GPU 가속 양자 시뮬레이션의 확장성 및 자원 활용 분석

박은수, 테오도라 아두푸, 김윤희
숙명여자대학교 컴퓨터과학과

psrkeunsoo@sookmyung.ac.kr, theoadufu@sookmyung.ac.kr, yulan@sookmyung.ac.kr

Scalability and Resource Utilization Analysis of GPU-Accelerated Quantum Simulations

Eunsoo Park, Theodora Adufu, Yoonhee Kim
Dept. of Computer Science, Sookmyung Women's University

요 약

GPU 가속 양자 시뮬레이션 프레임워크는 양자 알고리즘의 성능 평가에 널리 활용되고 있으나, 전체 성능은 연산 성능뿐 아니라 데이터 이동 오버헤드와 메모리 효율성에 크게 좌우된다. 본 연구에서는 PennyLane 프레임워크를 기반으로 대표적인 양자 워크로드를 대상으로 이기종 GPU 환경에서의 양자 시뮬레이션 성능 및 자원 효율성을 분석하였다. 실험 결과, 큐비트 수가 증가함에 따라 계산 처리량보다 메모리 대역폭 및 전력 제한에 따른 병목이 실행 성능을 지배함을 확인하였다. 이는 대규모 GPU 기반 양자 시뮬레이션의 확장성 확보를 위해 메모리 및 에너지 효율 최적화가 핵심적임을 시사한다.

Ⅰ. 서 론

양자 컴퓨팅은 중첩과 얽힘을 활용하여 고전 컴퓨팅의 한계를 극복할 잠재 기술로 주목받고 있으나, NISQ(Noisy Intermediate-Scale Quantum) 장치는 노이즈와 규모 확장 측면에서 제약을 갖는다. 이에 따라 실제 양자 하드웨어 이전 단계에서 양자 알고리즘의 성능과 확장성을 평가하기 위해 cuQuantum 과 같은 GPU 가속 양자 시뮬레이션 프레임워크의 활용이 필수적이다.[1] 본 연구는 PennyLane 프레임워크를 기반으로 단일 GPU 환경에서 VQE, QPE, Grover 알고리즘을 대상으로 큐비트 수 증가에 따른 실행 성능을 분석하고, RTX 4090 과 NVIDIA A30 GPU 의 자원 활용 특성을 비교함으로써 양자 시뮬레이션 인프라 구축 시 효율적인 하드웨어 선택을 위한 실증적 근거를 제시한다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 실험 설계, 평가지표 및 알고리즘 특성과 실험 결과를 제시하고, 제 3 장에서 결론을 맺는다.

Ⅱ. 양자 시뮬레이션의 실험 설계 및 분석

A. 관련 연구

기존 연구에 따르면, 양자 회로 시뮬레이션의 성능 병목은 알고리즘 복잡도와 파라미터 구성(예: 큐비트 수, 회로 깊이)에 따라 상이하게 나타난다. 특히 HPC 환경에서는 연산 성능보다 데이터 이동 오버헤드와 메모리 효율성이 성능을 지배하는 요인으로 보고되고 있으나[2], GPU 기반 양자 시뮬레이션 환경에 대한 체계적인 분석은 아직 제한적이다.

B. 실험 설계

본 연구는 PennyLane 프레임워크 기반 단일 GPU 환경에서 양자 알고리즘의 확장성과 자원 효율성을 평가한다. 먼저, 큐비트 수 증가에 따른 성능 변화를 분석하기 위해 2-30 큐비트 범위에서 알고리즘별 총 실행 시간을 측정하였다. 이후, 이기종 GPU 환경에서의 자원 활용 특성을 비교하기 위해 28 큐비트 설정에서 각 알고리즘을 반복 실행하며 연산, 메모리, 전력 사용량을 모니터링하였다. 실험에는 FP16 기준 165 TFLOPS 연산 성능과 165W 전력 제한을 갖는 NVIDIA A30 과, 83 TFLOPS 의 연산 성능, 1,008 GB/s 메모리 대역폭, 450W 전력 한계를 갖는 RTX 4090 을 사용하였다. 이를 통해 GPU 의 연산 성능, 메모리 대역폭, 전력 제한이 양자 시뮬레이션 확장성에 미치는 영향을 분석한다.

C. 선택한 양자 알고리즘

본 연구에서는 하드웨어 자원 변화에 따른 성능을 다각도로 분석하기 위해 연산 특성이 상이한 세 가지 양자 알고리즘을 선정하였다.

i)VQE: 분자의 기저 상태 에너지를 탐색하는 데 활용된다[3]. 본 실험에서 큐비트 수의 증가는 타겟 시스템의 규모 확장을 의미한다. 반복적인 최적화 과정을 수반하므로 GPU 리소스 점유 변화를 분석하기에 적합하다.

ii)QPE: 유니터리 연산자의 고윳값 위상을 추정한다[4]. 본 실험에서 큐비트 수는 위상 추정의 정밀도를 결정하는 보조 큐비트 수이다. 시스템 규모 확장에 따른 GPU 리소스 점유 변화 분석 지표로 활용한다.

iii)Grover: 비정렬 데이터베이스 내에서 특정 상태를 탐색하는 확률 증폭 알고리즘이다[5].

D. 실험 결과 및 분석

i) 큐비트 수에 따른 알고리즘 확장성 분석

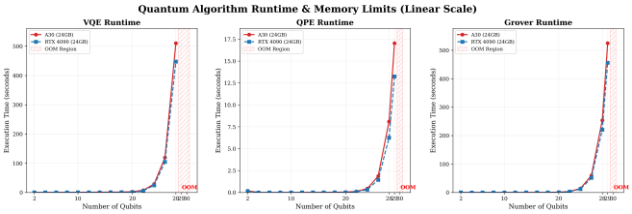

그림 1. RTX4090, A30:큐비트 수에 따른 실행시간 비교

모든 실험은 비교의 일관성을 위해 회로의 레이어 수를 100 으로 고정하여 수행되었다. 그림 2 의 실행시간 분석 결과, 모든 알고리즘에서 큐비트 수 증가에 따라 실행 시간이 지수적으로 상승하는 양상을 보였다. 이는 양자 상태 벡터 시뮬레이션의 계산 복잡도가 $O(2^n)$ 에

비례한다는 이론적 특성을 뒷받침한다. 메모리 한계로 인해 VQE 는 29 큐비트, QPE, Grover 는 30 큐비트에서 Out-of-Memory 가 발생한다.

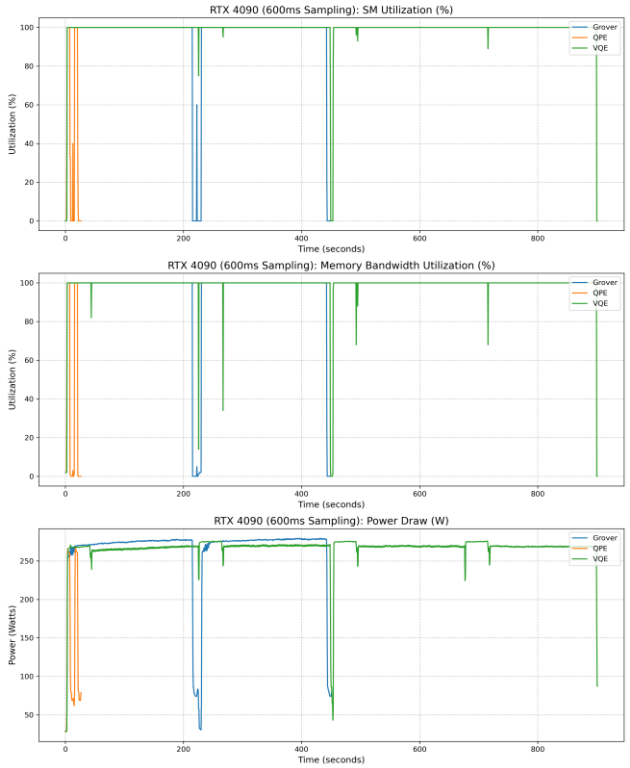

그림 2. RTX4090: 리소스 프로파일(Q28)

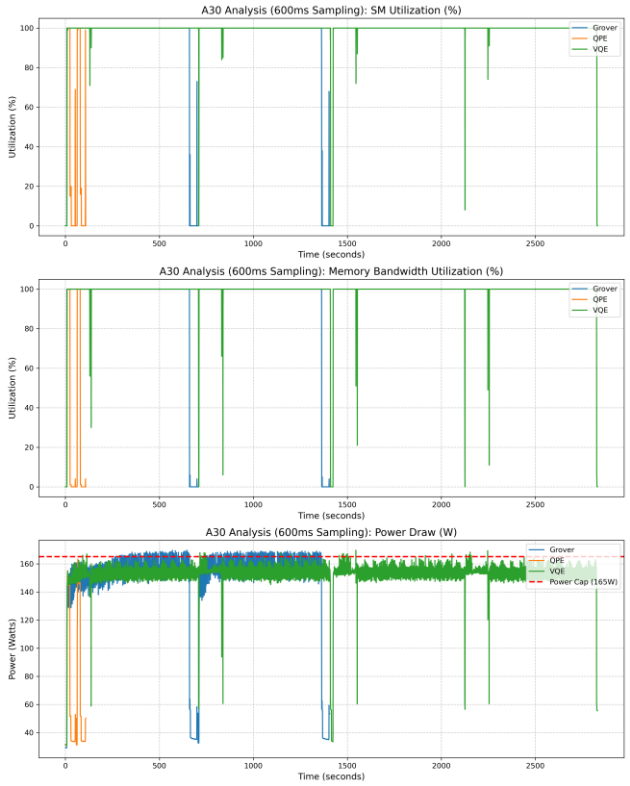

그림 3. A30: 리소스 프로파일(Q28)

ii) 이기종 GPU 에서의 자원 활용 특성 비교

그림 2 와 3 은 각 응용을 이기종 GPU 환경에서 2 번 반복 실행 후 도출된 리소스 프로파일이다. 29, 30 큐비트에서 OOM 이 발생하므로 안정적인 부하를 주기 위해 28 큐비트를 선정하였다. 응용의 1 회 실행 종료 시점에 SM 과 메모리 대역폭 사용률이 0 으로 떨어진다. 이론적인 연산 성능 측면에서 NVIDIA A30 은 FP16 기준 NVIDIA RTX 4090 대비 약 2 배 높은 수치를 보유하고 있다. 그러나 그림 1 의 실제 시뮬레이션 실행 속도는 RTX4090 이 전 구간에서 A30 을 상회하는 역전 현상이 발생한다. 본 연구에서는 이러한 성능 괴리의 원인을 다음 두 가지 하드웨어적 특성으로 분석하였다.

첫째, 메모리 대역폭의 영향이다. 양자 시뮬레이션은 대규모 행렬 연산을 수반하며, 이는 전형적인 메모리 집약적 워크로드의 특성을 가진다. RTX4090 은 1,008 GB/s 의 대역폭을 제공하여 A30(933 GB/s) 대비 초당 데이터 처리 성능이 우수하다. 따라서 이론적 연산 성능이 낮더라도, 데이터 전송 병목 현상을 보다 효율적으로 해소하는 RTX4090 이 실제 시뮬레이션 환경에서 더 높은 성능을 발휘한다.

둘째, 열 설계 전력에 따른 Throttling 현상이다. 그림 3 의 A30 전력 모니터링 결과, 큐비트 28 개로 실행시킨 결과 전력 소모량이 설계 임계치인 165W 에 즉각적으로 도달함을 확인하였다. 이후 연산 과정 전반에서 전력 소모량이 해당 임계치를 초과하지 못하고 포화 상태를 유지한다. 이는 A30 이 제원상 높은 연산 성능을 보유했음에도 불구하고 실제 시뮬레이션 속도에서 RTX4090 대비 열세를 보이는 원인으로 분석된다.

III. 결론

본 연구에서는 이기종 GPU 환경에서 큐비트 수 확장에 따른 주요 양자 알고리즘의 시뮬레이션 성능과 자원 활용 효율을 분석하였다. 실험 결과, 단순한 이론적 TFLOPS 수치보다 메모리 대역폭과 전력 공급 임계치가 실제 시뮬레이션 속도에 결정적인 영향을 미침을 확인하였다. 향후 대규모 양자 회로 시뮬레이션을 위한 최적의 하드웨어 가속기 선정 및 자원 할당 전략의 기초 자료로 활용될 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (과제번호: RS-2025-24534879).

참 고 문 헌

[1] “cuQuantum SDK: A High-Performance Library for Accelerating Quantum Science.” [Online]. Available: https://developer.nvidia.com/cuquantum-sdk

[2] Xin-Chuan Wu et al. “Full-state Quantum Circuit Simulation by Using Data Compression”. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC’19). New York, NY, USA: Association for Computing Machinery, 2019, pp. 1-24. doi:10.1145/3295500.3356155.url:https://doi.org/10.1145/3295500.3356155.

[3] Peruzzo, A., et al. (2014). “A variational eigenvalue solver on a photonic quantum processor.” Nature Communications, 5(1), 4213.

[4] Kitaev, A. Y. (1995). “Quantum measurements and the Abelian Stabilizer Problem.” arXiv preprint quant-ph/9511026.

[5] Grover, L. K. (1996). “A fast quantum mechanical algorithm for database search.” Proceedings of the 28th Annual ACM Symposium on Theory of Computing, 212-219.