업스테이지 컨소시엄과 래블업이 한국형 파운데이션 모델을 개발하기 위해 Backend.AI를 활용하는 방법

회복탄력적인 대규모 학습 환경 구축을 위한
래블업과 업스테이지의 엔드투엔드 접근 방식

lablup upstage

Case Study upstage

업스테이지 컨소시엄과 래블업이 한국형 파운데이션 모델을 개발하기 위해 Backend.AI를 활용하는 방법

"업스테이지는 인프라 단의 모든 것을 Backend.AI에게 맡겼습니다. 덕분에 인프라 관리에 시간을 할애하는 대신 거대언어모델(LLM) 개발에 매진할 수 있었습니다."

업스테이지
이승윤 기술이사

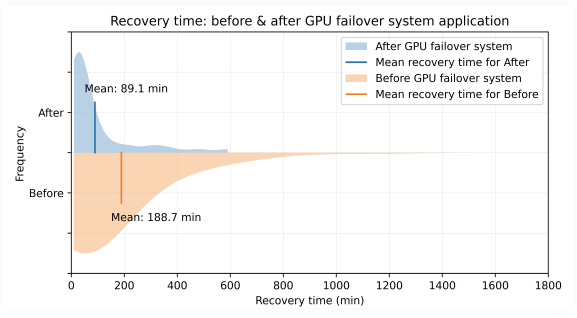
'독자 AI 파운데이션 모델 프로젝트'는 대한민국 정부가 한국적 정체성과 언어에 최적화된 경쟁력 있는 AI 모델을 개발하기 위해 추진하고 있는 국가 주도의 사업으로, 한국의 AI 경쟁력을 강화하고 데이터 주권을 확보하며 AI 산업 생태계의 지속 가능한 성장을 촉진하기 위한 전략적 노력입니다. 래블업은 거대언어모델 개발 역량을 인정받고 있는 업스테이지와 협력해 스타트업 중심의 컨소시엄을 구성했고, 치열한 경쟁을 거쳐 최종 5개 정예팀에 선정되어 프로젝트를 수행하게 되었습니다. 선정된 컨소시엄은 글로벌 수준의 성능을 달성할 수 있도록 GPU 자원, 데이터, 인재 등 핵심 자원을 지원받습니다.

GPU 클러스터 성능을 안정적으로 유지하기 위한 도전

독자 AI 파운데이션 모델 프로젝트는 제한된 기간 안에 거대언어모델(LLM)을 최대 효율로 개발해야 하는 과제이기에 비효율이나 시스템 중단이 발생하지 않아야 했습니다. 이러한 요구를 충족하려면 여러 노드에 걸친 고강도 워크로드 환경에서도 인프라가 끊김 없이 안정적인 성능을 유지해야 했습니다. 업스테이지는 Backend.AI의 컨테이너 수준 GPU 가상화 기술과 함께 AI 워크로드를 위해 설계된 Sokovan 스케줄러를 활용하여 AI 모델 개발 효율성을 확보하였으나, 대규모 클러스터 환경에서 발생하는 불가피하고 예측하기 어려운 GPU 장애 (Cui et al., arXiv:2503.11901)는 또 다른 문제를 유발하였습니다. 이를 해결하기 위해 래블업과 업스테이지는 사람이 개입하지 않아도 GPU 노드 장애를 감별하고 복원할 수 있는 엔드투엔드 복원 루프를 설계했습니다. 전통적인 NVIDIA DCGM 및 래블업이 자체 개발한 GPU 및 서버 모니터링 소프트웨어를 통해 실시간으로 장애를 감지하고, MS Teams와 Slack을 통해 사용자에게 자동으로 알림을 보냅니다. 감지된 장애가 워크로드를 중단시키는 결함이라면 장애 노드를 자동으로 격리하고 예비 풀에서 즉시 대체 노드를 투입한 뒤, 멀티노드 세션을 재시작하고 최신 체크포인트부터 학습을 이어가도록 구성했습니다. 이러한 설계는 사람이 개입해야 하는 과정을 최소화하고, 장기간 실행되는 학습 작업의 처리량을 유지하며, 연구자의 생산성을 보호하는 데 핵심적인 역할을 수행합니다.

멈추지 않는 모델 학습을 위하여: 대규모 훈련 안정성을 확보하는 Backend.AI의 자동 장애 복구

업스테이지는 GPU 500여장 이상으로 구성된 대규모 클러스터를 안정적으로 운영하기 위해 Backend.AI를 활용했습니다. Backend.AI는 분산된 수백여 대의 GPU 자원을 체계적으로 오케스트레이션하면서, 훈련 환경 전반의 투명한 가시성과 관리 권한을 유지할 수 있도록 지원합니다. 이를 통해 업스테이지는 학습 중단 후 재개까지 소요되는 시간을 **47%** 가까이 줄여 인프라 유지·보수 노력을 최소화할 수 있었으며, 모델 개발을 위해 주어진 시간을 온전하게 모델 학습에 활용할 수 있었습니다.



Backend.AI 자동 장애 복구 기능 적용 전후의 Failover 시간 비교