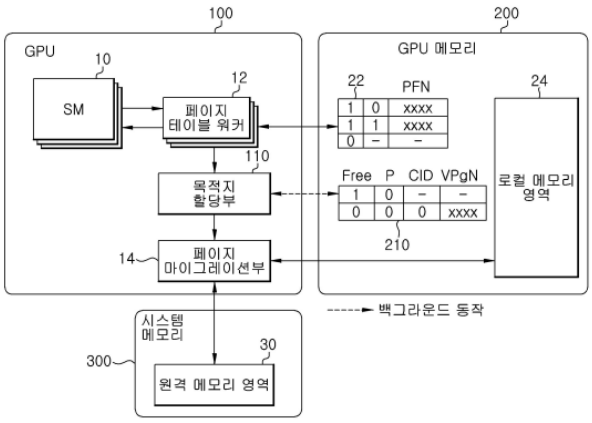
로컬 메모리 액세스 감지를 통한 페이지 마이그레이션 시스템

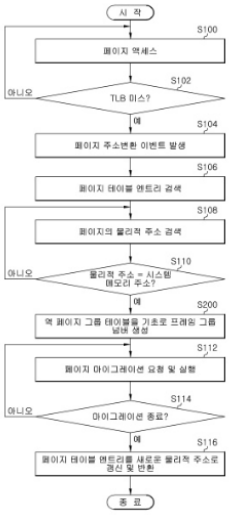
연구개발자: 반도체시스템공학과 김정래 교수

I 기술 개요

01 기술 요약



[GPU의 가상 메모리 관리 장치에 대한 블록도]



[GPU 가상 메모리 관리 방법의 흐름도]

- 본 기술은 로컬 메모리 접근을 하드웨어에서 감지해 소프트웨어 개입 없이 페이지를 자동 마이그레이션하는 GPU 통합 메모리 관리 기술에 관한 것으로, 소용량 GPU 메모리 환경에서도 성능 저하 없이 대규모 어플리케이션 실행이 가능하도록 하는 것을 특징으로 함

02 지식재산권 현황

No	발명의 명칭	출원번호	출원년도
1	손실 데이터 탐지하는 기법	2025-0019216	2025
2	메모리 소자와 그 동작 방법	2024-0080973	2024
3	메모리 소자와 반도체 장치 및 그 동작 방법	2024-0051648	2024
4	디램 소자와 반도체 장치 및 그 동작 방법	2024-0051623	2024
5	그래픽 처리 장치의 메모리 보호 장치 및 방법	2024-0136063	2024
6	디램 및 그 제어방법	2023-0150978	2023
7	단일 심볼 오류 정정 및 이중 비트 오류 정정을 위한 부호 생성 방법, 이를 위한 오류 정정 부호 생성 장치	2023-0078765	2023
8	이기종 시스템의 가상 메모리 관리 장치 및 방법	2022-0168002	2022
9	메모리 장치 및 메모리 리매핑 방법	2021-0096297	2021

로컬 메모리 액세스 감지를 통한 페이지 마이그레이션 시스템

03 기술의 우수성

- **하드웨어 기반 자동 마이그레이션**

-H/W가 로컬 메모리 액세스를 직접 감지하여 S/W의 개입 없이 페이지를 마이그레이션, 기존 S/W 기반 페이지 폴트 처리 시간을 제거

- **성능 향상 및 오버헤드 감소**

-페이지 폴트 처리 및 S/W 오버헤드를 H/W로 오프로딩하여 마이그레이션 소요 시간을 줄이고 전체 실행 성능을 향상

- **메모리 초과 사용 효율 극대화**

-메모리 초과 사용 시에도 성능 악화가 급격하게 증가하지 않아, 한정된 GPU 메모리로 더 큰 AI 어플리케이션 실행 가능

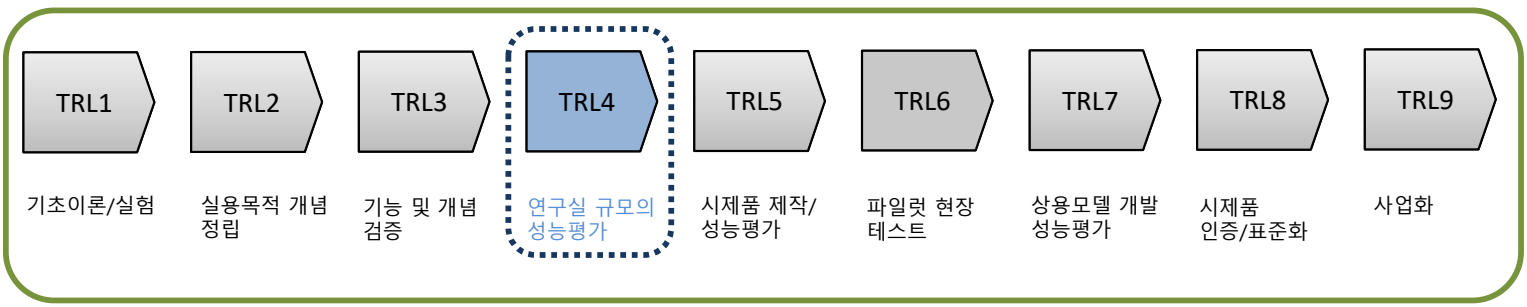
- **통합 메모리 장점 최대화**

-통합 메모리 환경에서 데이터 접근성을 향상시키고, 마이그레이션 비용을 최소화하여 통합 메모리의 장점인 메모리 초과 사용 비율을 높임

- **대규모 AI 학습에 최적화**

-대규모 데이터 셋을 다루는 인공지능 학습 분야에서 메모리 제약 없이 효과적인 활용이 가능하도록 설계되어 HPC/AI 시장의 핵심 요구 사항 충족

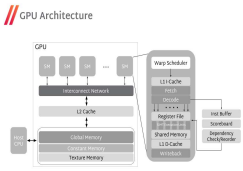
04 기술 개발 완성도

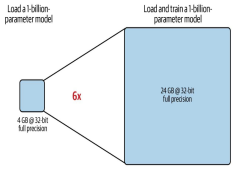


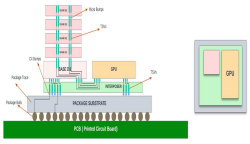
로컬 메모리 액세스 감지를 통한 페이지 마이그레이션 시스템

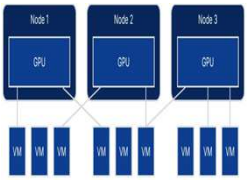
II 기술 동향

01 기술응용분야









[고성능 컴퓨팅 시스템]

슈퍼컴퓨터 또는 데이터센터의 대규모 병렬 GPU 환경에서 메모리 자원 배분 효율을 높여 연산 속도를 극대화

[AI 학습 및 추론]

LLM 등 메모리 집약적인 AI 모델의 학습 시, 제한된 GPU 메모리로 더 큰 모델을 안정적으로 운영

[그래픽 처리 장치 제조]

GPU 아키텍처 및 칩셋에 핵심 메모리 관리 기능으로 통합하여 제품 경쟁력 강화

[클라우드 컴퓨팅 서비스]

클라우드 GPU 인스턴스의 메모리 가상화 및 통합 메모리 서비스의 성능을 근본적으로 개선

[자율주행 및 임베디드 AI]

메모리 자원이 제한적인 자율주행 차량 또는 엣지 디바이스의 AI 프로세서에서 실시간 데이터 처리 효율을 개선

02 기술 동향

[~2020]

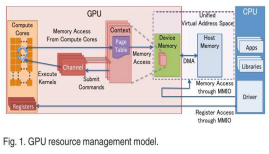
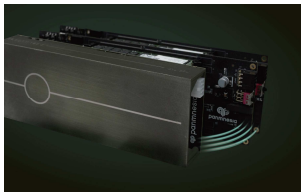


Fig. 1. GPU resource management model.

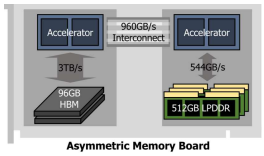
기존 GPU 메모리 관리, 페이지 폴트 발생 시 OS/Driver 개입을 통한 S/W 기반 마이그레이션으로 오버헤드 발생

[2021~2024]



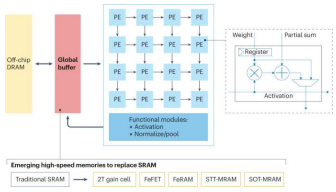
CXL 등 메모리 확장 기술과 결합하여 GPU 통합 메모리 기술이 발전

[2025]



하드웨어 기반 지능형 메모리 관리가 HPC/AI 가속기의 핵심 경쟁 요소로 자리 잡음

[향후 전망]



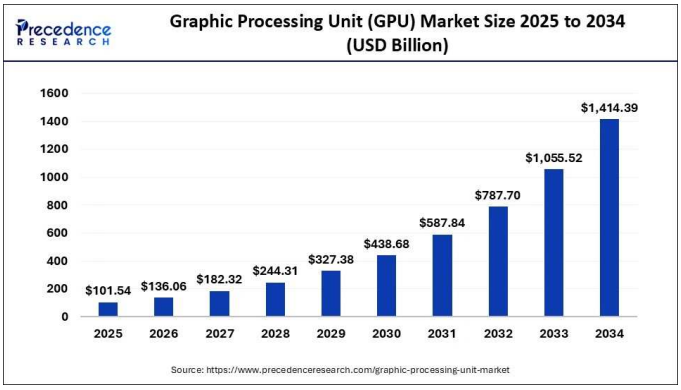
비대면, AI 기반 서비스 확산으로 대규모 데이터 처리 수요 급증

AI·HPC 환경에서 데이터셋 대형화로 메모리 초과 문제가 심화되고, CPU–GPU 통합 시스템의 빈번한 페이지 마이그레이션에 따른 성능 저하가 주요 병목으로 부각되고 있음. 본 기술은 소프트웨어 개입을 최소화한 하드웨어 가속 기반 페이지 폴트·마이그레이션 처리로 실행 시간 증가를 억제하여 대규모 AI 학습 및 HPC 분야에서 높은 시장 가치를 가질 것으로 전망됨

로컬 메모리 액세스 감지를 통한 페이지 마이그레이션 시스템

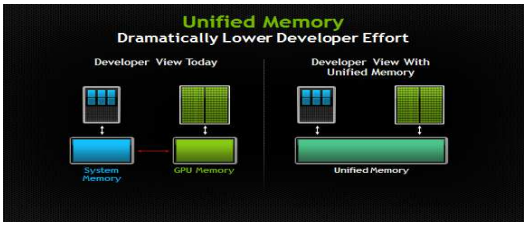
III 시장 동향

01 시장규모



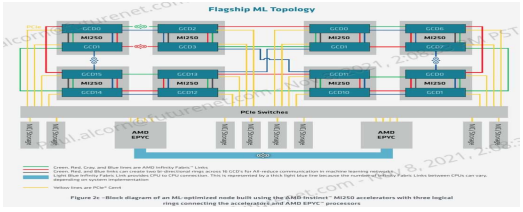
- 글로벌 GPU 시장은 2025년에 1,015억 4천만 달러로 추산되었으며, 2026년 1,360억 6천만 달러에서 2034년에는 약 1조 4,143억 9천만 달러로 증가하여 2025년에서 2034년까지 연평균 성장률 13.8%로 확대될 것으로 예측됨

02 주요 시장 참여자



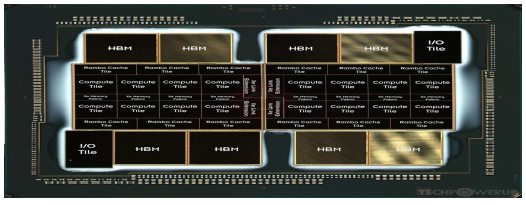
[NVIDIA 社 CUDA Unified Memory 제품]

- 글로벌 AI GPU 시장 선두 주자, 통합 메모리 기술 개선 및 H/W 가속 기반의 메모리 관리 기술 확보



[AMD 社 Infinity Architecture 제품]

- HPC 시장 점유율 확대를 위해 CPU-GPU 간 통합 메모리 성능 개선 및 효율 극대화 기술에 집중 투자



[Intel 社 One API / Ponte Vecchio GPU 제품]

- 이기종 컴퓨팅 환경에서의 메모리 관리 및 데이터 이동 최적화 기술 확보를 통해 HPC 시장 진입을 목표로 함

기술 이전 상담 및 문의

산학사업팀 담당자 김민경 / 031-290-5087 / mikkypat@skku.edu