

2025년 11월 27일 | 산업분석_Industry Report

Overweight

반도체

TPU 부각에 따른 영향

Meta가 Google의 TPU를 채용?

Meta가 2027년 가동에 들어갈 데이터센터에 Google의 TPU 사용을 검토하고 있다는 언론 보도가 있었다. Google의 TPU 연혁을 살펴보면, 2016년에 추론 전용칩인 TPU 1세대를 통해 구글 검색의 음성 인식 등 내부 서비스의 추론 작업 가속화에 집중해왔다. 2017년에는 훈련도 지원 가능한 TPU 2세대 칩을 출시했고, Cloud TPU 서비스를 통해 구글 클라우드 고객에게 최초로 공개했다. 이후 AI 모델 및 추론 영역을 강화해 왔고, 2025년 TPU 7세대 칩을 공개했다. 7세대 TPU는 이전세대대비 HBM 용량이 6배 증가한 192GB를 탑재해 대규모 모델 추론 효율을 극대화했다. 해당 칩이 이번 기사에서 언급된 칩으로 추정된다.

추론 AI는 예정된 경쟁. 훈련용에 생긴 균열?

Google의 자체 ASIC인 TPU를 포함해서 Microsoft, Amazon, Meta 등의 빅테크 업체들도 자사 ASIC 칩을 지속 개발 및 출하했다. 향후에도 추론 영역에서 자사 ASIC의 활용 범위를 확대하고자 하는 계획을 갖고 있었다. Nvidia 입장에서 추론 영역에서의 경쟁 심화는 예정된 수순이었다. 다만, Meta가 Google TPU 사용을 고민한다는 기사가 맞다면, 이는 훈련용으로 활용하려는 의도일 것으로 추정된다. 이는 Nvidia 입장에서 점유율에 일부 균열이 일어날 가능성에 노출된 것이다.

투자 비용 효율화에 대한 고민

Meta와 같은 빅테크 업체들은 Capex가 지속 상향되면서 비용 및 전력에 대한 부담이 가중되는 상황이었고, 효율적인 투자에 대한 하나의 대안을 시도하는 것으로 추정된다. 다만, Google의 TPU 역시 여타 ASIC과 마찬가지로 GPU처럼 범용, 호환성, 그리고 고정밀 연산 작업을 하기는 어려울 것이다. TPU는 AI 학습에 필수적인 행렬 연산을 위한 낮은 정밀도 연산에 최적화되었기 때문에 연산 속도 및 전력 소모 측면에서 유리하다. 반면에 GPU처럼 다양한 AI모델 학습, 과학 시뮬레이션이나 그래픽 렌더링과 같은 작업을 하기에는 한계가 있다. 결국 훈련용 AI 영역에서 GPU를 완전히 대체하기에는 무리가 있다고 판단한다. 다만, 일정 영역에서 TPU를 활용함으로써 비용 및 전력을 낮출 수 있을 것으로 전망한다. Nvidia 역시 추론 영역에서 에이전틱 AI 및 피지컬 AI 전용 개발자 플랫폼인 블루프린트, 옴니버스 등을 통해 선제적으로 고객사를 확보한 부분도 있기 때문에 훈련 및 추론 영역 모두에서 경쟁이 활발해질 것으로 예상된다.

메모리에 대한 영향은?

메모리 입장에서 보면 GPU가 HBM에 대한 의존도 및 스펙이 확실히 높다. 다만, ASIC 역시 스펙이 상향되며 고사양, 고용량 HBM이 탑재되고 있고, HBM의 수요처가 다변화된다는 측면도 고려해야 한다. 결국 HBM의 전체 수요는 우상향하는 것이기 때문에 부정적인 영향은 제한적이라 판단한다. 다만, 기존에 독보적인 지위를 갖고 있던 Nvidia의 GPU와 SK하이닉스의 HBM에 대한 프리미엄은 고민이 필요해진다. 새로운 도전자의 등장은 기존 강자들 입장에서 불편한 요소일 수밖에 없으며, 이런 경쟁을 통해 시장은 성장한다는 것을 상기해야 한다.



Analyst 김록호 roko.kim@hanafn.com
RA 김영규 kyg1019@hanafn.com

도표 1. Google의 TPU 연혁

세대	V1	V2	V3	V4	V5e	V5p	V6e	V7
공개	2016년	2017년	2018년	2021년	2023년	2023년	2024년	2025년
공정	28nm	16nm	16nm	7nm	5nm	5nm	3nm	3nm
크기	311mm ²	<625mm ²	<700mm ²	<400mm ²	350mm ²	700mm ²	790mm ²	-
메모리	8GB DDR3	16GB HBM	32GB HBM	32GB HBM	16GB HBM	95GB HBM	32GB HBM	192GB HBM
대역폭	34GB/s	600GB/s	900GB/s	1.2TB/s	819GB/s	2.8TB/s	1.6TB/s	7.2TB/s
초당 연산 성능 (int8)	23 TOPs	92 TOPs	246 TOPs	550 TOPs	394 TOPs	918 TOPs	1,836 TOPs	-
초당 연산 성능 (bfloat16)	11.5 TFLOP	46 TFLOP	126 TFLOP	275 TFLOP	197 TFLOP	459 TFLOP	918 TFLOP	-

자료: Google, 언론보도 종합, 하나증권

도표 2. 빅테크 AI 로드맵

CSP	자체 칩	노드	협력사	메모리 타입	단일 칩 용량	메모리 개수	메모리 총 용량	출시일
Google	TPU v5 Inference	5nm	Broadcom	HBM2e	8	2	16	2024년
	TPU v5 Training	5nm	Broadcom	HBM2e	16	6	96	2024년
	TPU v6 Inference	3nm	Broadcom	HBM3	16	2	32	2024년
	TPU v7 Training	3nm	Broadcom	HBM3e	24	8	192	2025년 이후
	TPU v7 Inference	3nm	MTK	HBM3e	36	6	216	2026년 하반기 이후
Amazon	Inferentia v2	7nm	Alchip	HBM2e	16	2	32	2024년
	Trainium v2/2.5	5nm	Marvell	HBM3/3e	24/36	4	96/144	2024년~2026년 상반기
	Inferentia v2.5	5nm	Marvell	HBM3	24	2	48	2025년
	Trainium v3	3nm	Alchip	HBM3e	36	8	288	2026년 1분기
Microsoft	Maia v1	5nm	GUC	HBM2e	16	4	64	2024년
	Maia v2	3nm	GUC	HBM3e	24	4	96	2026년 1분기 이후
	Maia v3	3nm	Marvell	HBM4	TBD	TBD	TBD	2027년
Meta	MTIA v1	7nm	Broadcom	LPDDR5	8	16	128	2024년
	MTIA v2	5nm	Broadcom	LPDDR5	16	16	256	2025년 4분기 이후
	MTIA v3	3nm	Broadcom	HBM3e	TBD	TBD	TBD	2026년 상반기

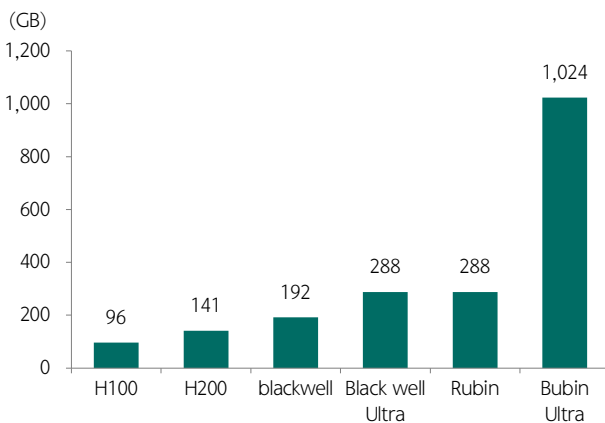
자료: Trend Force, 하나증권

도표 3. TPU vs GPU

	Google TPU v6e	Nvidia H100/H200
VRAM	32GB	80GB/141GB
HBM 속도	≈ 1.5TB/s	≈ 3~4.8TB/s
Interconnect	~450GB/s	~900GB/s
Compute(FLOPS)	≈ 2 PFLOP/s	~1~1.2 PFLOP/s
Cost	\$21.6/h(8 unit), 1유닛당 \$2.7	\$2.99~\$7.98/h
Software	TensorFlow/JAX	CUDA/Pytorch

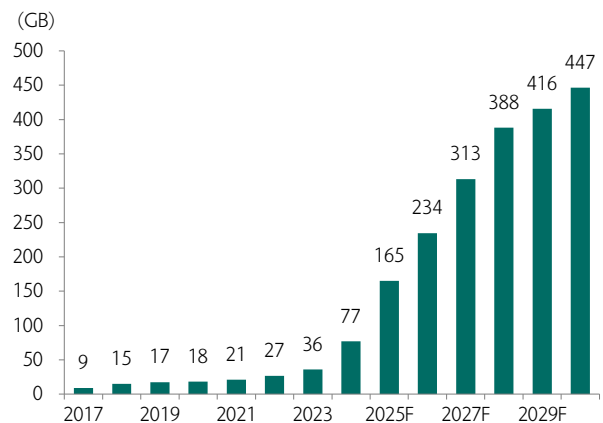
자료: Horizon IQ, 하나증권

도표 4. Nvidia GPU 탑재 HBM 용량



자료: Nvidia, 하나증권

도표 5. ASIC HBM 용량 전망 추이 및 전망



자료: 산업자료, 하나증권

Compliance Notice

- 당사는 2025년 11월 27일 현재 해당회사의 지분을 1%이상 보유 하고 있지 않습니다
- 본 자료를 작성한 애널리스트(김록호)는 자료의 작성과 관련하여 외부의 압력이나 부당한 간섭을 받지 않았으며, 본인의 의견을 정확하게 반영하여 신의성실 하게 작성하였습니다.
- 본 자료는 기관투자자 등 제 3자에게 사전 제공한 사실이 없습니다
- 본자료를 작성한 애널리스트(김록호)는 2025년 11월 27일 현재 해당회사의 유가증권을 보유하고 있지 않습니다

본 조사자료는 고객의 투자에 정보를 제공할 목적으로 작성되었으며, 어떠한 경우에도 무단 복제 및 배포 될 수 없습니다. 또한 본 자료에 수록된 내용은 당사가 신뢰할 만한 자료 및 정보로 얻어진 것이나, 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 최종결정을 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.

투자등급 관련사항 및 투자의견 비율공시

- 투자의견의 유효기간은 추천일 이후 12개월을 기준으로 적용

• 기업의 분류

BUY(매수)_목표주가가 현재주가 대비 15% 이상 상승 여력
Neutral(중립)_목표주가가 현재주가 대비 -15%~15% 등락
Reduce(비중축소)_목표주가가 현재주가 대비 15% 이상 하락 가능

• 산업의 분류

Overweight(비중확대)_업종지수가 현재지수 대비 15% 이상 상승 여력
Neutral(중립)_업종지수가 현재지수 대비 -15%~15% 등락
Underweight(비중축소)_업종지수가 현재지수 대비 -15%~15% 등락

투자등급	BUY(매수)	Neutral(중립)	Reduce(매도)	합계
금융투자상품의 비율	95.71%	4.29%	0.00%	100%

* 기준일: 2025년 11월 24일