ARM 기밀 컴퓨팅 아키텍처에 대한 성능 분석

서영욱*, 황명하, 주정현, 강희운, 권유진, 이현우**
*KENTECH (학부생), **KENTECH (교수), 한국전력공사 전력연구원

Performance Analysis of ARM Confidential Computing Architecture

Yeongwook Seo*, Myeong-Ha Hwang, Jeonghyun Joo,
Heewoon Kang, YooJin Kwon, Hyunwoo Lee**
*KENTECH (Undergraduate student), **KENTECH (Professor), KEPRI

요 약

엣지 디바이스 상에서 AI 서비스를 제공하는 온디바이스 AI 관련 시장이 빠르게 성장하고 있다. 다수의 가까운 엣지 디바이스가 제공하는 저지연 AI 서비스는 사용자의 일상 속 효율성과 만족도를 높일 것으로 기대된다. 하지만 엣지 컴퓨팅은 넓은 공격 표면 노출을 야기하기에 이를 해결할 수 있는 하드웨어 보안 기술인 ARM CCA 적용을 고려할 필요가 있다. 특히나 CCA가 제공하는 보안 영역에서 AI 모델을 실행하게 되면 모델의 파라미터를 보호할 수 있기에, 엣지 디바이스 상에 온디바이스 AI를 안전하게 구현할 수 있다. 이러한 보안 시스템을 현실화하기 위해서는 모델의 실행 과정에서 발생하는 비보안 영역과 보안 영역의 통신이 최소화 될 필요가 있다. 본 연구에서는 CCA 에뮬레이션 상에서 비보안 영역과 보안 영역 사이 전이 시간 지연을 일반 노트북 대비 저사양 엣지 디바이스에서 측정하였고, 기기 성능 차이가 전이시간에 미치는 영향이 거의 없다는 것을 확인하였다.

I. 서론

엣지 컴퓨팅은 사용자에게 사용자와 가까운 기지국이나 WiFi 액세스 포인트(AP, Access Point) 같은 엣지 디바이스에서 데이터 스토리지와 AI 서비스를 제공함으로써 클라우드 컴퓨팅에 비해 저지연 서비스를 제공한다 [1]. AI 기술이 고도화되면서, AI 모델을 말단 디바이스 위에서 돌리는 온디바이스 AI 개념이 등장하면서 관련 시장도 급속히 성장할 것으로 예상된다 [2]. 이에 따라 엣지 디바이스 위에서 AI 모델이 돌아가는 것도 조만간 현실화 될 것이다. 이러한 온디바이스 AI는 사용자 친화적이고 지능적인 서비스를 제공하여 삶을 윤택하게 만들 것으로 예상된다.

이러한 서비스의 성공을 위해서는 엣지 디바이스에서 AI 모델에 대한 보안성이 보장되어야 한다. 특히나 AI 모델 제공자가 모델 파라미터의 유출을 우려하기에, 이에 대한 대비책이 요구된다. 하지만 AI 모델이 상대적으로 저사양인 다수의 근접 엣지 디바이스에 탑재되어 넓은 공격 표면에 노출되기 때문에, 엣지 디바이스 자체의 신뢰성을 높일 방안이 필수적이다 [3].

이를 위해 엣지 디바이스에 메모리 암호화, 기밀 부팅, 워크로드 인증 등을 포함하는 하드웨어 보안 기술인 ARM Confidential Computing Architecture (CCA) 적용을 고려할 수 있다 [4]. ARM CCA 구조에서는 모델을 안전하게 실행하기 위해 비보안 영역으로부터 격리된 신뢰 실행환경인 보안 영역(realm)을 제공하고, 이 보안 영역 내에 모델을 로드하여 모델 파라미터의 유출을 방지한다 [4]. 따라서 비보안 영역의 데이터가 보안 영역으로 전송되는 과정이 필연적이다. 이 전이 시간이 신뢰 실행을 현업화하는데 중요한 고려 사항이 될 것으로 보이며, 특히나 저사양인 엣지 디바이스에서도 큰 부하가 되어서는 안된다. 이 때문에 본 논문에서는 비보안 영역과 보안 영역 간 전이 시간을 기기별로 비교 및 분석하고자 한다.

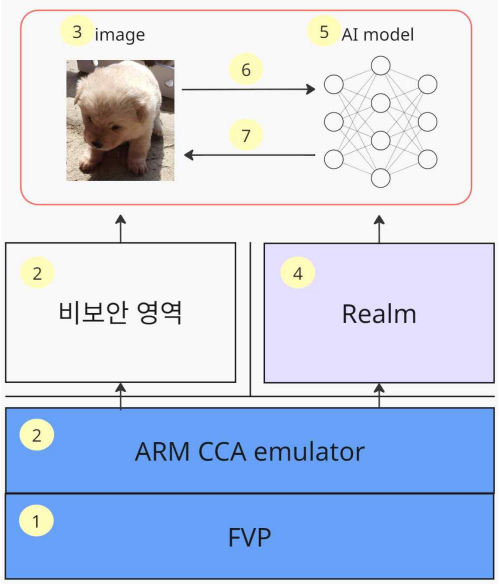

그림 1 실험 시나리오

II. 실험 환경

실험을 수행하기 위해 본 연구진은 다음의 시나리오를 고려하였다 (그림 1 참조). 엣지 디바이스에 이미지 분류를 수행하는 AI 모델을 활용하여 분류 작업을 수행하며, 이때 AI 모델을 보호하기 위해 엣지 디바이스에 CCA를 활용하였다. CCA는 ARMv9 이후로 하드웨어 기반으로 구현될 예정이기 때문에, 에뮬레이터로 구현된 CCA를 활용하여 기기 성능에 따른 전이 시간을 측정하고자 하였다.

ARM CCA 에뮬레이션 환경에서 비보안 – 보안 영역간 전이시간을 측정하지만 실제 하드웨어 성능과의 오차를 최소화하기 위해 ARM에서 제공하는 ARM Fixed Virtual Platform (FVP)를 사용했다. 이는 하드웨어에 가능한 가까운 속도로 메모리, 프로세서, IO 디바이스 등의 ARM 시스템을 시뮬레이션을 수행하도록 돕는다 [5]. 이러한 ARM FVP 환경 위에 ARM에서 제공하는 CCA 에뮬레이터를 결합하였다. 이는 ARMv9 구조를 기반으로 비보안 영역과 보안 영역을 제공한다 [3]. 해당 에뮬레이터는 비보안 영역에서 Kernel based Virtual Machine (KVM) 을 활용하여 보안 영역을 구성한다.

이 위에서 이미지 추론 모델을 로드하였다 [6]. 비보안 – 보안 영역의 전이 시간을 측정하기 위해 다음과 같은 과정을 40 차례 반복 수행한다 (그림 2 참조).

1) 이미지 하나를 비보안에서 보안 영역으로 전이
2) 보안 영역에서 대기 상태이던 모델은 이미지를 받으면 추론 수행
3) 모델은 추론 결과를 비보안 영역에 전달
4) 남은 이미지에 대해 1)부터 재수행

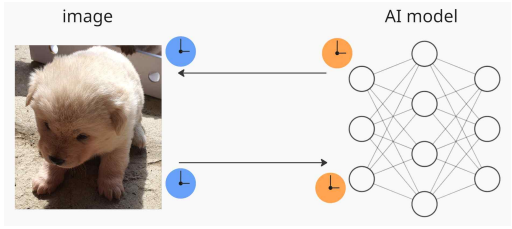

그림 2 응용 시나리오

비보안 영역에서 이미지를 보낸 시점과 모델의 추론 완료 신호를 받은 시점의 차이로 두 번의 기기간 전이 시간과 모델의 추론 시간의 합을 측정한다. 또한, 보안 영역에서 이미지를 받은 시점과 모델 완료 신호를 보낸 시점의 차이로 모델의 추론 시간을 구하여 두 측정값의 차이를 통해 전이 시간을 측정하였다.

비보안 영역과 모델 사이의 전이 및 전달은 Realm Management Monitor (RMM)가 보안 영역과 하이퍼바이저 사이의 통신에 관여하기에 시간 지연이 추가로 발생한다 [4]. 비보안에서 보안 영역으로 접근하는 것이 보안 영역의 격리성을 침해하여 보안상 취약점을 야기할 수 있기에 RMM이 관여한다.

III. 실험 결과

우리는 라즈베리파이 5를 엣지 디바이스로 선택하였고, 이것의 전이 시간을 두가지 종류의 일반 랩탑 기기의 전이 시간과 비교하였다. 각각의 기기에 대하여 실험 환경에서 언급한 에뮬레이터를 설치하여 실험하였다.

기기	평균 (초)	표준편차	clock*
NT550XED	2.567740	1.275026	4.0 GHz
NT950--QED	2.500793	1.276080	2.3 GHz
Raspberry Pi 5	2.528082	1.123956	2.4 GHz

표 1. 기기별 전이시간

위 표는 실험 시나리오 수행 결과로 얻은 기기별 전이 시간의 평균, 표준편차와 기기 스펙을 나타낸다. 라즈베리파이 5는 NT550XED에 비해 전이 시간이 1.6% 빨랐고, NT950QED에 비해서는 1.1% 느렸

* 시뮬레이션 실행 초기 cpu 클럭수 평균

다. 두 가지 일반 랩탑기기 간의 전이시간 차이는 2.7% 정도였다. 이는 단순한 구조의 모델을 에뮬레이터로 구현된 CCA 상에서 실행하는 경우 CPU 성능 차이에도 불구하고 전이 시간에서는 성능 차이가 거의 없다는 것을 보여준다.

IV. 결론

본 논문에서는 엣지 디바이스로부터의 AI 모델 유출을 방지하기 위해 ARM CCA 사용을 고려하며, 저사양의 엣지 디바이스 상에서 CCA가 큰 부하를 초래하는지 보고자 하였다. 우리 실험의 결론은 CCA가 특별한 부하를 초래하지 않아 안전한 저지연 서비스가 가능하다는 것을 보였다. 하지만 여전히 하드웨어 구현물에서 예상 못한 부하가 발생할 수 있기 때문에 이에 관한 추가적인 연구가 필요할 것이다. 향후 연구에서는 ARMv9 아키텍처를 탑재한 실제 하드웨어 상에서의 모델 실행 시간 측정을 통해, 엣지 디바이스에서의 저지연 서비스 구현 가능성을 더욱 구체화할 수 있을 것으로 기대된다.

Acknowledgment

본 연구는 한국전력공사의 2022년 착수 기초연구개발 과제 연구비에 의해 지원되었음(과제번호: R24XO01-3)

[참고문헌]

[1] K. Zhang, S. Leng, Y. He, S. Maharjan and Y. Zhang, "Mobile Edge Computing and Networking for Green and Low-Latency Internet of Things," in IEEE Communications Magazine, vol. 56, no. 5, pp. 39-45, May 2018

[2] Korea Copyright Commission, 온디바이스 AI 산업 현황 보고서, https://www.copyright.or.kr/information-materials/trend/tmis/view.do?brdctsno=53345, accessed May 12, 2025.

[3] S. Siby, S. Abdollahi, M. Maheri, M. Kogias, and H. Haddadi, “GuaranTEE: Towards Attestable and Private ML with CCA,” in Proceedings of the 4th Workshop on Machine Learning and Systems, Athens Greece: ACM, April. 2024

[4] Arm Ltd., Learn the architecture - Introducing Arm Confidential Computing Architecture, https://developer.arm.com/documentation/den0125/400/Overview, accessed May 12, 2025.

[5] Arm Ltd., Fixed Virtual Platforms Reference Guide, https://developer.arm.com/documentation/100966/latest/, accessed May 12, 2025.

[6] comet-cc, TFlite model, GitHub repository, https://github.com/comet-cc/TFlite-CCA, accessed May 12, 2025.