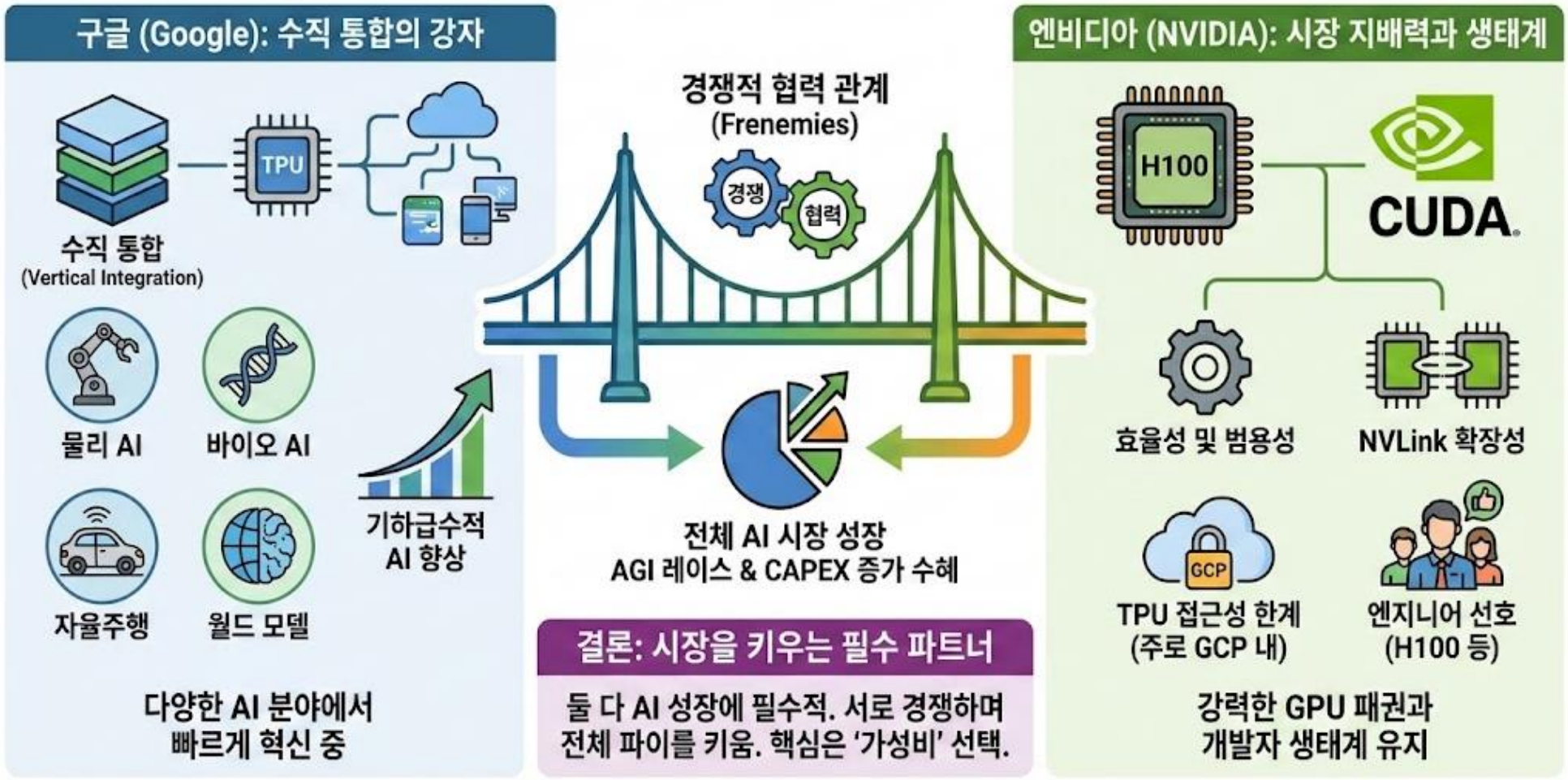
2025년 12월

Endless AI

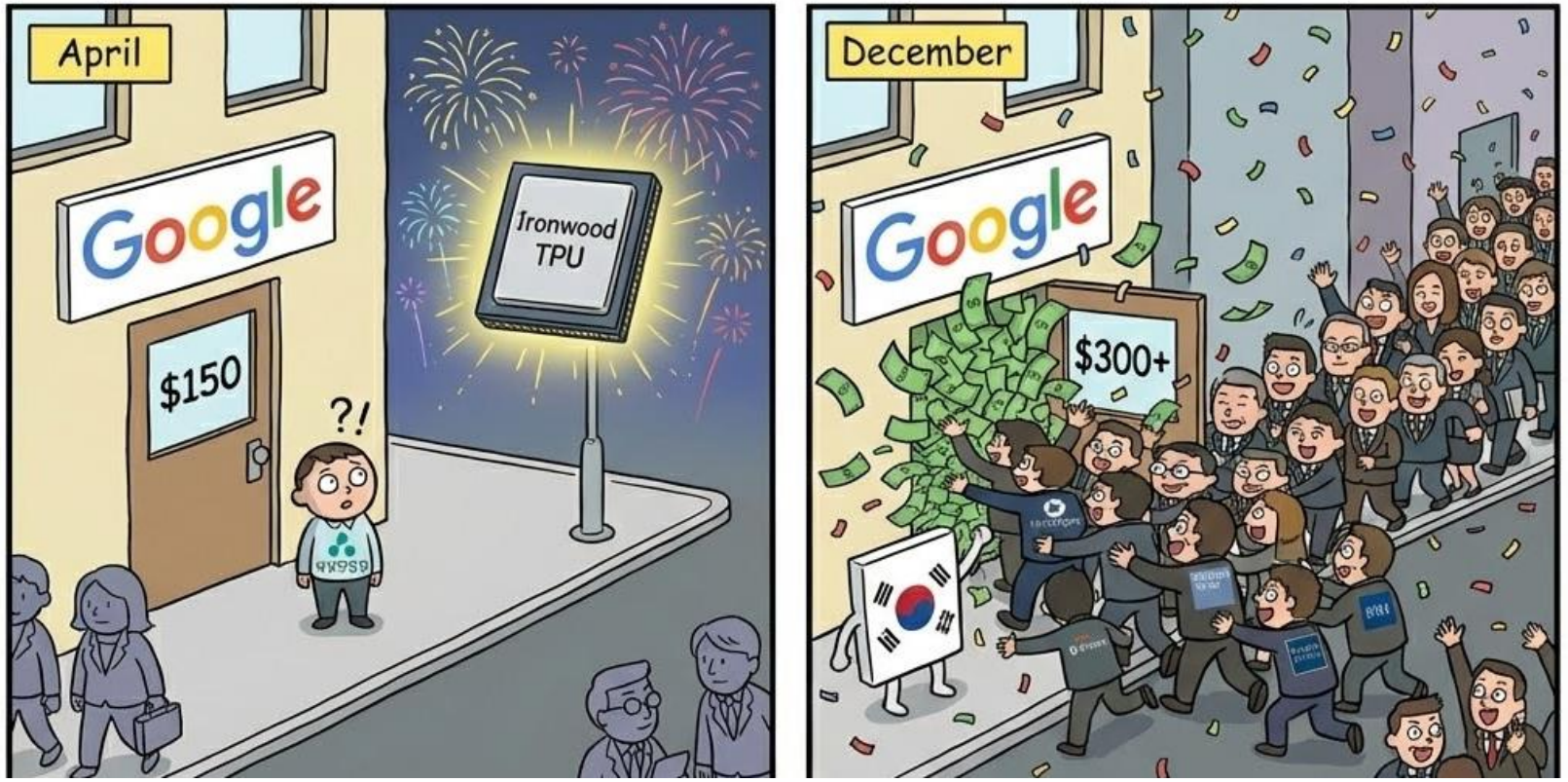
1화 Google & Nvidia : 애증의 동행

상상인증권 리서치센터
김경태 I AI/디지털자산 02-3779-3427, kt.kim@sangsanginib.com

Executive Summary



구글에 열광하는 시장, 그러나 연초부터 TPU의 위력은 이미 알려진 사실



과거 자료 복기: 최근 Google로의 쏠림이 심해진 주식시장

- Google: TPU의 높은 전성비와 수직계열화 및 Nvidia에 대항하는 하드웨어 시장 경쟁자로 급부상한 현 주도주
- 이미 4월에 Ironwood TPU의 성능이 공개된 이후 AI 관련 전문가들은 Google의 하드웨어 자체 공급처 확보 및 그에 따른 낮은 공급원가, 우수한 연구인력 보유에 따른 성능 향상에 주목하여 AI 패권 경쟁에서도 긍정적 전망을 내놓았음
- GPT5가 SOTA 자리를 유지하며 그러한 사실은 다른 이슈에 밀려 잠시 잊혀졌으나, Gemini 3.0 Pro-Preview 모델 및 Nano-Banana 2.0 출시 이후 시장은 Google 롱 – Nvidia 숏 분위기로 흘러감

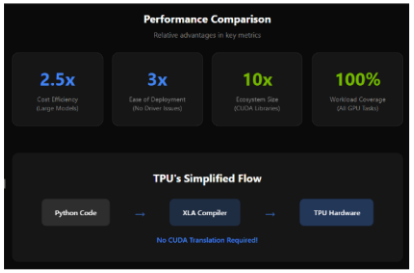
그림 1. 과거 AI In-Depth 자료 Google 전망 1

GPU vs ASIC

구글 : 10년 전부터 AI만을 준비해온 TPU의 독주 전망

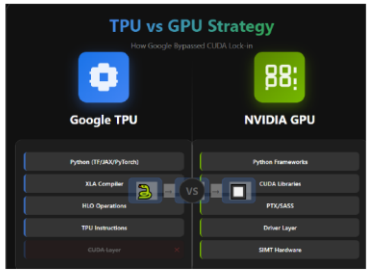
- Google은 자체개발 TensorFlow/JAX/XLA로 2015년부터 CUDA 생태계에서 독립
- TPU는 학습 & 추론에서 NVIDIA GPU 대비 훨씬 높은 전성비 실현. 이를 기반으로 중국, 미국 경쟁사보다 훨씬 낮은 단가로 API 서비스 & AI 서비스, 영상 생성까지 제공
- 체크포인트 분산 훈련으로 동시다발적 학습 가능 → 가파른 AI 성능 향상 속도 확보

그림 52. 동급 NVIDIA GPU 대비 TPU 성능 비교



자료: 상상인증권 리서치센터

그림 53. TPU vs GPU LLM 풀스택 비교



자료: 상상인증권 리서치센터

자료: 상상인증권 리서치센터

그림 2. 과거 AI In-Depth 자료 Google 전망 2

해외 기업분석

Google – TPU의 무시무시한 전성비 & 압도적인 연산력 체급

- 높은 GCP 성장률 & 가파른 Gemini 성능 향상 & Android 등 AI 전 영역 시너지 극대화
- TPU v5의 높은 전성비 / Tensorflow&JAX 지원으로 NVIDIA 의존도 빅테크 중 최저
- Waymo & Gemini 기반 RFM으로 Physical AI 시장 선점
- 우려 대비 안정적인 검색 광고 매출 + YouTube의 현금창출 능력은 여전
- 반독점 소송 우려로 P/E 20배 수준 저평가 → AI 실적 반영 시 밸류에이션 재평가 불가피

표 12. Google 주요 재무 지표

결산기	2022	2023	2024	2025E	2026E
매출액($B)	282.84	307.39	350.02	337.09	377.40
YoY(%)	9.8	8.7	13.9	-3.7	12.0
영업이익($B)	74.8	84.3	112.4	129.6	145.6
YoY(%)	-4.9	12.6	33.3	15.3	12.4
순이익($B)	60.0	73.8	100.1	118.4	124.9
YoY(%)	-21.1	23.0	35.7	18.3	5.4
EPS	4.56	5.80	8.04	9.80	10.48
ROE	23.6	27.4	32.9	29.8	26.5
PER	18.1	22.8	23.9	18.1	16.4
PBR	4.4	6.1	7.1	5.7	4.6
배당수익률	-	0.42	0.44	-	-

자료: 상상인증권 리서치센터

그림 75. 주요 AI 모델 별 API 호출 단가 순위

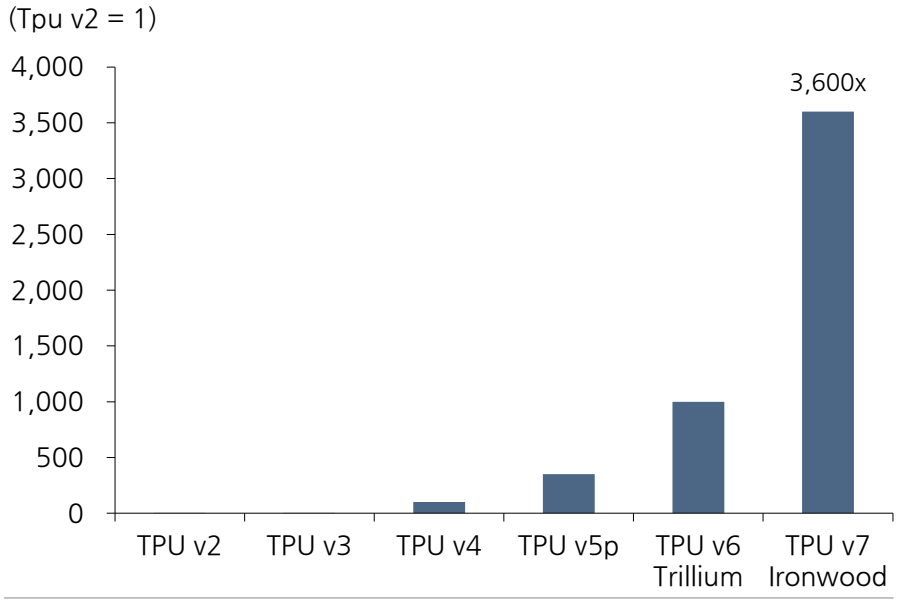
자료: ArtificialAnalysis, 상상인증권 리서치센터

자료: 상상인증권 리서치센터

Google TPU의 경이로운 성장은 AI 세월 축적의 성과

- Google은 단순 하드웨어 경쟁력 강화에 그치지 않고 CUDA 소프트웨어 생태계 침투를 위해 오픈소스 기반 CUDA 커널 지원을 본격화하여, 자체 엔지니어 인력으로 CUDA 기반 AI 모델을 TPU용으로 변환하는 포팅 서비스를 적극 지원
- 모델의 서비스 커버리지 및 API 호출량이 전세계 클라우드와 연결되며 데이터센터 수요와 연산량이 폭증한 현 시점이 GPU에서 TPU로의 포팅 비용이 정당화되는 변곡점으로 작용
- 기존 GPU/CUDA 기반 코드를 TPU/JAX로 전환하고 최적화하는데 소요되는 비용은 연간 수십억에서 수백억까지 발생 가능하며, 유지보수를 감안한다면 증가할 가능성도 상존하나, 장기 전력비용 절감 대비 감내 가능한 수준

그림 3. TPU 세대별 성능 증가 추이: 최대 연산량



자료: Google, 상상인증권 리서치센터

그림 4. Ironwood vs Blackwell 스펙 비교

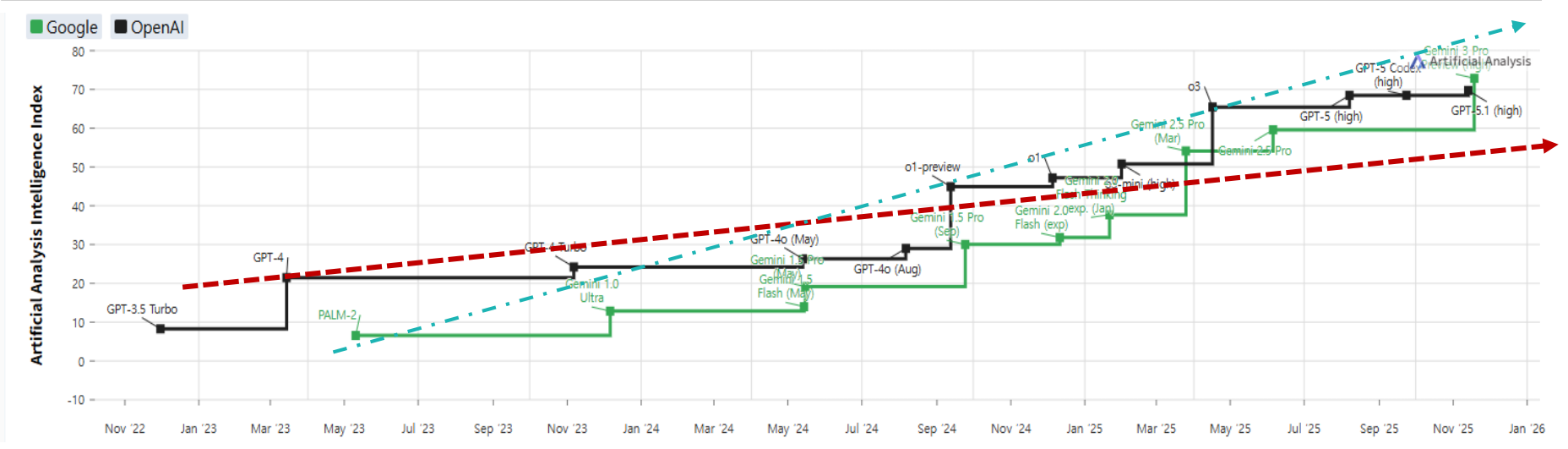
항목	Google TPU v7 (Ironwood)	NVIDIA B200 (Blackwell)
HBM 램 용량	192 GB (칩당)	192 GB (GPU당)
HBM 버전	HBM3e (8단)	HBM3e
메모리 대역폭	7.4 TB/s	8.0 TB/s
FP8 성능 (TFLOPS)	**약 4,614 TFLOPS**	**약 4,500 TFLOPS**
소비 전력 (TDP)	**약 600W (추정)***	**1,000W ~ 1,200W (설정 가능)**
네트워크 속도	ICI: 9.6 Tb/s (칩당)	NVLink: 14.4 Tb/s (1.8 TB/s)
단일 클러스터 규모	9,216 칩 (단일 Pod, OCS 연결)	72 칩 (NVL72, NVLink 도메인)
TCO (총 소유 비용)	GB200 대비 약 30~44% 저렴	H100 대비 가성비 12배 향상
주요 특징	전력 효율 극대화 액체 냉각 필수 아님(선택) 최대 9천 개 연결	최고의 단일 칩 대역폭 NVLink를 통한 72 GPU 단일 랙

자료: Google, Nvidia, 상상인증권 리서치센터

Gemini 성능 자체보다 상승 기울기가 위협적인 상황

- Gemini의 현 성능 수준보다 성능 개선 기울기 자체가 시장에 위협으로 인식되는 상황이며 시장의 Google 집중과 OpenAI의 코드 레드 선언은 이러한 Gemini의 가파른 성능 향상 궤적에 기인
- Google은 OpenAI와 달리 LLM 단일 영역에 집중하지 않고 World Model/Gemini Robotics, AlphaFold, Waymo, Google Glass 등 Physical AI, 제약 바이오, 자율주행, 컴퓨터 비전 및 Wearable AI 전 영역에서 동시 진격 중
- 일각에서는 Pre-training 단계에서 원 모델 성능 자체가 경쟁사 대비 우위에 있다고 평가하나 실제로는 TPU 전성비를 극대화한 추론 최적화 기법으로 체감 성능을 극한까지 끌어올리는 방식에 기인한다는 분석이 유력
- 이는 하드웨어/소프트웨어 수직 통합을 통해 추론 효율을 극대화하는 고유의 전략이 우위를 보여주는 포인트

그림 5. GPT vs Gemini 성능 격차 및 성능 상승 속도

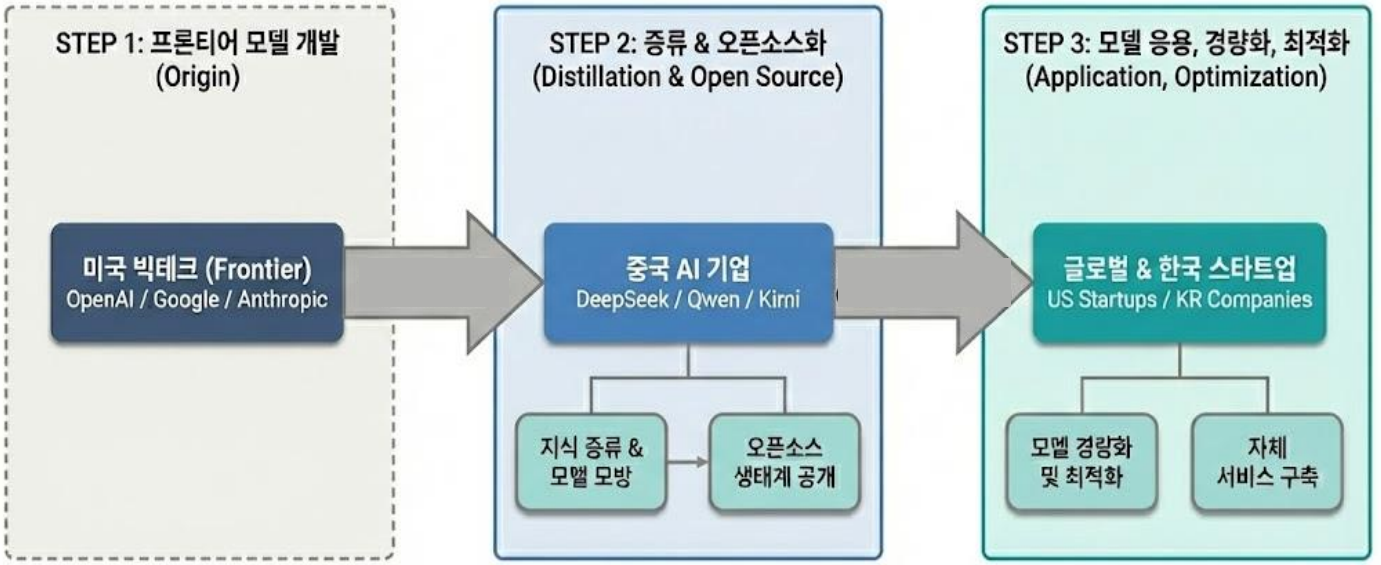


자료: AA, 상상인증권 리서치센터

Google의 장기전 경쟁력은 명확/중국 때문에 독점은 어렵다

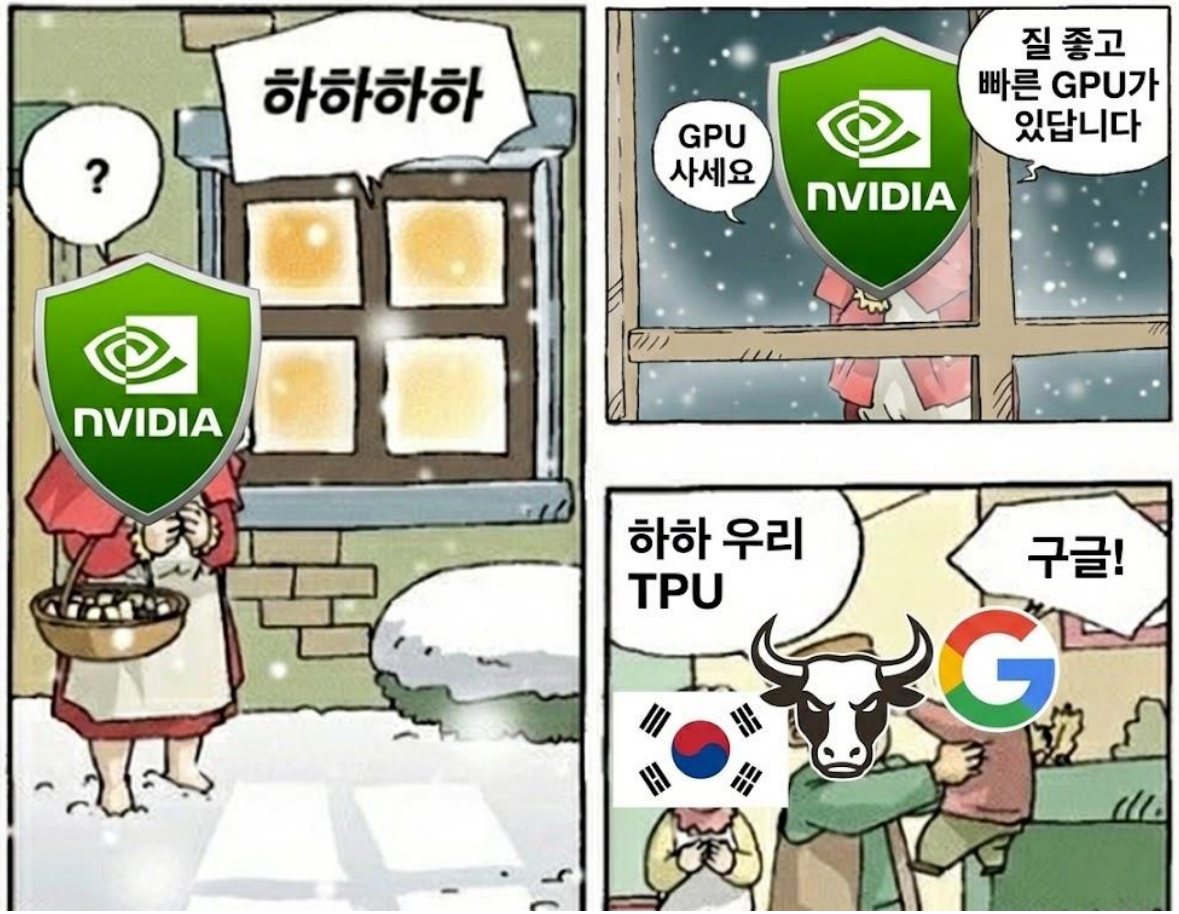
- LLM은 외견상 유사해 보이나, 고성능 모델일수록 모델 간 사용자 선호도, 고유의 강점과 특화 영역이 분화
- 관련 핵심 기술은 개발자 간 교류와 중국발 오픈소스를 통해 핵심 부분은 공개된 상태이며 향후 경쟁력은 데이터 확보 역량, 인력 전문성 및 노하우, R&D 방향성, 설비투자 규모가 좌우
- TPU는 외부 판매를 개시했으며, 장기적으로 비용상 유리. 그러나 Google Cloud 락인 가능성과 시장 점유율 확산으로 경쟁 빅테크 입장에서는 TPU 종속을 수용하기 어려운 상호 견제 역학으로 인해 TPU 확산에는 태생적 한계가 존재
- 이러한 신규 플레이어 참여 & 소규모 모델 범람은 범용 GPU 수요를 뒷받침하는 구조적 요인

그림 6. OpenAI, Gemini, Claude -> 오픈소스 -> 소규모 특화 모델 범람



자료: 상상인증권

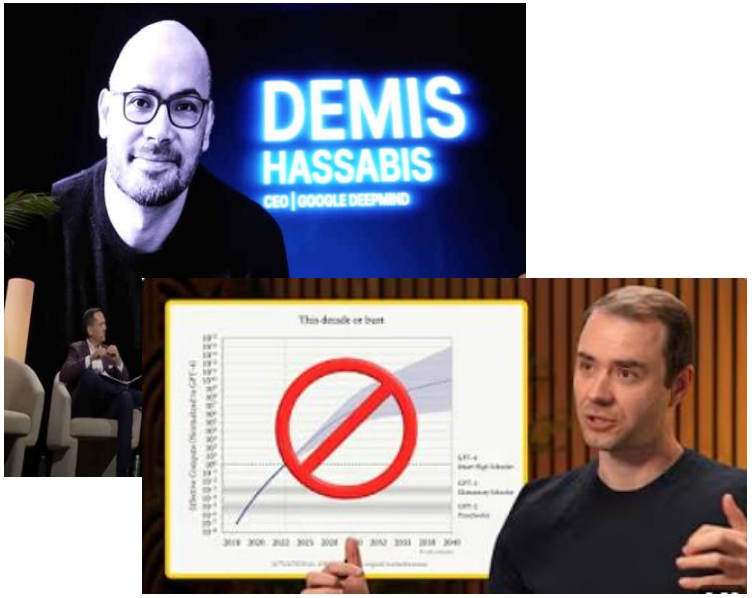
그럼에도 GPU를 다시 봐야하는 이유



AGI까지는 “Breakthough” 필요–GPU 범용성의 가치 높음

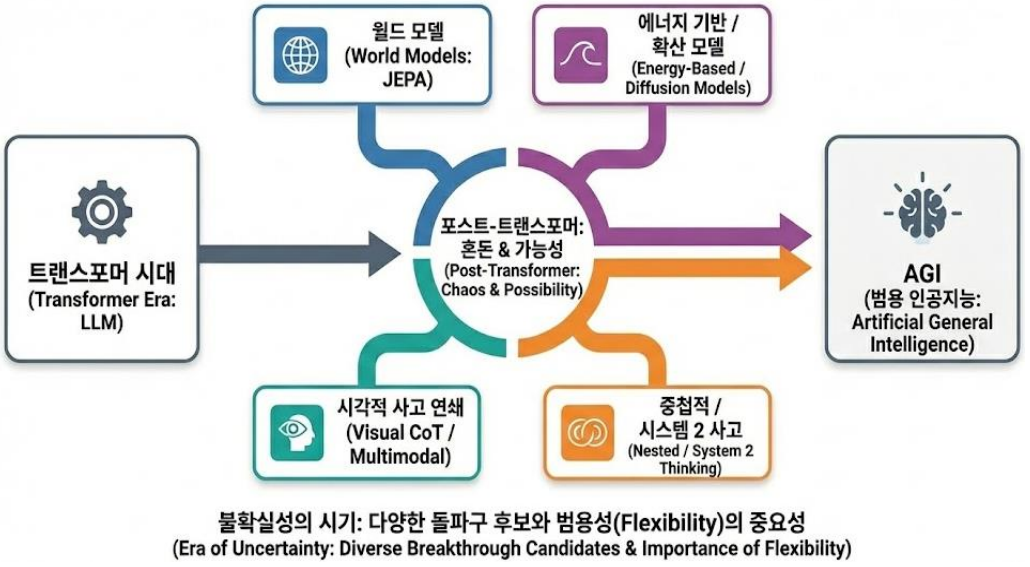
- 월드 모델, AGI, ASI(Super Intelligence)까지 도달하려면 추론과 기억 영역에서 여전히 큰 퀀텀점프가 필요한 상황
- 확률적 사고에서 벗어난 기억 및 추론의 안정화를 통한 일관성 확보, 결과에만 의존하고 과정을 중요시하지 않는 현재 훈련 구조 탈피, 데이터가 부족한 예외, Corner case에 대한 대응안 마련 등 해결 과제가 산적
- 이를 보완하기 위해 다양한 방식의 AI 모델 아키텍쳐가 실험되고 있는 상황에서는 실험용 수요가 중요
- 완제품을 돌리는 데 최적화되어 있으며 기능이 제한적인 TPU 등 ASIC이 시장의 대세가 된다는 주장은 시기상조
- 특히 신규 진입 플레이어가 늘어날 가능성이 높은 만큼 GPU의 호환성과 범용성의 가치는 미래에도 유지됨

그림 7. 전문가들은 스케일링 한계 우려



자료: Google, Anthropic, 상상인증권 리서치센터

그림 8. 현재 AGI 발전 경로는 누구도 예측하기 어려운 상황



자료: 상상인증권 리서치센터

Nvidia의 성능 상승폭과 전력효율성 증가에 주목할 필요

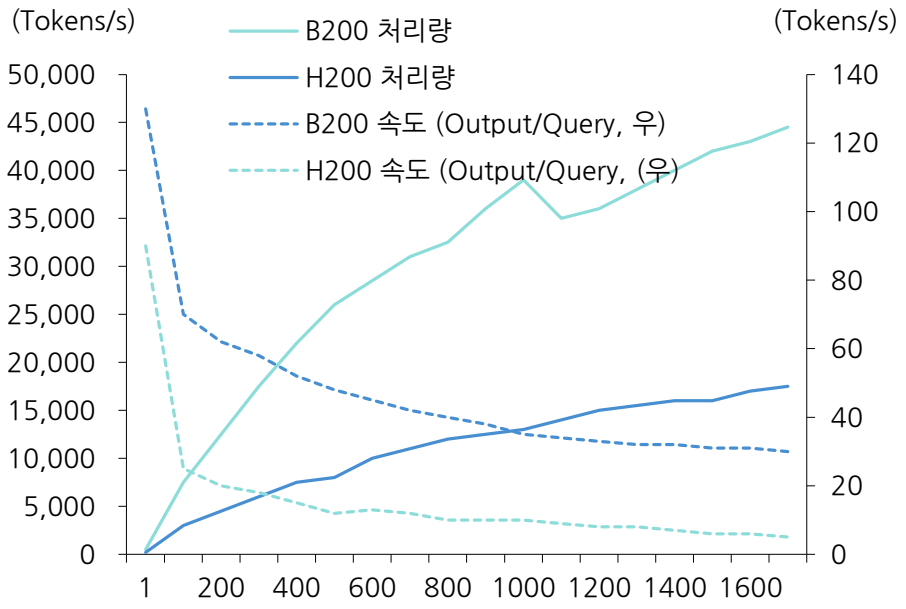
- 중국, 한국, 유럽 등 AI 스타트업들이 AI 경쟁에 본격 가세하면서 구형 칩에 대한 수요도 일정 수준 유지되는 상황
- TPU의 전성비 증가에도 GPU는 전성비와 절대 성능 양면에서 동시에 개선되며 화답 중이며 ASIC이 전력 소모가 적은 것은 사실이나 이제는 성능 상승폭이 전력 사용량 증가분을 상회하며 ASIC의 전성비를 추격하는 국면
- 과거 코드 호환성과 CUDA 소프트웨어 생태계의 관성을 감안하면 ASIC 점유율 증가 추세는 당분간 유지되겠으나 증가 속도는 가속이 아닌 감속 시나리오가 유력하며 GPU 추월은 난망
- 현 가격대에서도 GPU 완판이 지속되고 최신 모델 품귀 현상이 이어지는 한 마진폭 축소 또한 쉽지 않을 전망

표 1. Nvidia 세대별 스펙 비교

구분	A100 (Ampere)	H100 (Hopper)	B200 (Blackwell)	B300 (Blackwell Ultra)	R100 (Vera Rubin)
FP4 연산 (AI 추론)	미지원	미지원	20,000 TFLOPS	30,000 TFLOPS	미정
FP8 연산 (AI 학습/추론)	미지원	3,958 TFLOPS	10,000 TFLOPS	15,000 TFLOPS	미정
FP16 연산 (AI 학습)	624 TFLOPS	1,979 TFLOPS	5,000 TFLOPS	약 7,500 TFLOPS	미정
메모리 용량	80 GB	80 GB	192 GB	288 GB	미정
메모리 규격	HBM2e	HBM3	HBM3e (8-Hi)	HBM3e (12-Hi)	HBM4
메모리 대역폭	2.0 TB/s	3.35 TB/s	8.0 TB/s	8.0 TB/s	미정
트랜지스터	542억 개	800억 개	2,080억 개	2,080억 개	미정
TDP (소모전력)	400 W	700 W	1,000 W	1200 ~1400 W	미정
NVLink	NVLink 3 (600 GB/s)	NVLink 4 (900 GB/s)	NVLink 5 (1.8 TB/s)	NVLink 5 (1.8 TB/s)	NVLink 6 (3.6 TB/s)

자료: Nvidia, 상상인증권 리서치센터

그림 9. Nvidia H200 vs B200 LLM 처리량 및 속도

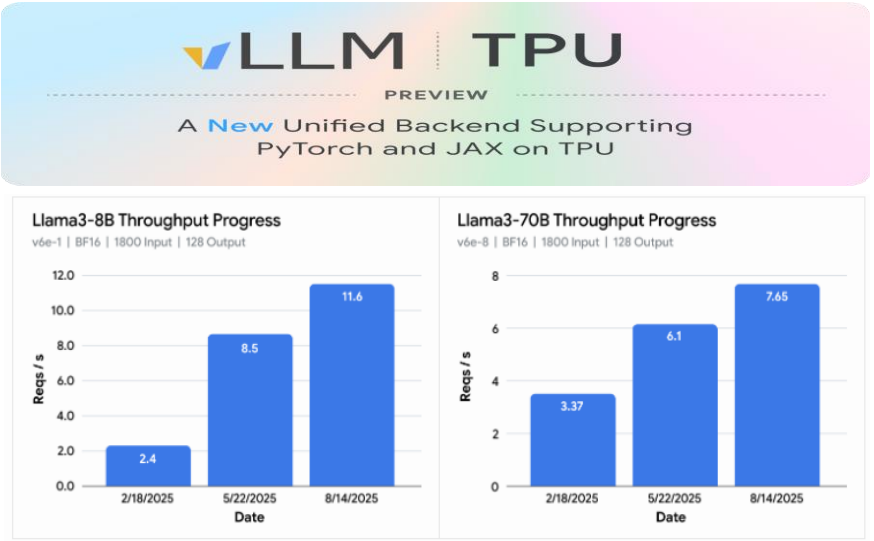


자료: AA, 상상인증권 리서치센터

CUDA가 약해졌다지만... 생각만큼 TPU 가성비가 높지 않다

- CUDA 생태계 약화의 핵심 논거인 vLLM의 TPU 지원은 이제 막 시작 단계이며 최적화가 미흡하여 벤치마크 기준 TPU는 Nvidia GPU 대비 금액/성능비가 최대 5배 이상 낮은 상황, vLLM 및 Red Hat에서도 TPU 지원은 아직 실험적 단계이며 JAX/PyTorch 통합에도 시간이 소요된다고 최근까지 인정한 바 있음
- Google은 기술 유출 방지를 위해 XLA:TPU 컴파일러, TPU 런타임, MegaScaler 등 핵심 코드를 오픈소스로 공개하지 않아 GPU처럼 자체 디버깅 및 유지보수가 제한되며 GCP 사용을 유도하는 구조로 오픈소스를 표방하나 핵심은 비공개
- Vera Rubin 아키텍처 출시가 26~27년으로 임박했으나 Nvidia는 생산 및 배포 측면에서 여전히 최소 1년 격차를 유지 중이며 빅테크 AI 엔지니어들은 특별한 유인책이 없는 한 익숙한 GPU를 선호하는 관성이 지속될 전망

그림 10. 10월에 “Preview” 버전을 발표한 vLLM



자료: vLLM, 상상인증권

표 2. 구형 GPU vs 구형 TPU 시장에서는 GPU 우세

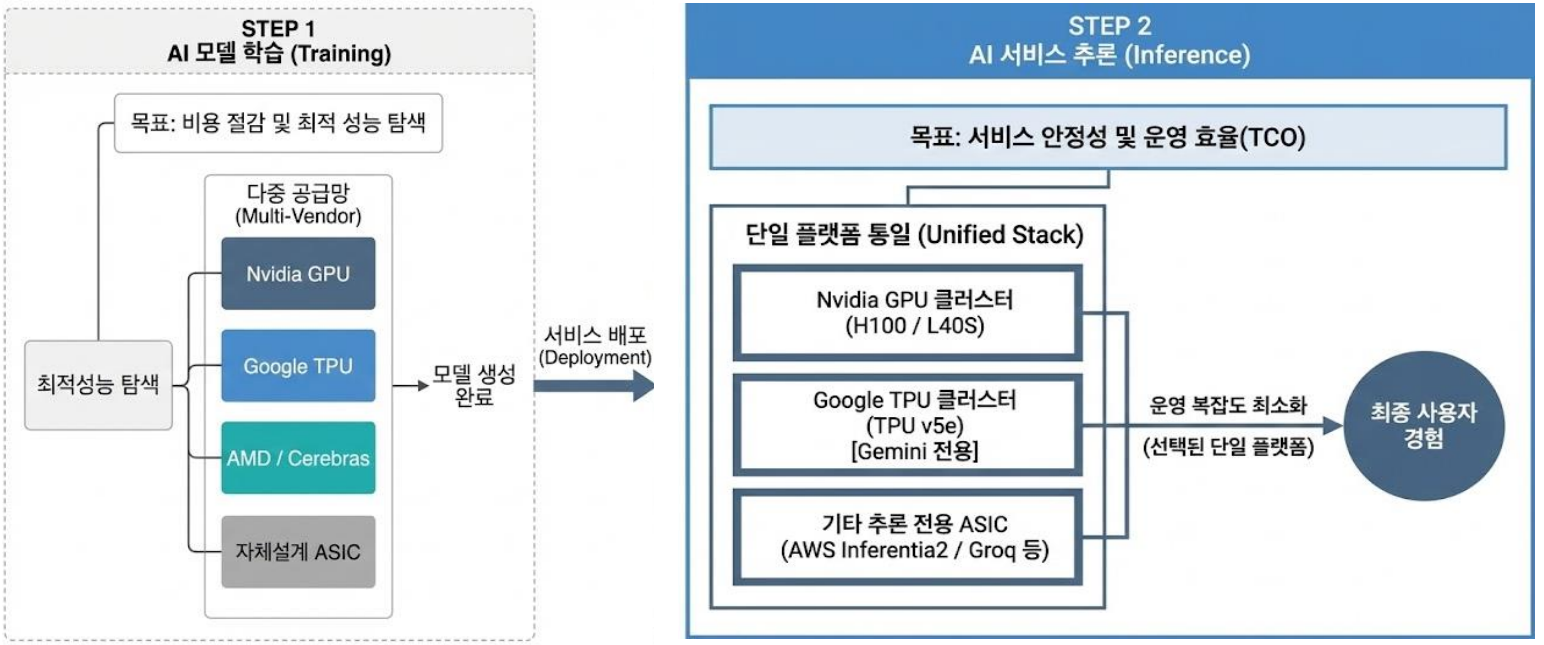
항목	우위	비고
VRAM (메모리)	NVIDIA H100	80GB vs 32GB
인터커넥트 (연결성)	NVIDIA H100	NVLink가 더 높은 대역폭을 제공함
첫 토큰 생성 시간 (TTFT)	TPU v6e	동시 요청이 적은(Low-concurrency) 추론 환경에서 더 빠름
비용 효율성 (가성비)	NVIDIA H100	성능 대비 비용($/Performance) 비율이 더 낮음
가용성 (접근성)	NVIDIA H100	GCP 외에도 다양한 클라우드에서 이용 가능
처리량 (Throughput)	NVIDIA H100	대규모 추론 시 초당 토큰 처리량 (Tokens/sec)이 더 많음

자료: HorizonIQ, 상상인증권

GPU가 추론의 시대에서도 널리 쓰이는 이유

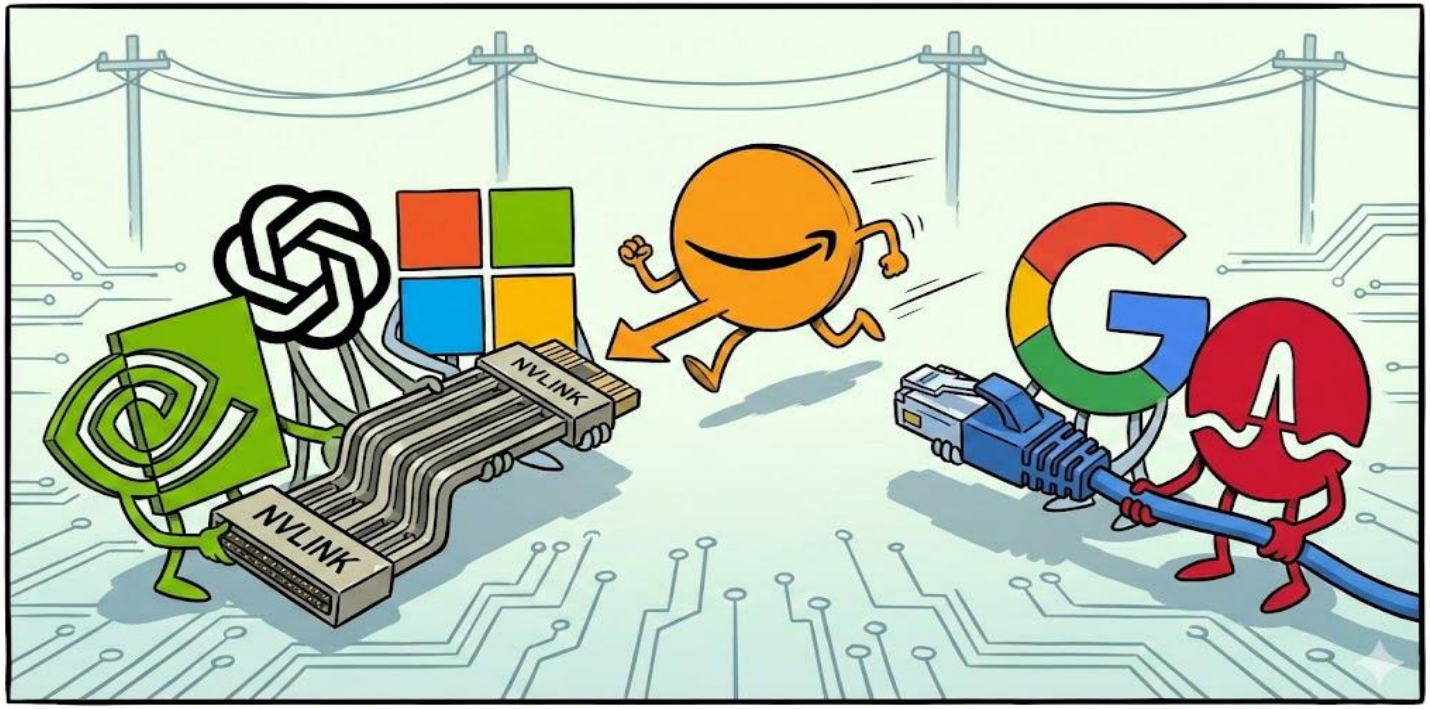
- Nvidia는 매 분기 호실적에도 불구하고 2024년 중순 이후 밸류에이션 레벨이 지속적으로 하락하는 중
- Nvidia, TPU 외에도 AMD, Sambanova, Cerebras, Rivos, Matx 등 훈련용 칩 옵션은 다양해졌으나 안정적인 LLM 서비스 특히 추론은 사용자 경험과 직결되기 때문에 통일된 하드웨어 옵션 선호 경향이 강함
- 칩 하나당 단가가 아닌 시스템 전체 단가 관점에서 보면 Nvidia 시스템은 가장 긴 Track Record를 보유한 검증된 방식
- 성능 및 관련 설비 지원 측면에서도 Nvidia가 우위에 선 상황이며 이는 GPU 진영에 유리하게 작용

그림 11. 추론 스케일링 시대의 하드웨어 통일성



자료: 상상인증권 리서치센터

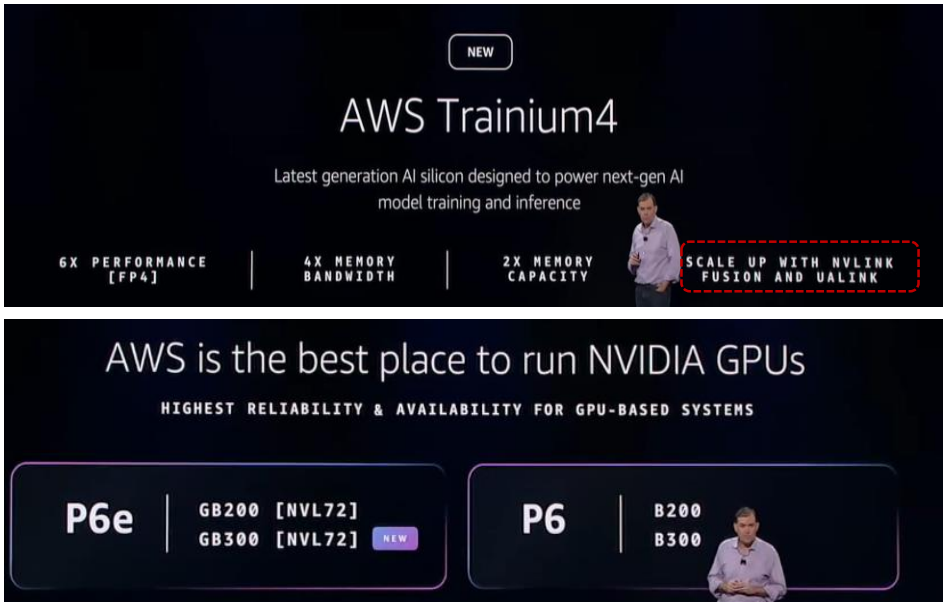
왕관의 무게를 견뎌야 하는 Nvidia



GPU의 NVLink/Infiniband vs ASIC의 RoCE/Ethernet

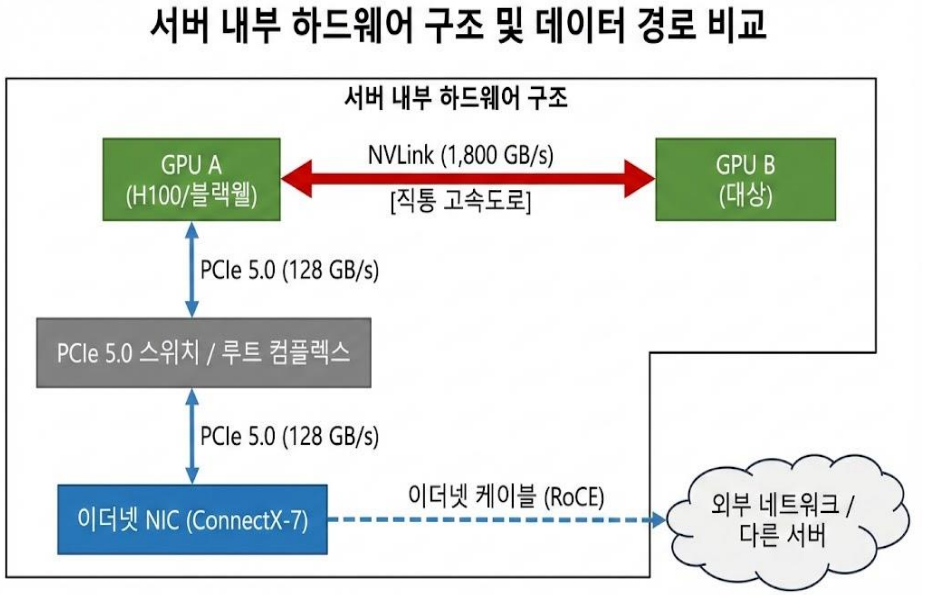
- 데이터센터 규모 확대에 따라 GPU vs TPU/ASIC 진영 간 핵심 전장 중 하나는 칩 간 네트워크 방식이며 그간 TPU/ASIC 진영은 자체 규격인 NVLink 대신 Ethernet 기반의 RoCE/iWARP 등 오픈소스 규격을 채택해왔음
- 그러나 데이터 통신 규모가 확대될수록 NVLink의 데이터 접근성과 GPU 호환성이 Ethernet 방식 대비 최종 속도 및 성능에 미치는 영향이 구조적으로 증가하는 추세. AI 순간 통신대역폭이 1~5TB/s에 달하는 상황
- GPU 의존도가 높은 OpenAI/Microsoft Azure는 이미 NVLink를 활용 중이며 AWS 또한 AI 데이터센터에서 기존 Marvell Technology + Ethernet 방식에서 NVLink 방식으로 전환 결정

그림 12. AWS, NVLink 통합을 선언



자료: Amazon

그림 13. RDMA방식은 PCIe 슬롯 병목 등 현실적 문제에 봉착

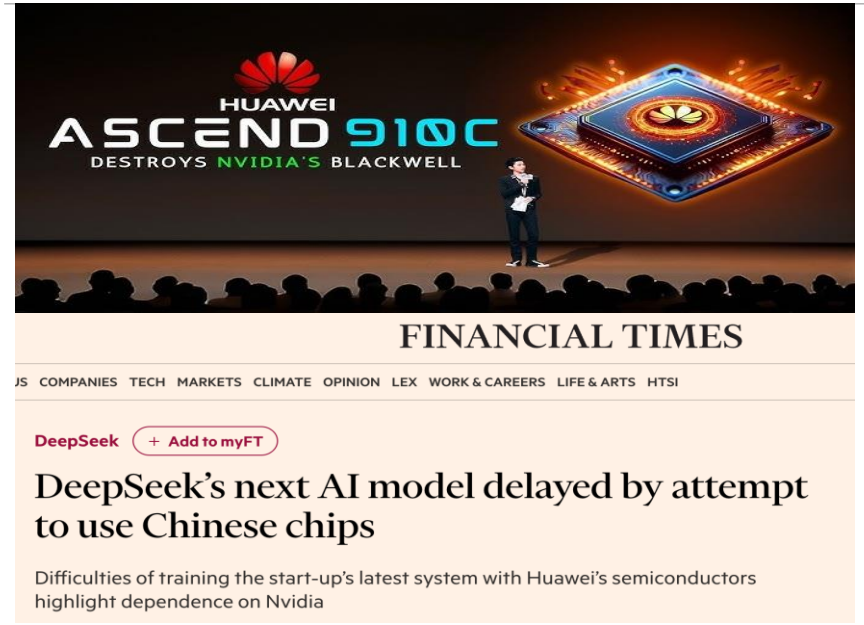


자료: 상상인증권 리서치센터

신규 경쟁자 등장 우려는 항상 존재했다는 점을 기억하자

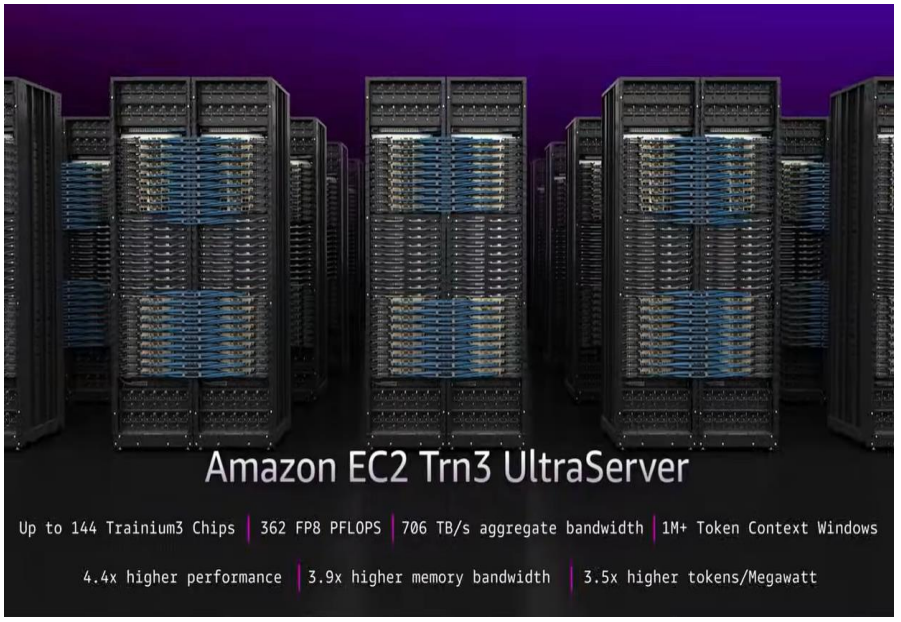
- Nvidia는 호실적에도 불구하고 현 Fwd PER ~23 근방을 유지. 주가 부진의 주 원인은 Nvidia의 사실상 독점 스토리와 CUDA 해자(Moat) 훼손 우려이며 Google 외에도 다양한 칩 경쟁자들이 지속 등장 중
- 다만 이는 AI 고도화에 따른 필연적 수순. 미국은 Amazon, Qualcomm, Tesla, Groq, Apple 등 Edge 추론 중심으로 중국은 화웨이 Ascend 등 한국은 퓨리오사, 리벨리온 등 AI 시장 성장과 함께 다양한 수요에 맞춘 공급이 창출되는 중
- Nvidia의 마진률이 장기 지속 불가하다는 우려는 꾸준히 제기되어 왔으며 Deepseek 사태, 화웨이 Ascend, Google TPU 발표 등으로 주가가 부진할 때마다 도리어 매수 기회로 작용해왔음을 상기할 필요

그림 14. 시장이 크게 반응했었던 화웨이 칩은 결국 실패



자료: Financial Times, Huawei

그림 15. Amazon Trainium 3 Ultraservers, AWS도 참전

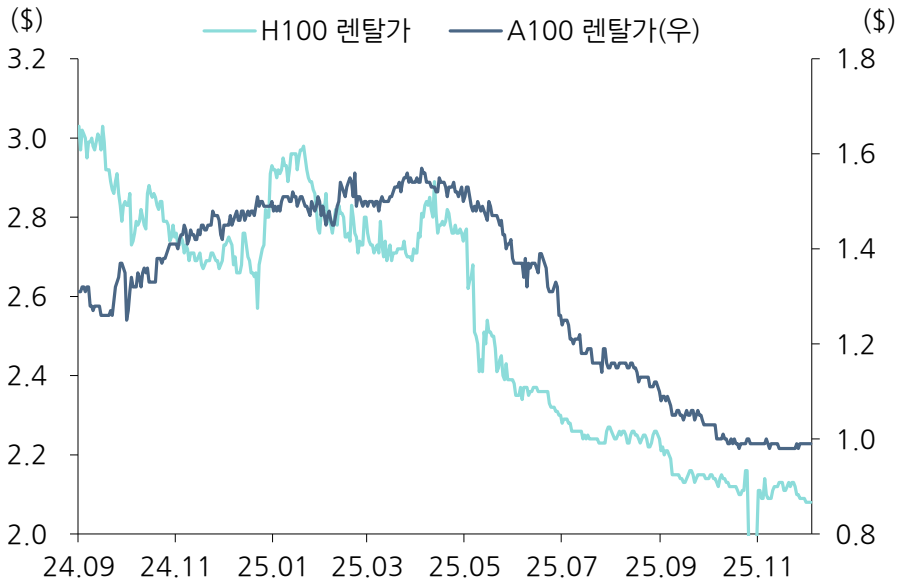


자료: Amazon

GPU 감가가 짧아지면 교체주기가 오히려 짧아진다

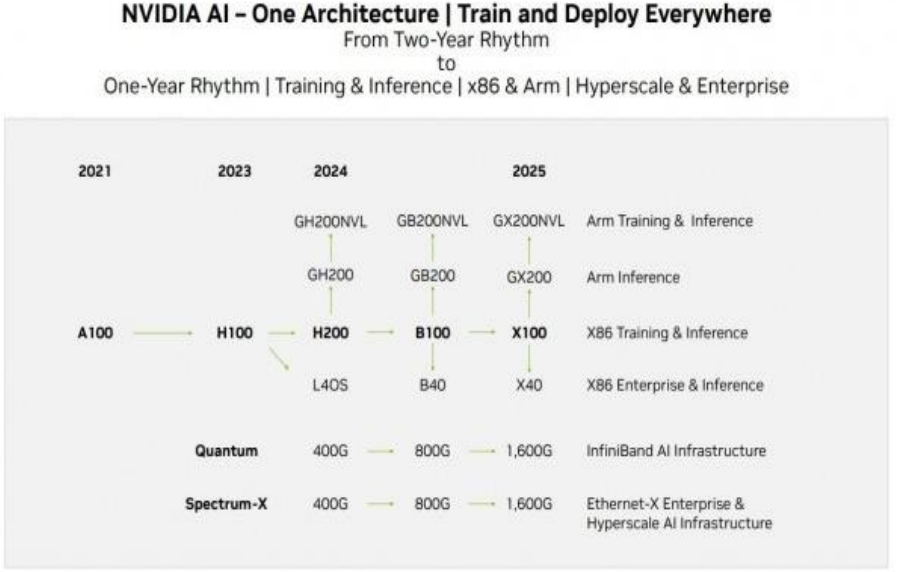
- 마이클 버리는 Oracle, Nebius 등 업체의 GPU 감가상각이 과소 반영되어 있다고 지적하며, 이를 통한 GPU 담보 자금 조달 및 순이익 과대계상 등 AI 밸류에이션이 대체로 과열되고 있다고 경고 (GPU 감가 3년->5년, 실제로는 2년 미만)
- GPU 수명이 단축되고 있다는 지적 자체는 일리가 있으며, 실제로 구세대 GPU는 시장에서 빠르게 퇴출되는 추세
- 그러나 감가상각 주기가 짧아지고 세대별 성능 및 전성비 격차가 벌어질수록 GPU 교체 수요를 앞당기는 요인
- 교체 주기 장기화 이후 업체들의 영업이익과 주가 모멘텀이 둔화되었던 전례를 감안하면 Nvidia에 오히려 긍정적

그림 16. GPU 감가 곡선: A100, H100



자료: Bloomberg, 상상인증권 리서치센터

그림 17. 블랙웰 울트라, 베라 루빈 출시일정 2년 -> 1년 단축

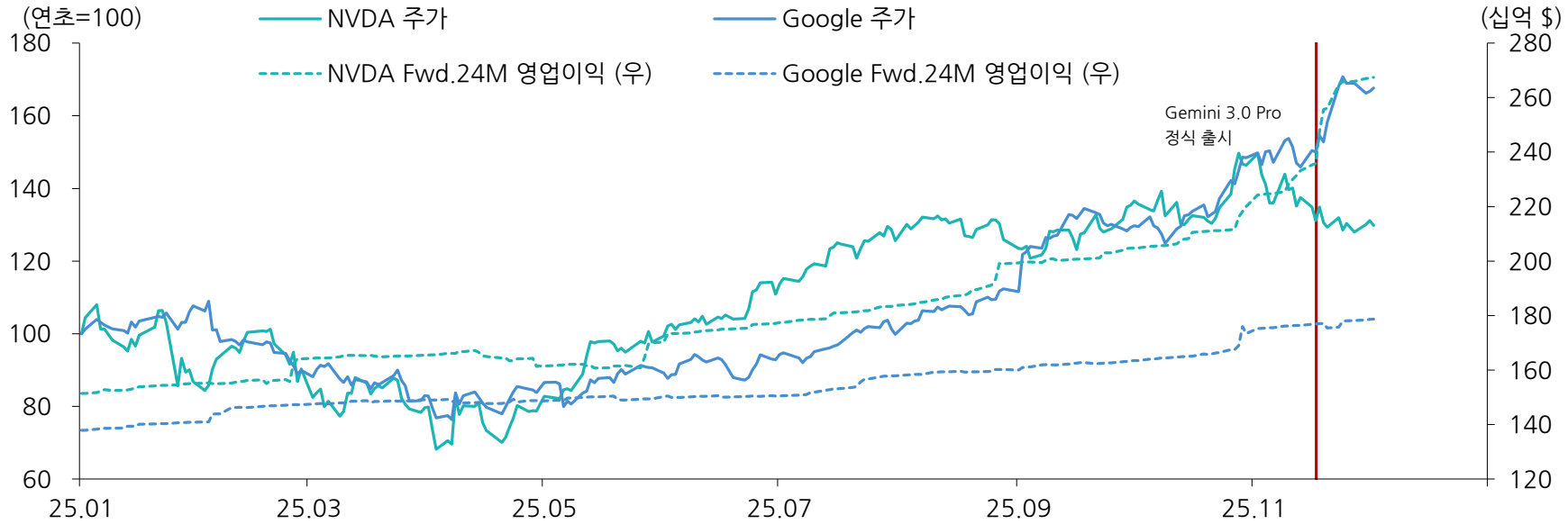


자료: Nvidia, 상상인증권 리서치센터

Nvidia, Google은 결국 둘 다 곡괭이 & 삽

- 연초 미 IB의 곡괭이 & 삽(Picks and Shovels) 전략 상기 필요: 구조적인 반도체 & AI 슈퍼사이클, 대세 상승장에서는 AI 인프라 격인 AI Accelerator와 클라우드 플랫폼 기업은 포트폴리오의 주축이 되어야 함
- 다양한 AI 칩 제작업체의 등장은 AI 시장 성숙 과정에서 시장의 다양한 수요에 맞추는 자연스러운 수순
- 전체 파이 확대 국면에 Nvidia와 Google은 'Frienemies' 관계로 협력하면서도 건전한 경쟁을 유발하는 존재가 될 것
- 따라서 AI 신모델 출시, 경쟁제품 등장 등 단기적 뉴스보다는 AI 인프라 시장 전체 성장에서 양사 모두 최대 수혜자라는 점에 주목하며, 둘 중 저평가 되어있는 종목에 대해 비중을 확대하는 전략이 가장 유효하다고 판단

그림 18. 연초 이후 Nvidia, Google 주가, 영업이익 추이



자료: Bloomberg, 상상인증권

Compliance Notice

- 본 자료에 기재된 내용들은 작성자 본인의 의견을 정확하게 반영하고 있으며 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다. (작성자: 김경태)
- 본 자료는 고객의 증권투자를 돕기 위한 정보제공을 목적으로 제작되었습니다. 본 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보를 바탕으로 작성한 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 참고자료로만 활용하시기 바라며 유가증권 투자 시 투자자 자신의 판단과 책임하에 최종결정을 하시기 바랍니다. 따라서 본 자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.
- 본 자료는 당사의 저작물로서 모든 저작권은 당사에게 있으며 어떠한 경우에도 당사의 동의 없이 복제, 배포, 전송, 변형될 수 없습니다.
- 동 자료는 제공시점 현재 기관투자가 또는 제3자에게 사전 제공한 사실이 없습니다.
- 동 자료의 추천종목은 전일 기준 현재당사에서 1% 이상 보유하고 있지 않습니다.
- 동 자료의 추천종목에 해당하는 회사는 당사와 계열회사 관계에 있지 않습니다.

상상인증권
sangsangin securities