

IBKS Spot Comment

IT/반도체/디스플레이

김운호

02) 6915-5656

unokim88@ibks.com

[DeepSeek AI 모델 발표]

GPU 및 HBM 수요 단기적으로는 중립적 이벤트

What's New: DeepSeek R1 발표

2023년 11월 DeepSeek Coder.

2024년 5월 DeepSeek V2 발표.

2024년 12월에 DeepSeek-v3. 발표 이후 2025년 1월 DeepSeek R1 발표.

R1은 미국 수학경시대회 벤치마크 테스트에서 Open AI의 o1을 능가.

V3 개발에 57.6만 달러(약 80억원) 투자. 인건비 운영비 제외. \$2/hour*2 months. 실제 비용은 \$4~8로 보는게 정상(Alibaba, Tencent Cloud 가격).

V3는 H800 2048개. 2,788,000 GPU Hour. 671B 모델 학습.

Open Source를 지양하기 때문에 무료 서비스. 이로 인해 Chat GPT 보다 많은 구독자 확보.

Analysis: DeepSeek 사용 GPU는 H800, H100의 Spec down 제품

DeepSeek에서 사용한 GPU는 H800 2,048개로 학습했다고 밝히면서 H100 이상 High End GPU에 대한 무용론이 제기되고 있음.

H100과 H800을 비교하면 HBM 용량은 동일하고, 초기 H800은 HBM2e를 사용했으나 현재는 HBM3가 적용 중. 대역폭과 전력 소모량에서 차이가 있음.

H800 2개와 H100 1개가 동일한 성능.

표1. H100과 H800 주요 사양 비교

구분	H100	H800
아키텍처	Hopper	Hopper
CUDA 코어수	16,896	16,896
메모리 용량	80GB HBM3	80GB HBM3
메모리 대역폭	3.35TB/s	2.9 TB/s
NVLink 대역폭	900GB/s	400GB/s
전력 소모	700 W	350 W

자료: NVIDIA, IBK투자증권

So What?: GPU, HBM 수요는 지속적으로 확대될 것

DeepSeek는 NVIDIA의 GPU H800으로도 충분한 성능을 확보할 수 있다는 것을 증명했다고 하나 투자 금액 규모에 대해서는 논란의 여지가 있음.

일부 언론에서는 H800 가격을 \$10,000 내외로 소개하고 있으나 인터넷 쇼핑몰에서는 \$30,000에서 거래되기도 함. 이는 물론 H200, B200에 비해서는 저렴. 1시간 대여료도 투자 금액 산정할 때 \$2/시간으로 되어 있으나 실제 알리바바, 텐센트 클라우드 비용은 \$4~7. H100대비 10~30% 저렴. 인건비와 운영비는 제외한 금액.

추후 DeepSeek와 같은 AI 개발업체가 지속적으로 출현할 것으로 예상하지만 선도업체인 Open AI를 모든 면에서 능가하기는 쉽지 않을 것으로 판단. 가성비면에서는 어느 정도 증명한 것으로 판단.

현재 투자를 확대하고 있는 AI 업체에게 새로운 시각을 제공할 수도 있을 것으로 판단. 이 경우 이전의 투자 계획을 하향 조정할 가능성도 있음. H800 같은 저사양 GPU를 확보하려는 수요보다는 High End GPU B200 확보 규모에 대한 재고를 의미함.

단기적으로는 GPU 물량 및 HBM에 대한 전망치를 변경할 이벤트는 아니라고 판단함.

High End GPU 수요 증가 및 신제품 채택에 대한 의구심은 꾸준히 제기되고 있지만 후발 주자들의 AI 모델은 Open AI, Google, Meta 등이 연구 개발한 LLM을 바탕으로 보다 쉽게 개발할 수 있었던 것으로 판단.

결국 서버 중심으로 LLM 모델을 구축하고 있는 대형 CSP업체들과 이를 추종하는 업체들 간의 접근 방법에서 차이가 발생하고 있음. 후발업체들의 투자 비용이 선발업체들에 비해서 훨씬 낮은 수준.

AI 모델 개발업체가 증가하고, 연산 능력 향상에 대한 수요가 지속되는 한 NVIDIA의 GPU와 이를 지원하는 HBM에 대한 수요는 꾸준히 증가할 것으로 기대.