CUDA MPS 기반 GPU 자원 분배가 Multi-cell 5G PHY 계층 처리 성능에 미치는 영향 분석

전가겸, 나지현, 김남이
한국전자통신연구원
{jkk83, jhna, namikim}@etri.re.kr

Performance Analysis of GPU Resource Allocation Based on CUDA MPS in Multi-cell 5G PHY Layer Processing

Kakyeom Jeon, Jee-Hyeon Na, Nam-I Kim
Electronics and Telecommunications Research Institute

요 약

AI-RAN(Artificial Intelligence Radio Access Network)은 인공지능을 활용하여 네트워크 자원 관리 및 성능 최적화를 목표로 하며, 이를 위한 GPU 가속과 cloud-native 방식의 SW-defined RAN으로의 전환이 요구되고 있다. 본 연구에서는 NVIDIA의 CUDA 기반 AI Aerial 플랫폼을 활용하여, GPU에서 PHY layer 기능을 수행하는 cuPHY 모듈의 성능을 분석한다. 특히, CUDA MPS(Multi-Process Service)를 통해 GPU의 SMs(Streaming Multi-processors)를 논리적으로 분할하고, 시뮬레이션을 통해 Multi-cell 5G PHY layer 처리 시 자원 분배가 시스템 지연 시간(latency)에 미치는 영향을 평가한다.

Ⅰ. 서론

최근 RAN(Radio Access Network)에 대한 연구가 활발히 진행되면서, AI-RAN(Artificial Intelligence RAN) 기술에 대한 관심이 급증하고 있다. AI-RAN은 인공지능 기술을 활용하여 무선 네트워크의 자원 관리, 성능 최적화, 에너지 효율 개선 등을 달성하는 것을 목표로 한다[1]. 이를 위해 기존의 HW 중심의 RAN 구조를 벗어나 유연하고 확장 가능한 SW-defined RAN으로 전환이 요구된다. 이러한 AI-RAN의 목적을 달성하기 위해, GPU 가속과 cloud-native 방식을 채택하여 L1과 L2의 fully in-line GPU 가속을 지원하는 SW 기반의 RAN 구조가 제안되고 있다. 기존의 전용 HW 기반의 가속기와 달리, SW-defined 방식은 HW 교체 없이 단순한 SW 업데이트만으로 시스템을 개선할 수 있는 높은 유연성과 확장성을 제공한다. 이를 통해 비용 절감과 네트워크 인프라의 수명 연장이라는 실질적인 이점을 확보할 수 있다.

NVIDIA는 이러한 기술적 흐름에 맞춰, GPU 가속 기술을 활용한 CUDA 기반의 AI Aerial 플랫폼을 제안하였다[1]. AI Aerial 플랫폼은 cuPHY와 cuMAC의 SW-defined 모듈을 통해 L2 계층 이하의 RAN 기능을 GPU에서 수행할 수 있도록 설계되었다. 특히, 계산 집약적인 신호 처리 작업을 GPU로 off-load 함으로써, 별도의 HW 확장 없이도 높은 처리 성능을 제공할 수 있다. 이를 통해 AI-RAN 환경에서 요구되는 고성능 처리 능력을 충족시키고, 스펙트럼 효율성 향상, 사용자 증가 대응, 광대역 서비스 제공 등의 미래 지향적인 네트워크 확장을 가능하게 한다. 이러한 기술의 발전은 AI를 활용한 네트워크 예측, load balancing, 트래픽 제어 등의 지능형 네트워크 관리 기능을 실현할 수 있으며, 기존 RAN 환경에서는 구현하기 어려웠던 스펙트럼 관리, 간섭 완화, QoE(Quality of Experience) 최적화와 같은 새로운 접근 방식을 지원할 수 있을 것으로 기대된다.

본 논문에서는 NVIDIA Aerial 플랫폼의 High-PHY layer를 구현하는 핵심 모듈 cuPHY를 활용하여, Multi-cell 환경에서 5G PHY layer 처리를 수행할 때 CUDA MPS(Multi-Process Service) 기반 GPU 자원 분할이 시스템 성능에 미치는 영향을 분석한다. 특히, GPU의 SMs(Streaming Multi-processors) 자원을 논리적으로 분할하여 할당함에 따라, 5G PHY layer의 주요 채널들을 처리하는 과정에서 발생하는 지연 시간(latency)이 어떻게 변화하는지 시뮬레이션을 수행하고, 이를 통해 AI-RAN 환경에서의 GPU 자원 최적화 가능성을 논의한다.

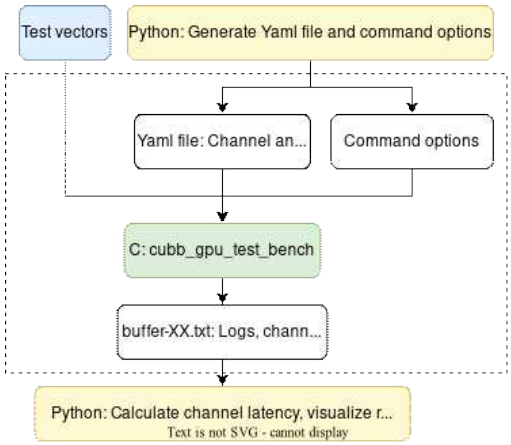

그림 1. cuPHY 성능 측정을 위한 테스트 벤치 프레임워크

Ⅱ. 본론

본 논문에서는 NVIDIA AI Aerial에서 제공하는 *aerial_sdk/testBenches*를 활용하여, multi-cell/multi-channel 환경에서 CUDA 기반 MPS(Multi-Process Service)를 동작시켜 cuPHY의 독립적인 성능을 평가한다. MPS는 NVIDIA GPU에서 여러 CUDA 애플리케이션 또는 스레드가 동시에 GPU를 공유할 수 있도록, GPU의 핵심 연산 유닛인 SMs를 논리적으로 분할하여 사용할 수 있게 하는 스케줄링 시스템이다. 이를 통해 여러 채널이 하나의 GPU를 사용할 때, SM 자원을 효율적으로 할당받아 CUDA kernel의 병렬 실행이 가능하도록 한다.

성능 테스트는 그림 1과 같이 진행되며, NIC(Network Interface Controller)나 L2 layer와의 데이터 교류가 없고, TV(Test Vector)와 Yaml file, 그리고 명령어 옵션을 입력으로 사용하여 cuPHY의 고유 성능을 검증한다. 테스트 명령어를 통해 GPU의 클럭 주파수, 전력 소모량, stream/graph 모드, 셀 및 슬롯 수 등 기본 환경을 설정할 수 있으며, 각 채널 [PRACH, PDCCH, PUCCH, PDSCH, PUSCH, SSB]별로 sub-context를 생성하여 요청하는 SM 개수를 지정할 수 있다. 이를 통해 SM 자원 사용량을 조절하고, 성능과 자원의 trade-off를 최적화할 수 있다. 시뮬레이션 결과는 TV 로그, 각 채널의 시작 및 종료 시각, 디버그 정보가 포함된 buffer-XX.txt 파일로 생성되며, 이를 분석하여 각 채널의 latency 및 자원 사용 특성을 평가한다.

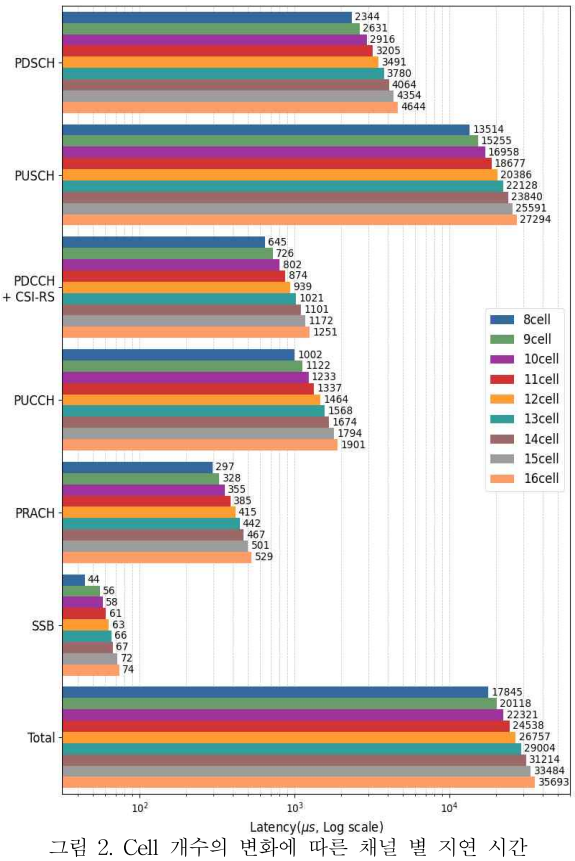

그림 2. Cell 개수의 변화에 따른 채널 별 지연 시간

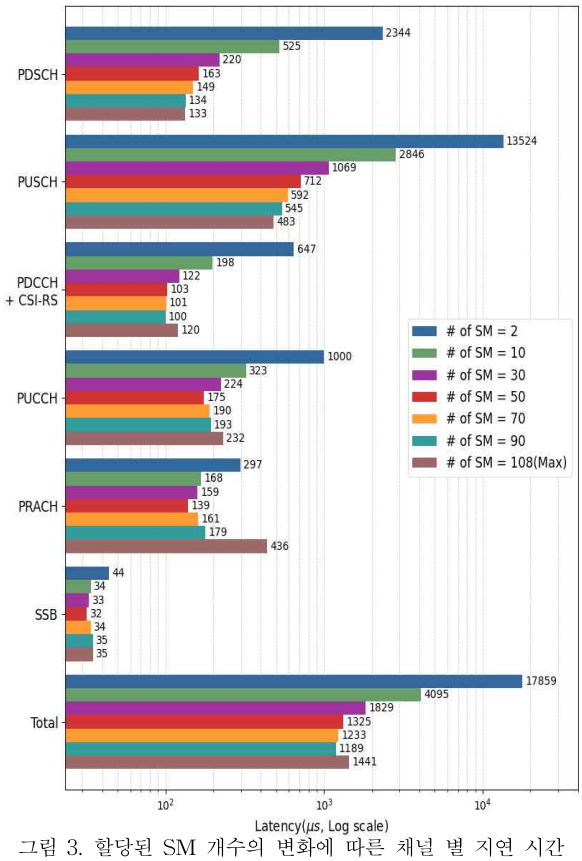

그림 3. 할당된 SM 개수의 변화에 따른 채널 별 지연 시간

표 1. 시뮬레이션 환경 및 파라미터

Parameters	Value
Server	Dell R760
CPU	Intel Xeon Platinum 8580
GPU	A100X
NVIDIA Aerial Release	24-1
GPU Power	250W
GPU Clock Frequency	1530MHz
Iteration	1000
Number of Cells	[8, 9, 10, 11, 12, 13, 14, 15, 16]
Number of SMs	[2, 10, 30, 50, 70, 90, 108]
Running Mode	Stream mode

Ⅲ. 시뮬레이션 결과

본 시뮬레이션에서는 표 1과 같이 NVIDIA가 명시한 시뮬레이션 환경을 구축하고, NVIDIA Aerial SDK가 제공하는 사전 정의된 테스트 케이스를 활용하였으며, 4T4R 기반의 슬롯 패턴(F08, TDD: DDDSUUDDDD) 하에서 SM 분배에 따른 성능 변화를 분석하였다.

그림 2는 SM 수를 2로 고정한 상태에서 셀 수를 8개에서 16개 까지 증가시킬 때, 각 채널의 지연 시간 변화를 나타낸다. 셀 수 증가에 따라 전체 지연 시간도 점진적으로 증가하였으며, 특히 PDSCH와 PUSCH 채널에서 현저하게 높은 지연 시간이 관측되었다. 이는 두 채널이 LDPC 디코딩, MIMO 처리, 변조/복조 채널 추정 등의 고부하 연산을 수행함으로써, 상대적으로 많은 SM 자원이 있어야 하기 때문이다.

그림 3은 셀 수를 8개로 고정한 상태에서 각 채널에 할당된 SM 수를 2에서 108까지 변화시킨 경우의 지연 시간 결과를 보여준다. 시뮬레이션 결과 전체 작업 시간은 SM이 최대 개수(108)가 아닌 90개일 때 가장 짧은 지연 시간을 보였다. 이는 PDSCH와 PUSCH는 SM 할당이 증가할수록 지연 시간이 급격히 감소하는 반면, 그 외 PDCCH, PUCCH 등의 채널은 SM이 일정 수준 이상으로 과도하게 할당되었을 때 오히려 지연 시간이 증가한 것에 의한 결과로 확인되었다. 이는 낮은 연산 부하의 채널에 대해 과도한 양의 SM이 할당되어 오버헤드, 메모리 접근 지연 등의 병목 현상을 유발하여 나타난 것으로 해석된다.

결과적으로, 채널 특성에 따른 SM 자원의 균형 잡힌 분배가 전체 시스템의 성능을 높이고 지연 시간을 줄이는 데 중요한 역할을 한다는 점을 확인할 수 있다.

Ⅳ. 결론

본 연구에서는 NVIDIA AI Aerial 플랫폼의 cuPHY 모듈을 기반으로, CUDA MPS를 활용한 GPU 자원 분할이 Multi-cell 5G PHY 계층 처리 성능에 미치는 영향을 분석하였다. 시뮬레이션을 통해, SM의 자원 할당 방식이 각 채널의 지연 시간에 미치는 영향을 정량적으로 평가하였다.

시뮬레이션 결과, 고연산 부하 채널은 할당된 SM 수가 증가할수록 지연 시간이 현저히 감소하며, 충분한 SM 자원을 확보하는 경우 효율적인 병렬 처리를 통해 성능 최적화가 가능함을 확인하였다. 반면, 저연산 부하 채널은 과도한 SM 자원 할당 시 오히려 오버헤드 및 자원 경쟁으로 인해 지연 시간이 증가하는 병목 현상이 발생하였다. 이는 채널 특성에 따라 적절한 자원 분배 전략이 필수적임을 의미한다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2024-00396828, AI 기반 저전력 5G-A O-DU/O-CU 기술 개발)

참 고 문 헌

[1] NVIDIA Corporation, Aerial CUDA-Accelerated RAN Documentation, Release 24-1, Jul. 2024. [Online]. Available: https://docs.nvidia.com/aerial/archive/cuda-accelerated-ran/24-1/index.html

[2] N. A. Khan and S. Schmid, "AI-RAN in 6G Networks: State-of-the-Art and Challenges," in IEEE Open Journal of the Communications Society, vol. 5, pp. 294-311, Jan, 2024