공학 분야 오픈 소스 소프트웨어들의 GPU 계산환경 활용 동향 및 사용 방안

윤태호*
*한국과학기술정보연구원
e-mail:thyoon@kisti.re.kr

Trends and usage plans of GPU computing environments for open source software in the engineering field

Tae Ho Yoon*
*Center for Advanced Scientific Computing, Korea Institute of Science and Technology Information

요 약

본 논문에서는 최근 HPC와 함께 AI 계산 등 다양한 워크로드에 대응하기 위한 GPU 장착 HPC 시스템에 대한 공학 분야 오픈 소스 소프트웨어들의 GPU 솔버 개발 활동 및 이의 활용을 분석하였다. 이를 위해 공학 분야 길이 스케일 기준에 따른 전체 영역을 포괄할 수 있도록 총 6개의 오픈 소스 소프트웨어(Code_Aster, GROMACS, LAMMPS, OpenFOAM, Quantum ESPRESSO, SIESTA)들로 제한하여 검토하였다. 오픈 소스 소프트웨어 진영에서도 원자나 분자의 동역학 해석 분야로의 GROMACS, LAMMPS, Quantum ESSPRESSO 등의 적극적인 공식 GPU 솔버가 포함된 소프트웨어가 있는 반면 Code_Aster나 OpenFOAM 등 연속체 역학 분야에서는 GPU 솔버의 공식 지원이 미흡한 상황임을 확인하였다. 공학 분야 오픈 소스 소프트웨어들이 GPU 자원을 효과적으로 활용할 수 있는지를 사전 충분히 검토해 보기 위해 GPU 구축 방식을 확인하고 기존 구축된 GPU 기반 클러스터로 서비스 중인 뉴론에서의 해당 GPU 솔버 모듈들에 대해 테스트 하였다. 이를 통해 공식 또는 비공식적으로 적용 가능한 각 GPU 솔버들이 CPU 기반 솔버 대비 효율적인 계산 성능을 확보할 수 있음을 확인하였다. 슈퍼컴퓨터들의 점진적 GPU 기반 HPC 시스템으로 전환 추세임을 고려할 때 추후 연속체 역학분야 오픈 소스 소프트웨어 등에 대한 적극적인 공식 GPU 솔버 지원에 기여하는 한편 각 오픈 소스 소프트웨어별로 구현된 GPU 솔버들의 분석 결과를 바탕으로 자체적인 통합 GPU 개선 솔버로의 확장을 고려해 볼 수 있을 것이다.

1. 서론

최근 슈퍼컴퓨터는 전통적인 물리기반 계산과학 문제 해결 등을 위한 고성능컴퓨팅(HPC) 이외에 데이터 기반의 문제 해결을 위한 빅데이터, AI 등의 여러 워크로드들에 대한 계산 영역으로 그 사용 범위가 확장되고 있다. 특히 GPU 컴퓨팅에 특화되어 있는 AI 문제 영역을 다루는 연구자들이 많아짐에 따라 GPU 장착 슈퍼컴퓨터가 늘어가고 있는 추세이다. 더구나 계산과학 영역에서 필수로 고려하는 FP64의 연산 성능에서도 GPU가 CPU 대비 더 높은 계산 성능을 제공해 주고 있다.[1] 이는 GPU 자원 활용이 계산과학 솔버에게도 CPU 자원 대비 더욱 효율적인 상황으로 되었다. 하지만 앞으로는 HPC 자원이 상대적으로 AI의 연구자들로부터 많은 영역이 점유될 것으로 예상됨에 따라 계산과학 영역에서의 슈퍼컴퓨터 활용 기회가 점점 축소될 수도 있다는 우려가 있다. 또 다른 방향에서는 PINN[2]이나 DeepONet[3] 등 AI를 계산과학 문제와 연계한 하이브리드 계산 영역으로도 일부 사용자층이 확장되고 있기도 하다.

계산과학 영역에서 ANSYS Fluent 등의 상용 소프트웨어 개발업체들은 Nvidia 등 GPU 개발업체와의 전략적인 GPU 솔버 장착 준비를 일찍부터 추진해 왔다.[4] 이와는 달리 공학 분야 오픈 소스 소프트웨어 진영에서는 NVIDIA, AMD, INTEL 등의 GPU에 적용하기 위한 공식적인 서비스로의 제시나 개별적인 개발자의 연구가 추진되고 있다.

본 논문에서는 공학 분야에 대한 오픈 소스 소프트웨어들 중에서 오랜 개발 기간 동안 탄탄한 사용자층을 확보하고 있으며, 길이 스케일 기준에 따른 원자나 분자 영역, 연속체 영역의 전체 영역을 포괄할 수 있도록 추가적으로 CPU 기반에서 활용도가 높은 구조해석 솔버로 Code_Aster(https://code-aster.org/)를 포함하여 총 6개의 오픈 소스 소프트웨어(Code_Aster, GROMACS, LAMMPS, OpenFOAM, Quantum ESPRESSO, SIESTA)들을 공학 분야 주요 소프트웨어[5]로 하여 GPU 솔버에 대한 집중적인 활용성을 검토하였다. 또한 관련 분야 상용 소프트웨어들의 GPU 솔버 적용 상황도 함께 세부적으로 확인해 보고자 하였다. 누리온 슈퍼컴퓨터와 함께 국내 연구자들에게 서비스되고 있는 NVIDIA의 V100, H100, H200, GH200 등의 GPU 기반 뉴론 시스템(https://www.ksc.re.kr/byjw/neuron)을 통

해 앞에서 제시된 오픈 소스 소프트웨어들의 GPU 솔버를 각 소프트웨어별로 간단한 예제를 사용하여 테스트 하였다. 이는 2025년부터 본격적으로 국내 연구자들을 대상으로 서비스하게 될 HPE 업체에서 구축 예정인 600PF급의 GPU 기반 6호기 국가슈퍼컴퓨터[6]를 대비하기 위해 준비하였다.

본 연구를 통해 오픈 소스 소프트웨어에 적용 가능한 다양한 GPU 솔버 모듈에 대해 파악하고 이를 통한 기존 미흡하게 적용된 일부 오픈 소스 소프트웨어에도 적용 가능한 통합적인 GPU 솔버로의 추후 개발 방향을 고려해 보고자 한다.

2. 공학분야 오픈 소스 소프트웨어의 GPU 기반 동향

2.1 슈퍼컴퓨터의 GPU 기반 시스템 상황

2024년 11월 기준 전 세계 슈퍼컴퓨터 순위(Top500 list:https://www.top500.org/lists/top500/2024/11/)에서 살펴보면 엑사스케일 성능을 넘는 1~3위의 슈퍼컴퓨터에는 1위인 El Capitan(LLNL)와 2위인 Frontier(ORNL)에 모두 AMD GPU가 장착되어 있다. 그리고 3위인 Aurora(ANL)에 INTEL GPU가 포함되어 있다. 슈퍼컴퓨터 순위 10위까지로 확장해 보면 AMD GPU가 5개, NVIDIA가 3개, INTEL GPU 1개, 그리고 일본 Fugaku만이 단지 CPU 독립 시스템으로 남아 있다. 즉, 상위 계산 성능의 슈퍼컴퓨터는 대부분이 GPU 기반 시스템으로 새롭게 정착되었음을 알 수 있다. 2022년 11월~2024년 11월까지 총 3년에 걸쳐 GPU 기반 슈퍼컴퓨터의 비중을 확인해 본 결과 각각 36%, 37%, 42%로 매년 증가하고 있음을 알 수 있다. 즉, 기존 CPU 기반 슈퍼컴퓨터를 점진적으로 대체되고 있는 것이다. 최근의 슈퍼컴퓨터 순위(2024년 11월 기준)에서 확인해 보면 NVIDIA의 GPU가 183개로 AMD의 19개와 INTEL GPU의 5개 대비 압도적으로 가장 많이 장착되어 있음을 알 수 있으며, 전체적으로는 이 3개 GPU가 거의 모든 슈퍼컴퓨터에 장착된 GPU에 해당된다.

2.2 공학 분야 6종의 오픈 소스 소프트웨어 파악

공학분야에서 길이 스케일별로 구분하게 되면 원자나 분자의 동역학을 고려하기 위한 오픈 소스 소프트웨어(GROMACS, LAMMPS, Quantum ESPRESSO, SIESTA)들과 연속체 역학 분야 오픈 소스 소프트웨어(Code_Aster, OpenFOAM)가 있다. 표 1에서 보면 원자나 분자 스케일을 다루는 GROMACS, LAMMPS, Quantum ESPRESSO, SIESTA는 대규모 모델링을 통한 해석이 필요하여 많은 자원을 요구하는 빈도가 높을 것이기 때문에 공식적인 GPU 솔버 모듈의 장착이 이뤄진 것으로 판단되며 그만큼 GPU 계산자원에 기대는 사용자들의 수요가 많음을 알 수 있다. 표 1과 같이 공식적인 GPU 솔버 적용 노력도 일부 확인되나 별도로 일부 개발자들의 자체 GPU 솔버 개발 추진 중이다. 제시된 6종 소프트웨어 모두 코드 배포가 20년 이상 개발되고 있는 주요 소프트웨어들로 이중 GROMACS, LAMMPS, Quantum ESPRESSO는 GPU 솔버에 대해서는 공식적인 서비스로 확인되고 있다. SIESTA는 공식적으로 GPU 모듈을 제공하지는 않고 있으나 ELSI-ELPA로의 외부 ELPA 모듈로부터 ELSI 라이브러리를 구성하여 GPU 가속이 가능한 상황이다. 연속체 역학을 다루는 길이 스케일의 경우 상대적으로 대규모 해석 시뮬레이션 상황에서는 병렬 수준을 높이면 계산 포화 상황이 발생될 여지가 큼에 따라 선형적 병렬성이 확보되기 어려운 문제를 포함하고도 있다. 이에 따라 아직까지는 GPU 자원에 대한 큰 수요로 진전되지는 않은 것으로 보여 GPU 솔버 모듈에 대한 공식적인 지원이 적용되지는 않았을 것이라 짐작된다. OpenFOAM의 경우 개별적으로 상당한 GPU 솔버 모듈들을 통한 계산 접근이 이뤄져 왔다. 하지만 유동해석에 비해 상대적으로 대규모 계산이 확연하게 적용되지 않고 있는 구조해석 분야에서의 Code_Aster는 개별적인 GPU 솔버에 대한 개발 노력도 찾아보기 쉽지 않은 상황이다.

[표 1] 오픈 소스 소프트웨어 GPU 지원

오픈 소스 소프트웨어	공식 GPU 솔버 모듈 (GPU 제공년도)	CPU 솔버 배포시작(년도)
Code_Aster	X	2001
GROMACS	O(2013)	1901
LAMMPS	O(2004)	2004
OpenFOAM	X	2004
Quantum ESPRESSO	O(2019)	2001
SIESTA	X	1996

3. 오픈 소스 소프트웨어의 GPU 솔버 모듈 활용성

3.1 GPU 솔버 모듈 테스트를 위한 뉴런 시스템

공학 분야 오픈 소스 소프트웨어들이 GPU 자원을 효과적으로 활용할 수 있는지를 사전 충분히 검토해 보기 위해 기존 구축된 GPU기반 서버로 서비스 중인 뉴론 시스템에서의 해당 일부 GPU 솔버 모듈들에 대해 테스트 하였다. 2.1절에서도 제시되었듯이 대부분의 슈퍼컴퓨터에서는 NVIDIA의 GPU가 상당 부분을 차지하고 있고 현재 뉴론 시스템에는 현재 V100, H100, H200, GH200 등의 GPU 카드가 장착되어 있다. 각 GPU 카드 그룹별로 작업 큐가 지정되어 있어 선택적 GPU 카드 아키텍쳐에 대해 성능 테스트가 가능하다. 이러한 다중 NVIDIA의 GPU 카드 셋팅 큐 중에서 곧 구축될 국가슈퍼컴퓨터 6호기의 GPU일수도 있는 상황이다. 따라서 현 시점에서는 뉴론을 통한 사전 공학분야 오픈 소스 소프트웨어의 GPU 솔버 모듈의 활용성을 검토해 보는 것은 가장 효율적인 방법으로 판단된다.

3.2 GROMACS 예제 테스트

원자수가 약 8만개의 작은 시스템인 benchMEM의 예제로 독일 막스 플랑크 연구소에서 제공하는 벤치마크(https://www.mpinat.mpg.de/grubmueller/bench)를 적용했다.. Neuron의 V100 2개 노드(cas_v100_2)와 V100 4개 NVLINK노드(cas_v100nv_4)를 사용하여 테스트하였다. cas_v100_2의 경우 86.277 ns/day의 성능이 확인되었으며, cas_v100nv_4에 대해서는 76.269 ns/day로 성능 결과가 나타났으며, 이는 적용 시스템의 크기가 작아서 발생된 원인으로 판단된다.

다음으로는 bechPEP의 예제로 원자수가 1,200만개 수준인 상태에서 다시 동일 큐에서 테스트 하였다. cas_v100_2의 경우 1.633ns/day와 cas_v100nv_4에서는 1.917ns/day의 성능으로 일정 수준 v100 GPU 카드를 2배 추가하였으나 해당 성능이 2배 수준으로는 확보되지는 않았다.

3.3 LAMMPS 예제 테스트

총 10,000개의 물 분자로 이뤄진 시스템에 대해 CPU+GPU의 하이브리드 환경에서 테스트 하였다. 뉴론 서버의 V100 NVlink GPU로 4개 GPU를 Intel Xeon Gold 6230 40 코어 CPU와 하이브리드 해석을 적용하였다. 또한 8개 GPU를 Intel Xeon Gold 6226R 32 코어 CPU와 하이브리드 해석에 사용하였다. 기존 2개의 노드로 128개의 CPU를 사용(1,612초)했을 때 대비 4개의 GPU를 사용하였을 때는 약 두 배 정도 성능(905초)이 높아졌으며, 8개 GPU를 사용한 경우는 4개 GPU 대비 2배 이상의 성능(391초)이 추가적으로 확보됨을 확인하였다.

3.4 Quantum ESPRESSO 예제 테스트

본 예제는 실리콘의 슈퍼셀에 대한 계산을 수행한 것으로 기본 셀(primitive cell)은 내부에 2개의 실리콘 원자를 가지고 있다. 이에 계산 수준을 높이기 위해 x, y, z축으로 각 6개씩 추가로 연결시켜 $6 \times 6 \times 6$으로 슈퍼셀을 제작하였다. 단위 셀(unit cell) 내부에 총 432개($2 \times 6 \times 6 \times 6$)의 원자를 가지는 큰 시스템으로 고려하였다. Amd_a100nv_8 큐를 사용하여 테스트를 수행한 결과 SKL의 4개 노드 CPU 기반 해석(233초) 대비 1개 노드에서의 2개 GPU를 사용한 해석에서 279초 계산 결과 및 2개 노드에서의 2개 GPU로는 227초로의 성능이 확보됨을 확인되었다. 추가로 4개 노드의 2개 GPU의 경우는 144초로 성능을 높일 수 있었다.

4. 결론

본 연구를 통해 총 6가지 오픈 소스 소프트웨어에 대한 GPU 솔버의 활용성을 검토하였다. 오픈 소스 소프트웨어 진영에서도 원자나 분자의 동역학 해석 분야로의 GROMACS, LAMMPS, Quantum ESSPRESSO 등의 적극적인 공식 GPU 솔버가 포함된 소프트웨어가 있으며 반면 Code_Aster나 OpenFOAM 등 연속체 역학 분야에서의 기존 CPU 기반 모듈만이 아직까지 공식적으로 제공되고 있고 GPU 솔버의 공식 지원 준비가 미흡한 공학 영역도 확인하였다.

그리고 현 단계에서 적용 가능한 공식 또는 비공식 GPU 솔버 모듈이 포함된 오픈 소스 소프트웨어에 대해서는 GPU 기반인 뉴런 클러스터로 분석하였다. 이를 통해 Quantum ESSPRESSO, GROMACS, LAMMPS 등 공식 GPU 솔버로 제공되는 경우 GPU 솔버의 계산효율이 CPU 기반 솔버 대비 효과적임을 확인할 수 있었다.

이를 통해 추가로 슈퍼컴퓨터가 점진적 GPU 기반 HPC 시스템으로 바뀌는 추세임을 고려할 때 미흡한 상황에 있는 연속체역학 분야의 오픈 소스 소프트웨어 등에 대해 적극적인 공식 GPU 솔버로 확보될 수 있도록 지원하는 한편 각 오픈 소스 소프트웨어별로 구현된 GPU 솔버들의 분석 자료로부터 자체적인 통합 GPU 개선 솔버로의 확장도 추진해 보고자 한다.

후기

이 논문은 2025년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다.(과제번호: K25L2M2C3)

참고문헌

[1] H Jack Dongarra, John Gunnels, Harun Bayraktar, Azzam Haidar, Dan Ernst, ardware Trends Impacting Floating-Point Computations In Scientific Applications, November 2024, https://doi.org/10.48550/arXiv.2411.12090

[2] RaissiM.,PerdikarisP.,KarniadakisG.E., “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” J. Comput. Phys.,378(2019), pp.686-707

[3] Lu, J., Shen, Z., Yang, H., and Zhang, S. (2021a).Deep network approximation for smooth functions.SIAM Journal on Mathematical Analysis, 53(5):5465-5506.

[4] Robert Strzodka, Accelerated ANSYS Fluent: Algebraic Multigrid on a GPU, Presendation: https://developer.download.nvidia.com/GTC/PDF/GTC2012/PresentationPDF/RobertStrzodka_Accelerated_ANSYS_Fluent_SC12.pdf

[5] 윤태호, “고성능컴퓨팅 기술분류체계 구성 및 활성화 연구”,
한국기술혁신학회 2023년 추계학술대회, 2023
[6] KISTI 슈퍼컴퓨터 6호기 시스템 구축
제안요청서(RFP), ver.1.31, 2024