Jetson nano 환경에서의 경량 암호학적 난수발생기 병렬 구현

박영재[2], 유현도[2], 강주성[1,2], 염용진[1,2]*
국민대학교 정보보안암호수학과[1] / 금융정보보안학과[2]
{dudwo9696, dbguseh111, jskang, *salt}@kookmin.ac.kr

A Parallel Implementation of Lightweight Cryptographic Random Number Generator in Jetson nano

Yeongjae Park[2], Hyeondo Yoo[2], Ju-Sung Kang[1,2], Yongjin Yeom[1,2]*
Dept. of Information Security, Cryptology, and Mathematics[1]/
Financial information security[2], Kookmin Univ.

요 약

사물인터넷의 사용량 증가에 따라 최근 경량 암호의 중요성이 부각되고 있으며, 양자 키 분배 프로토콜에서 단시간에 많은 난수를 요구하고 있다. 본 논문에서는 경량 암호학적 난수발생기인 CHAM-CTR-DRBG를 경량 환경 Jetson nano에 병렬 구현하여 분석한다. 먼저 GPU 최적화 구현에 필요한 도구인 CUDA Occupancy Calculator의 결과를 분석한다. 다음으로, 기 연구된 PC 환경에서의 CHAM-CTR-DRBG GPU 구현된 코드를 경량 환경에서 GPU 구현하여 성능 분석한 결과를 중심으로 논한다. 마지막으로, 구현된 CHAM-CTR-DRBG가 경량 환경에서 효율적으로 동작하였음을 확인한다.

Ⅰ. 서 론

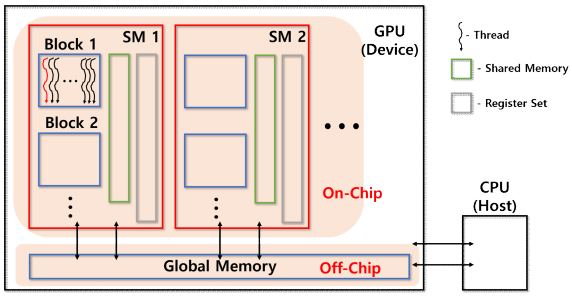
사물인터넷(IoT)은 4차 산업혁명의 핵심 기술로 CCTV, 가전 제품 등 다양한 사물 간 정보의 교환과 원격 제어가 가능해져 최근 각광받고 있다. 하지만 시장 규모가 날로 증가함에 따라 임베디드 기기 간 개인정보의 안전한 교환이 요구되고 있다. 동시에 임베디드 기기에 적용 가능한 경량 암호 알고리즘에 대한 연구도 지속적으로 이루어지고 있으며, 최근에는 양자컴퓨터 공격 관점에서 경량 암호 알고리즘이 분석되고 있다[1]. 특히, 경량 암호에 사용할 난수가 필요함에 따라, 경량 환경에서 암호학적으로 안전한 난수발생기에 대한 연구도 중요해지고 있다.

BB84 프로토콜은 양자키분배(QKD) 기술 중 이론적으로 안전성이 가장 잘 증명된 프로토콜이다. 위 프로토콜은 펄스 전송 속도에 따라 많게는 초당 1GB의 난수를 공급해야 한다. 한편, 많은 난수를 공급하기 위해 경량 의사난수발생기를 PC 환경에서 병렬 구현하는 연구도 이루어졌다[2].

본 논문은 위 연구[2]의 구현 결과물인 경량 블록암호 CHAM을 암호학적 의사난수발생기 CTR-DRBG에 적용한 CHAM-CTR-DRBG를 경량 환경 Jetson nano에 구현해 분석한다. 먼저, 알고리즘을 최적화 구현하기 위해 CUDA Occupancy Calculator가 계산하는 GPU 하드웨어 자원에 대해 면밀히 분석한다. 이후, 경량 환경에서의 난수 생성 속도를 분석하고, GPU가 50%와 100% 자원을 사용할 때 커널 함수의 성능을 비교한다. 끝으로, 경량 환경에서 난수발생기가 효율적으로 동작하였음을 확인한다.

Ⅱ. CUDA Occupancy Calculator

NVIDIA는 GPU용 연산 라이브러리인 CUDA를 제공하고 있다. CUDA 라이브러리를 활용하여 알고리즘을 최적화 및 병렬화하면 CPU보다 효율적으로 연산을 수행할 수 있다. [그림 1]에 CUDA가 GPU(Device)의 메모리를 GPU 내부에서만 접근 가능한 온칩(On-chip) 메모리와 CPU(Host)도 접근 가능한 오프칩(Off-chip) 메모리로 나누어 관리하는 것을 나타냈다. 사용자는 알고리즘의 성능 향상을 위해 최대한 온칩 메모리를 활용하고, 커널(Kernel) 함수의 파라미터인 블록(Block)과 스레드(Thread) 개수를 조정하여 SM(Streaming Multiprocessor)에 가능한 많은 와프(Warp)를 할당해야한다.



[그림 1] GPU의 온칩 메모리와 오프칩 메모리

와프는 32개의 스레드로 구성되는 GPU의 최소 실행 단위이다. 각 SM에 많은 와프를 지정하기 위해 할당 가능한 최대 와프와 최대 스레드 수를 고려해야하는데, 이때 사용가능한 도구가 바로 CUDA Occupancy Calculator이다. 해당 도구를 이용하여 GPU의 SM에 할당 가능한 최대 와프 수와 할당된 와프 수를 비율로 계산한다[3]. 위 비율을 계산하는 연산은 식 (1)과 같다.

$$(\text{Blocks per SM}) * (\text{Warps per block}) = \text{Warps per SM} \quad (1)$$

사용자가 Occupancy Calculator에서 조정 가능한 자원은 [표 1]과 같다. 먼저 GPU의 물리적인 코어, SM의 개수 등을 정하기 위해 Compute Capability를 확인하여 입력한다. 다음으로 알고리즘의 커널 함수 동작에 사용할 블록 당 스레드(Threads per block), 스레드 당 레지스터(Register per thread), 블록 당 공유 메모리(Shared memory per block)를 입력하면 자동으로 식 (1)을 계산한다. 이때, 값이 100%에 가까울수록 SM에 많은 와프를 할당한 것이고, GPU가 효율적으로 동작함을 의미한다.

[표 1] CUDA Occupancy Calculator 입력 자원

1. Compute Capability (CC)		2.0 - 7.5	(NSight Compute) 8.0 - 9.0
2.	Threads per block	최대 1,024개 (2의 거듭제곱)	
	Registers per thread	최대 255개	
	Shared memory per block	최댓값은 CC에 따라 변동 (bytes)	

Ⅲ. 경량 암호학적 난수발생기 CHAM-CTR-DRBG

CTR-DRBG는 암호학적 난수발생기 중 하나로, 국제 표준 ISO/IEC 18031에 제시된 의사난수발생기이다[4]. CTR-DRBG는 엔트로피 소스(entropy source)를 입력하면 내부갱신함수(update function)와 출력생성함수(generate function)를 통해 의사난수를 생성한다[5].

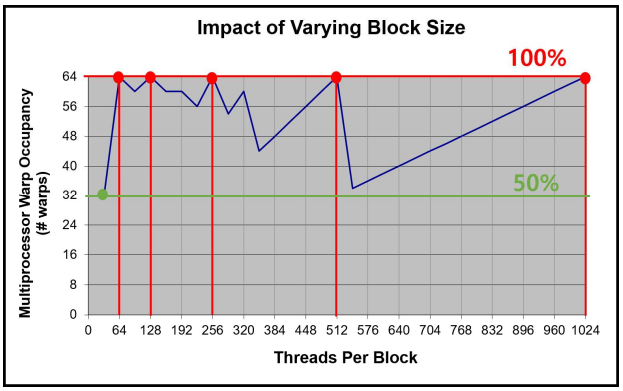
블록암호 CHAM은 2017년 국가보안기술연구소에서 개발하였다. CHAM의 경우에는 메모리를 총 640비트만 사용하는 반면, AES 32비트 구현의 경우 총 36,688비트의 메모리를 사용한다. 추가로, 경량 환경에서 CHAM이 AES보다 암호화 속도가 약 3.5배 빨라 블록암호 CHAM이 AES보다 경량 환경에 적합함을 보였다[5].

CHAM-CTR-DRBG는 출력생성함수에 경량 블록암호 CHAM을 CTR 방식으로 적용하는 난수발생기를 말한다. 유현도 등 3명은 경량 난수발생기 CHAM-CTR-DRBG의 출력생성함수를 PC 환경(NVIDIA TITAN Xp)에서 병렬화 구현하였다. 위 저자는 16 bytes의 입력 평문 V와 32 bytes의 라운드 Key를 공유 메모리에 저장하고, 하나의 스레드가 한 번의 암호화를 실행하도록 구성하였다. 결론적으로, 스레드 당 26개의 레지스터와 48 bytes를 공유 메모리로 사용하여 고속 구현하였다[2].

Ⅳ. 경량 환경에서의 구현 및 실험

기존에 PC 환경에서 GPU 병렬 구현한 CHAM-CTR-DRBG[2]를 경량 환경 Jetson nano에 병렬 구현하였다. 경량 환경에서도 3장과 동일하게 스레드 당 26개의 레지스터와 48 bytes의 공유 메모리를 할당하였다. 이후 경량 환경에 구현된 알고리즘의 커널 함수에 사용되는 최적의 블록과 스레드 개수를 찾기 위해 Occupancy Calculator를 사용하였다.

실험 환경인 Jetson nano의 GPU 사양은 Compute Capability 5.3이며, 128개 코어를 가지는 Maxwell GPU이다[5]. [그림 2]는 Jetson nano GPU에서 블록 당 스레드 수에 의해 변동하는 와프의 수를 분석한 결과이다. 해당 그림을 통해 실험 환경에서 CHAM-CTR-DRBG를 동작할 때, 블록 당 스레드 값을 32로 적용하면 64, 128, 256, 512, 1024 스레드를 사용하였을 때보다 성능이 낮을 것으로 예상할 수 있다.

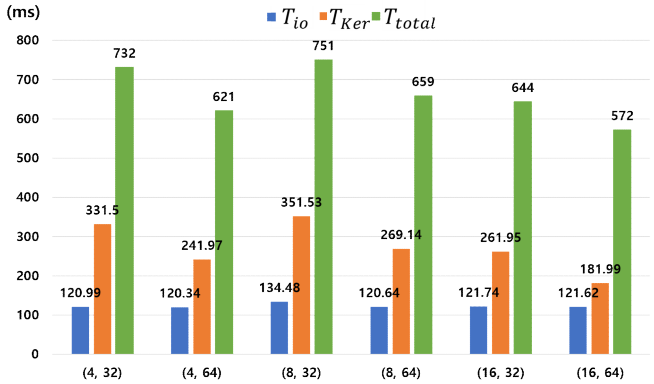


[그림 2] 블록 당 스레드에 따른 SM당 와프 개수

추가로, 경량 난수발생기의 최대 난수 생성 속도를 구하기 위해 커널 함수의 블록은 4, 8, 16, 32, 64, 128개, 스레드는 32, 64, 128, 256, 512개로 총 30가지로 나눠 실험하였다. 그 결과, 최대 속도는 16MB, 32MB 난수를 출력할 때 약 3Gbps 속도를 달성하는 것을 확인하였다.

앞서 확인한 [그림 2]의 내용을 검증하기 위해 32개의 스레드, 64개의 스레드인 경우에 난수발생기의 동작 시간을 분석하였다. [그림 3]은 커널 함수가 각각 4, 8, 16개의 블록과 32, 64개의 스레드인 경우에 16MB 난수 10개를 출력하여 시간을 측정한 결과이다. T_{io}는 CPU와 GPU 간 데이터 총 이동 시간, T_{Ker}은 10번의 커널 함수 동작 시간, T는 T_{io}와 T_{Ker}을 제외하고 오버헤드를 포함한 나머지 동작시간을 나타내는 값이다. 다시 말해, 알고리즘의 총 실행시간 T_{total}은 식 (2)와 같다.

$$T_{total} = T + T_{io} + T_{Ker} \quad (2)$$



[그림 3] 16MB 난수 10개 생성 시간 측정 결과

커널 함수가 64개 스레드에서 효율적으로 동작함을 확인하기 위해 T_{Ker}/T_{total} 값을 비교한다. 각각 4개 블록일 때 0.45 → 0.39로 13%, 8개 블록일 때 0.47 → 0.41로 13%, 16개 블록일 때 0.41 → 0.32로 22%만큼 커널 함수의 성능이 증가함을 확인하였다.

Ⅴ. 결 론

본 논문은 CHAM-CTR-DRBG를 경량 환경인 Jetson nano에 병렬 구현하여 성능을 분석하였다. 이전 연구의 결과로, 32비트 구현 기준 AES가 CHAM의 약 57배에 달하는 메모리를 사용하고, 속도는 CHAM이 약 3.5배 빨라 CHAM은 경량 환경에 적합함을 보였다. CHAM-CTR-DRBG에 사용되는 48 bytes 크기의 Key와 V를 공유 메모리로 사용하여 실험 분석한 결과, 약 3Gbps 속도를 달성함을 확인하였다. 특히, 커널 함수는 64개의 스레드를 사용하였을 때 32개의 스레드보다 최대 22%가량 성능이 향상됨을 보였다. 이전의 PC 환경(TITAN Xp)에서의 실험은 최소 2.7, 최대 7.9Gbps였음을 감안하면, 경량 환경인 Jetson nano에서 GPU 난수발생기가 준수한 성능을 가진다는 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-기후변화대응기술개발 사업의 지원을 받아 수행된 연구임(No.2021M1A2A2043893)

참 고 문 헌

[1] 장경배, 김현지, 송경주, 양유진, 임세진, 서화정, “사물인터넷을 위한 경량암호와 양자컴퓨터.,” 한국정보보호학회지, 2022.

[2] 유현도, 강주성, 염용진, “경량 암호 CHAM을 사용한 암호학적 난수발생기 GPU 병렬 구현.,” 한국통신학회 학술대회논문집, 2021.

[3] https://docs.nvidia.com/cuda/cuda-occupancy-calculator/index.html

[4] ISO/IEC, “ISO/IEC 18031:2011(E).,” International Standard, 2011.

[5] 유현도, 임형신, 강주성, 염용진, “경량 암호 CHAM을 사용한 암호학적 난수발생기 구현.,” 한국통신학회 학술대회논문집, 2021.

[6] https://developer.nvidia.com/cuda-gpus#compute