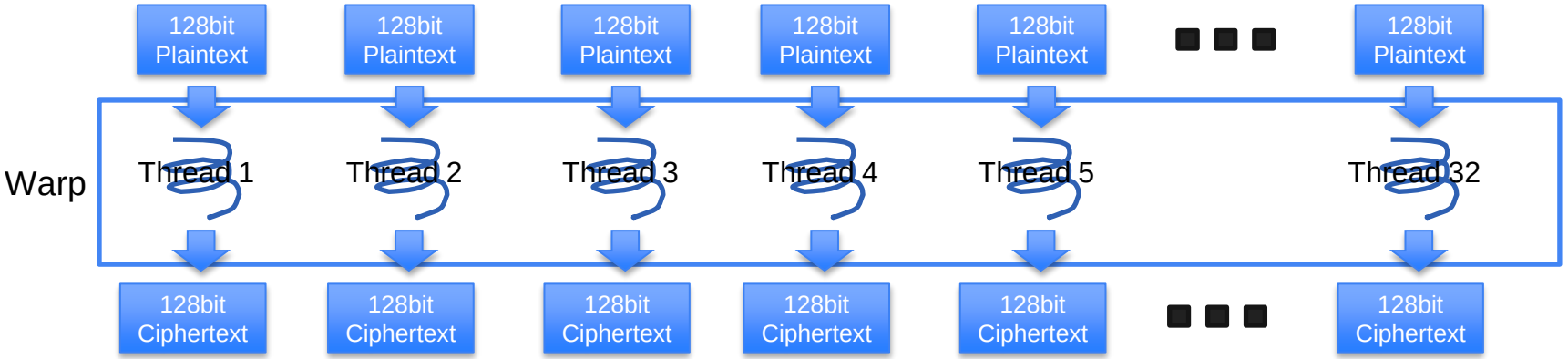
- Introduction/Background
- Motivation
- Memory hierarchy 와 접합크기에 따른 비례관계 변화
- Dummy memory request로 인한 영향
- 결론

AES Implementation on GPU

128bit의 평문을 block으로 encryption 진행

Table-lookup 방식으로 각 Table의 Size는 1KB를 가짐

각 Warp의 안의 Thread들이 각각의 block을 Encryption 수행



Side-channel Timing Attack

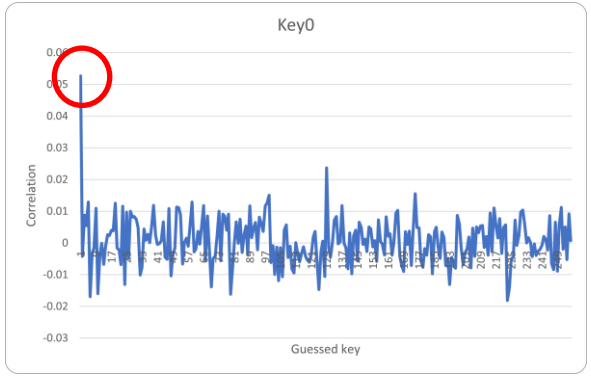
A complete key recovery timing attack on a GPU [Jiang HPCA'16]

AES의 총 10개의 라운드에서 마지막 라운드가 공격에 취약

메모리 요청과 실행시간사이의 비례관계를 공격에 이용

마지막 라운드를 이용하여 메모리 요청 횟수를 추측한 뒤 실행시간과 Correlation하여 Key값을 탈취

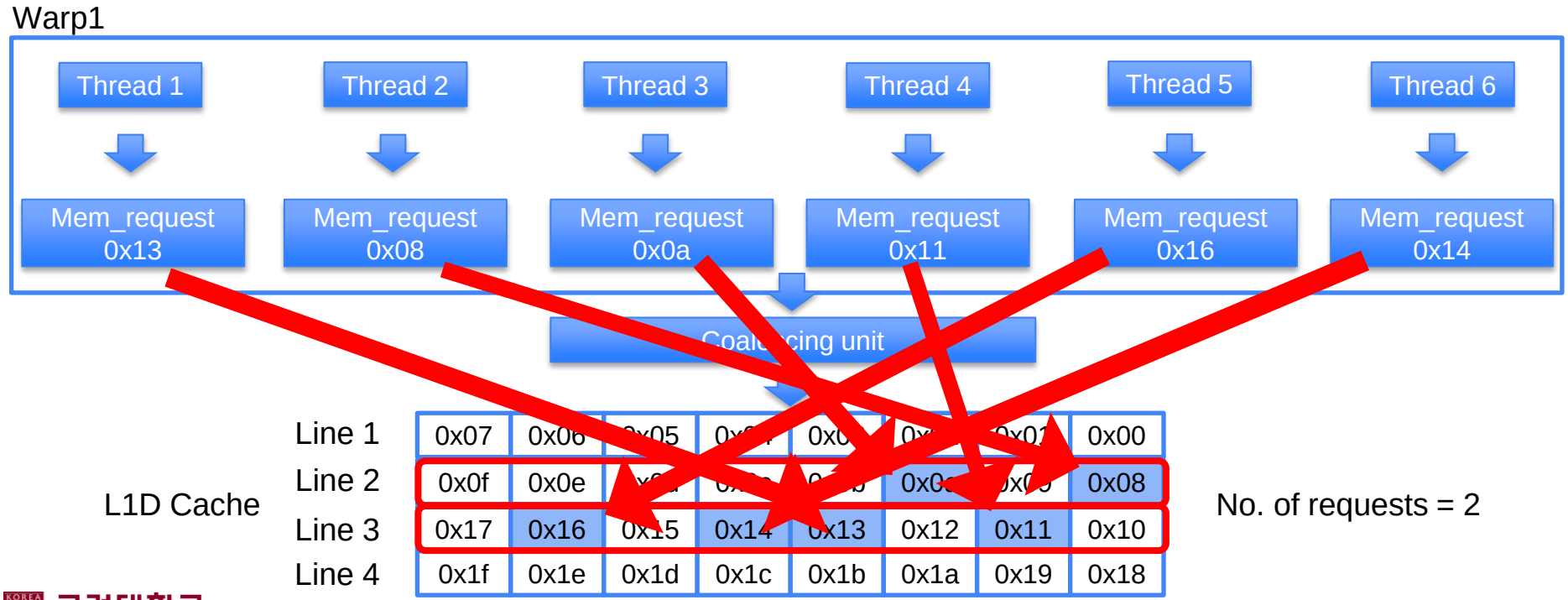
Gussed key	Guessed number of memory request	Execution Cycle
First key byte	Sample #1, Sample #2, … Sample #N	Execution cycle #N
0x00	15 14 14 13 11 12 15 13 15 … 14	69718
0x01	13 14 15 15 14 13 14 15 14 … 14	69621
0x02	14 13 16 14 14 15 14 13 14 … 12	69617
0x03	12 14 16 14 14 14 14 14 14 … 15	69571
0x04	14 15 14 15 14 15 14 12 13 … 14	69602
…	…	…
0xFF	15 15 16 14 13 14 13 13 13 … 13	69513



Memory Coalescing

GPU는 warp가 실행되면서 각 thread 에서 나오는 메모리 요청이 접합 크기에 따라 합쳐짐

이로 인하여 warp에서 실행되는 thread의 개수보다 적은 수의 메모리 요청이 cache line에 접근할 수 있음



Motivation

- GPU에 대한 side-channel timing attack은 커널의 실행 시간을 이용
- GPU의 AES 커널의 execution time에 연관되는 다양한 요소 존재
- 가장 직접적인 요소 : 메모리별 접근 시간
  - 따라서, 메모리 접근 시간 차이에 대한 공격 난이도 분석 필요
- 접근 시간에 영향을 주는 또 한가지 요소 : 접합 크기
  - 따라서, 접합 크기로 인한 공격난이도 차이에 대한 분석 필요

실험환경

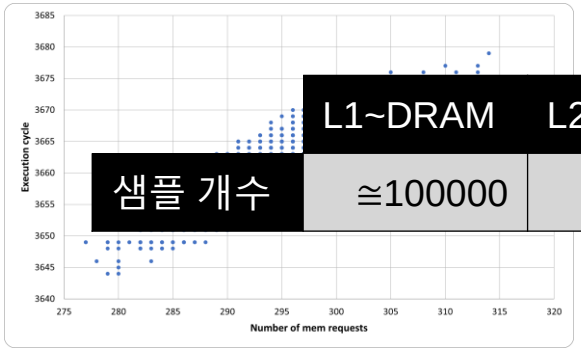
- **AES-128, ECB mode**
- **Sample size : 16 Byte, 32 encryption block**
- **GPGPU-SIM**

Core configuration	15 SMs, 32 threads/warp
L1 Cache	128 Byte Cache line, 4-way, 32 sets
L2 Cache	128 Byte Cache line, 8-way, 64 sets * 12
DRAM	GDDR5 with 6 memory partitions

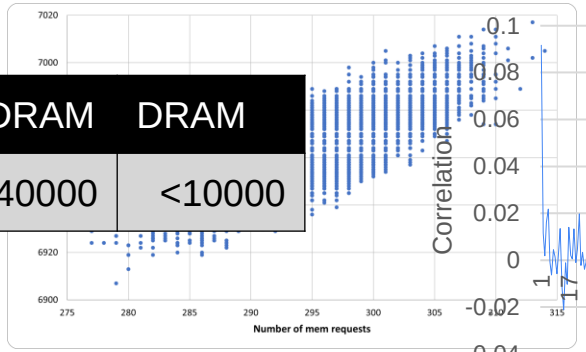
Case 1 : 접합크기 64 byte

- 메모리 계층이 올라갈수록 공격의 난이도 증가
  - 메모리 접근 시간 감소
  - 짧은 실행시간과 메모리 요청 개수 사이의 correlation 찾기 어려움
- 접합 크기가 작기 때문에 그로 인한 메모리 요청 횟수의 다양성 증가

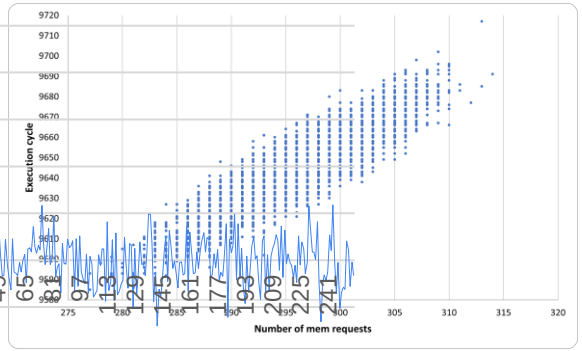
	L1~DRAM	L2~DRAM	DRAM
샘플 개수	≅100000	<40000	<10000



L1 ~ DRAM 사용, Correlation = 0.89



L2~DRAM 사용, Correlation = 0.82

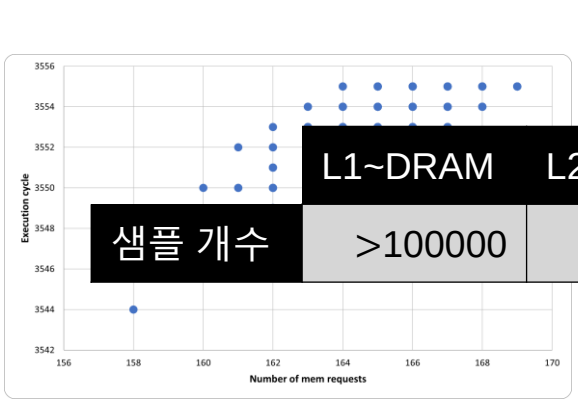


DRAM 사용, Correlation = 0.89

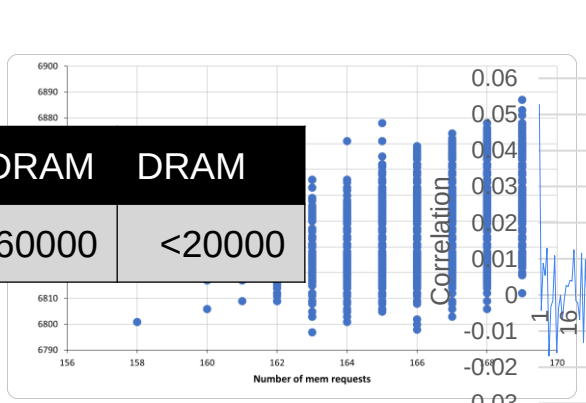
Case 2 : 접합크기 128 byte

- **L1 cache에서 table을 접근할 경우 high positive position을 가짐**
  - **테이블에 할당된 cache Line의 개수 : 8개**
  - **그로 인하여 실행 사이클의 다양성이 작아지고 점들의 겹침이 증가**
- **접합 크기가 크기 때문에 메모리 요청 횟수의 다양성 감소**

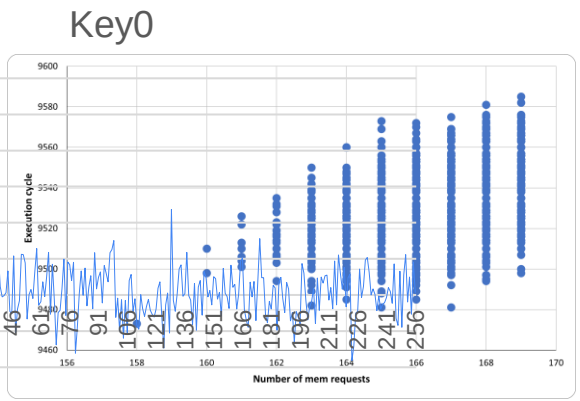
	L1~DRAM	L2~DRAM	DRAM
샘플 개수	>100000	<60000	<20000



L1 ~ DRAM 사용, Correlation = 0.74



L2~DRAM 사용, Correlation = 0.43



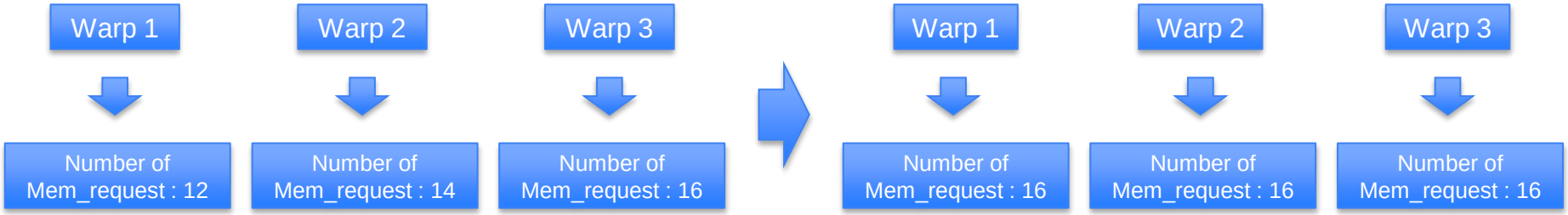
DRAM 사용, Correlation = 0.43

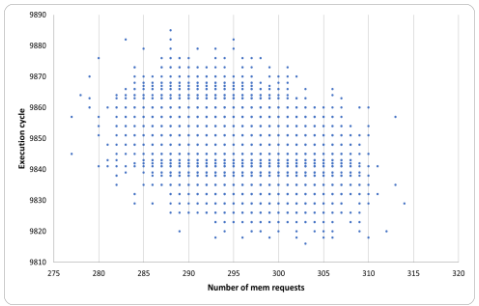
Side-Channel Attack의 난이도 결정 요소

- **부채널 공격의 난이도에 영향을 주는 요소**
  - **메모리 접합 크기로 인한 메모리 요청 횟수의 변화**
  - **메모리 계층에 따른 메모리 접근 시간**

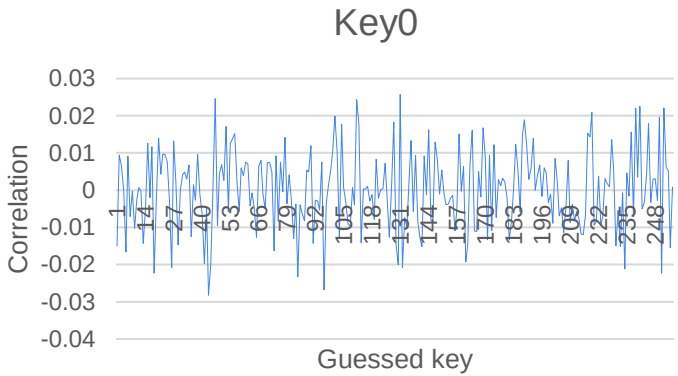
Coalescing width	L1~DRAM	L2~DRAM	DRAM
64-byte width	≅100000	<40000	<10000
128-byte width	>100000	<60000	<20000

Generating Dummy Memory Requests





DRAM 사용, Dummy memory request 사용,
Correlation = 0.027



Conclusion

- 부채널 공격의 난이도에 영향을 주는 요소
  - 메모리 접합 크기로 인한 메모리 요청 횟수의 변화
  - 메모리 계층에 따른 메모리 접근 시간
- Dummy memory requests를 이용할 시 비례 관계 사라짐
- 메모리 요청 개수와 실행시간 사이의 비례관계를 없애며 performance overhead를 줄일 수 있는 HW/SW 구조 연구 필요

Thank you