iM증권 리서치본부

[반도체]

Overweight

Nvidia에 대한 위협 요소와 대응

2026.03.09

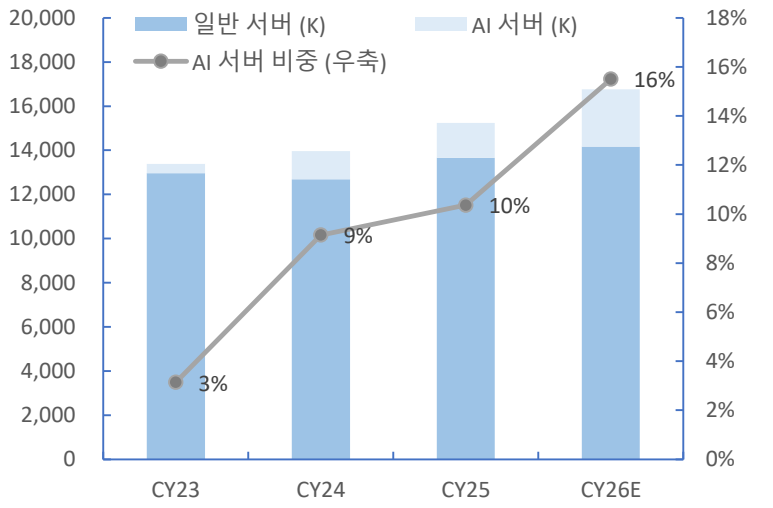
In-Depth Report

[반도체]
송명섭 2122-9207
mssong@imfnsec.com

CY26: AI 반도체 시장 급속 성장하나 Nvidia M/S 소폭 하락

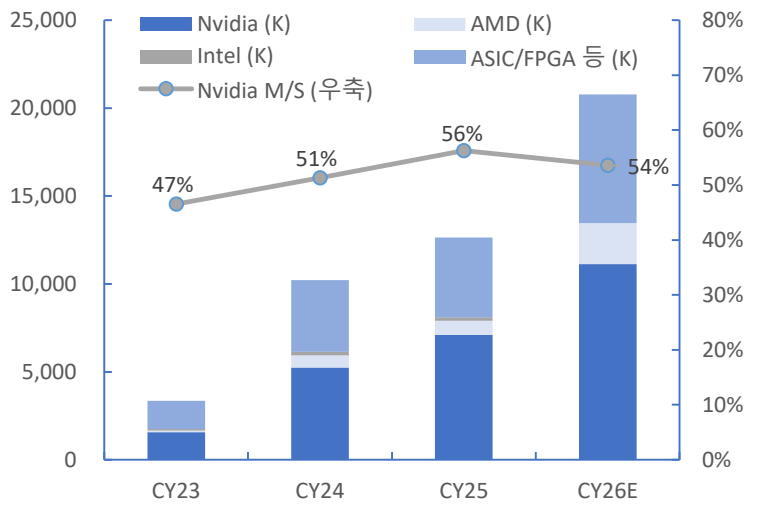
- CY26 전체 서버 출하량은 16.8백만대 (YoY +10%), AI 서버 출하량은 2.6백만대 (YoY +65%)를 기록하여 AI 서버 비중이 전년의 10%에서 16%로 대폭 상승할 전망임
- 최근 빅테크 업체들의 Capex 전망치 상향 조정과 Memory 반도체 구매량 대폭 증가를 감안 시, 올해 전체 서버 및 AI 서버의 출하량 증가율은 당사 현재 전망치를 상회할 가능성도 존재함
- 이에 따라 CY26 AI 가속기 반도체 출하량도 대폭 증가하나, 최근 수년간 지속 상승해 온 Nvidia의 M/S는 CY26에 54%를 기록해 전년의 56%에서 소폭 하락할 것으로 전망됨
- Nvidia에 대한 위협 요소는 전체 AI 서버 및 가속기 반도체 출하량의 증가율이 하락할 경우와 동사 자체 경쟁력이 하락할 경우로 나눠볼 수 있음

<그림1> 서버 출하량 및 AI 서버 비중



자료 : iM증권 리서치본부

<그림2> 가속기 반도체 출하량 및 Nvidia M/S



자료 : iM증권 리서치본부

Server: CY26 서버 출하 증가율이 10% 이상에 도달할 듯

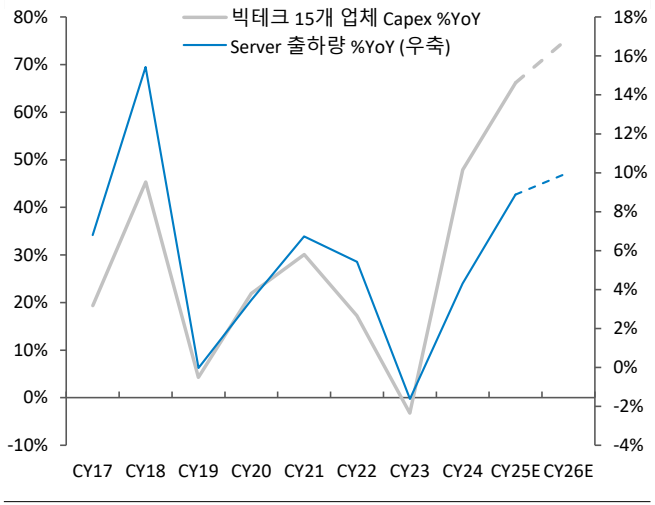
- 2월 25일 집계된 Neo Cloud 업체 포함 빅테크 업체 15개사의 FY25 Capex 증감률은 11월 20일 집계치 +64.6%에서 +68.7%로 상승. 당사는 이를 반영하여 CY25 서버 출하 증가율 전망치를 8.7%에서 9.2%로 상향
- Memory 가격 급등에도 불구하고 Trendforce, IDC, Digitimes 등은 CY26 서버 출하량이 약 10 ~ 13% 증가할 것으로 예상하며 기존 증가율 전망치를 최근 상향
- 2월 25일 현재 시장의 FY26 Capex 증감률 전망치는 +66.3%로 CY25 11월 20일 전망치 32.1%에서 대폭 상승한 것이고 향후 76% 이상까지 상향 조정될 가능성이 있는 것으로 보임. 빅테크 Capex 증가율과 서버 출하량 증가율 간의 비례 관계를 고려 시 76% 이상의 Capex 증가율이 달성된다면 CY26 서버 출하 증가율은 10%에 도달할 전망
- 또한 중국 업체들은 CY26 서버 생산량을 100% 증가시킬 것으로 언급 중이므로, Memory 가격 급등에 따른 비용 증가 요소를 감안해도 CY26 서버 출하량 증가율은 10% 이상에 도달할 가능성이 충분한 듯

<표1> 주요 빅테크 업체들의 Capex 추이 및 전망

Capex ($mn)	FY16	FY17	FY18	FY19	FY20	FY21	FY22	FY23	FY24	FY25	FY26E 11월 20일	FY26E 2월 25일
Google	10,212	13,184	25,139	23,548	22,281	24,640	31,485	32,251	52,535	91,447	114,615	180,502
Apple	12,734	12,451	13,313	10,495	7,309	11,085	10,708	10,959	9,447	12,715	13,988	13,509
Meta	4,491	6,733	13,915	15,102	15,163	18,690	31,431	27,266	37,256	69,691	108,373	123,629
Amazon	7,804	11,955	13,427	16,861	40,140	61,053	63,645	52,729	82,999	131,819	144,264	184,732
Microsoft	8,343	8,129	11,632	13,925	15,441	20,622	23,886	28,107	44,477	64,551	97,114	105,034
IBM	3,567	3,229	3,395	2,286	2,618	2,062	1,346	1,245	1,048	1,617	1,570	1,607
Oracle	1,189	2,021	1,736	1,660	1,564	2,135	4,511	8,695	6,866	21,215	35,804	50,386
Paypal	669	667	823	704	866	908	706	623	683	852	1,000	958
eBay	626	666	651	508	463	444	449	456	458	525	544	555
Salesforce	710	464	534	595	643	710	717	798	736	658	721	695
Alibaba	1,706	2,609	4,509	7,399	6,517	6,379	8,311	5,014	4,594	11,907	15,066	17,594
Tencent	1,492	1,801	3,356	3,927	5,719	4,808	3,451	3,017	9,675	13,193	12,791	13,938
Baidu	631	708	1,327	931	738	1,689	1,232	1,580	1,130	1,894	1,577	1,736
CoreWeave	-	-	-	-	-	-	72	2,943	8,702	14,864	25,840	26,519
NeBius	144	212	453	318	341	606	750	1,073	808	4,066	9,265	12,192
Total Capex	54,318	64,829	94,211	98,258	119,803	155,831	182,699	176,756	261,414	441,014	582,532	733,586
미국 IDC Capex	50,345	59,499	84,565	85,684	106,488	142,349	168,884	163,129	236,505	395,090	517,993	661,606
중국 IDC Capex	3,828	5,117	9,193	12,256	12,973	12,876	12,993	9,611	15,399	26,994	29,434	33,268
Neo Cloud Capex	144	212	453	318	341	606	822	4,016	9,510	18,930	35,105	38,711
Total %YoY	19.8%	19.4%	45.3%	4.3%	21.9%	30.1%	17.2%	-3.3%	47.9%	68.7%	32.1%	66.3%
미국 업체 %YoY	20.7%	18.2%	42.1%	1.3%	24.3%	33.7%	18.6%	-3.4%	45.0%	67.1%	31.1%	67.5%
중국 업체 %YoY	11.7%	33.7%	79.6%	33.3%	5.9%	-0.7%	0.9%	-26.0%	60.2%	75.3%	9.0%	23.2%
Neo Cloud %YoY	-32.9%	47.1%	113.2%	-29.9%	7.4%	77.5%	35.7%	388.6%	136.8%	99.1%	85.4%	104.5%

자료 : Bloomberg

<그림3> 빅테크 Capex와 Server 출하량 증감률 비교



자료 : Bloomberg, IDC, 대만 서버 ODM, iM증권 리서치본부

Server: CY26 CoWoS 할당량 및 AI 가속기 반도체 생산은 당초 예상을 상회

- TSMC는 당초 보수적으로 CoWoS 설비를 확장할 계획이었으나 가속기 반도체 업체들의 요청에 따라 확장 속도를 앞당긴 듯
- 가속기 반도체에 대한 CY26 CoWoS Capa 할당량은 CY25의 635K에서 1,380K (TSMC 1,080K, ASE 216K, Amkor 84K)로 117% 증가할 전망임. CY26 Nvidia에 대한 CoWoS Capa 할당량도 828K로 110% 증가할 듯. AMD, Broadcom에 대한 CY26 CoWoS Capa 할당량은 50K -> 193K, 76K -> 166K로 증가 예상
- Nvidia에 대한 CoWoS Capa 할당량과 각 GPU의 Net Die 수를 감안하면 동사 CY26 AI GPU 생산량은 YoY 57% 증가할 듯. Net Die 수가 적은 Rubin의 생산 비중이 당초 예상보다 낮을 전망인 점은 동사 AI GPU 생산량 증가율 상승 요소. 전체 가속기 반도체 생산량은 YoY 65% 증가할 듯

<표2> CoWoS Capa 할당량에 기반한 CY26 가속기 반도체 생산량 추정

CY26	Nvidia AI GPU 생산량	Net Die	B200	15	
			B300A	30	
			B300	15	
			중국 전용 (H200)	29	
			Rubin	9	
		CoWoS (K)		828	110%
		비중	B200	0%	
			B300A	5%	
			B300	40%	
			중국 전용 (H200)	5%	
			Rubin	50%	
		Chip 수 (K)	B200	0	
			B300A	1,242	
			B300	4,968	
			중국 전용 (H200)	1,201	
			Rubin	3,726	
			Total	11,137	57%
	AMD AI GPU 생산량	Net Die	MI350X	14	
			MI400X	10	
		CoWoS (K)		193	286%
		비중	MI350X	50%	
			MI400X	50%	
		Chip 수 (K)	MI350X	1,352	
			MI400X	966	
			Total	2,318	190%

CY26	Broadcom AI ASIC 생산량	Net Die	TPU	20	
			MTIA	30	
			Bytedance	40	
			OpenAI등	16	
		CoWoS (K)		166	118%
		비중	TPU	45%	
			MTIA	30%	
			Bytedance	15%	
			OpenAI등	10%	
		Chip 수 (K)	TPU	1,490	
			MTIA	1,490	
			Bytedance	994	
			OpenAI등	265	
			Total	4,239	70%
	기타 업체 생산량	Net Die	Avg.	16	
		CoWoS (K)		193	68%
		Chip 수 (K)		3,091	60%
	Total AI 가속기 생산량	Chip 수 (K)		20,786	65%

자료 : 대만 서버 ODM, TSMC, iM증권 리서치본부

HBM 업황: CY26 수급 안정 예상

- CY25 HBM 실수요량은 21.7억 GB로 YoY 107% 증가한 것으로 보임. CY25 업계 생산량은 28.8억 GB (SK하이닉스 16.3억, 삼성전자 7.5억, Micron 5.0억)로 연초 계획치인 37.5억 GB에서 대폭 하향되었으나 여전히 수요량을 대폭 상회. CY25에 HBM 가격이 하락한 이유

- 현재 가속기 반도체 생산 전망을 감안 시, CY26 Nvidia의 HBM 실수요량은 28.6억 GB로 YoY 96% 증가하고 전체 실수요량은 48.9억 GB로 126% 증가할 전망임. 이는 가속기 반도체 생산량 전망치 상향에 따라 당초 전망치보다 대폭 상향 조정된 것임

- 현재 CY26 생산량 전망치는 46.9억 GB (SK하이닉스 25.0억, 삼성전자 13.8억, Micron 8.1억) 수준으로 수요량을 하회해 수급 및 가격 안정 예상. 삼성전자는 HBM Capa를 모두 활용할 경우 16.3억 GB 생산이 가능하나 이익률이 더 높은 Legacy DRAM 생산량 증가를 우선시. 또한 HBM4 생산에 어려움을 겪고있는 것으로 알려진 Micron이 HBM을 줄이고 Legacy DRAM을 늘릴 가능성도 존재

<표3> CY25, CY26 HBM 실수요량 추정

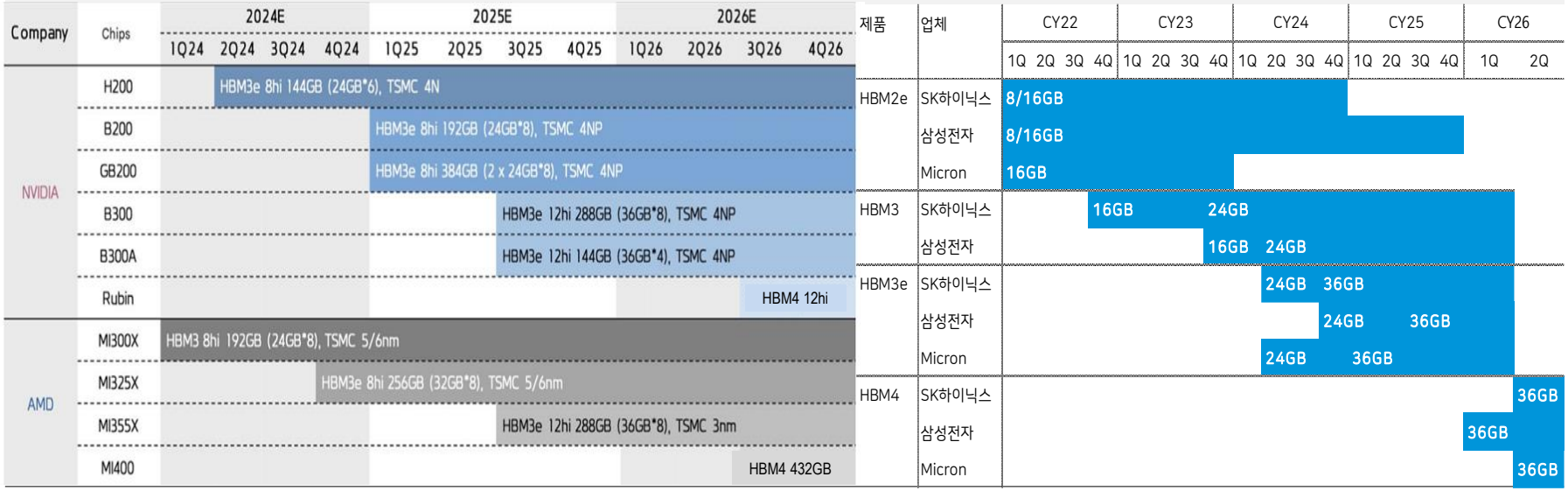
	고객	가속기 반도체		HBM					
		제품명	생산량 (K)	제품명	개당 용량 (GB)	가속기 당 개수	가속기 당 총용량	수요 (M GB)	% YoY
CY25	Nvidia	H100	0	HBM3	16	5	80	0	
		H200	1,257	HBM3E	24	6	144	181	
		New H20	0	HBM3E	24	6	144	0	
		B200	2,482	HBM3E	24	8	192	477	
		B300A	1,182	HBM3E	36	4	144	170	
		B300	2,187	HBM3E	36	8	288	630	
		Total	7,108					1,458	136%
	AMD	MI325X/350	800	HBM3E	36	8	288	230	70%
	Intel	Gaudi3	200	HBM2E	16	8	128	26	18%
	기타	ASIC/FPGA	4,525	HBM3/3E	20	5	100	453	67%
	Total		12,633				171	2,166	107%
CY26	Nvidia	B200	0	HBM3E	24	8	192	0	
		B300A	1,242	HBM3E	36	4	144	179	
		B300	4,968	HBM3E	36	8	288	1,431	
		중국 전용 (H200)	1,201	HBM3E	24	6	144	173	
		Rubin	3,726	HBM4	36	8	288	1,073	
		Total	11,137					2,856	96%
	AMD	MI350X	1,352	HBM3E	36	8	288	389	
		MI400X	966	HBM4	48	8	384	371	
		Total	2,318					760	230%
	Broadcom	TPU	1,490	HBM3E/4	27	10	261	389	
		MTIA	1,490	HBM3E/4	24	6	134	200	
		Bytedance	994	HBM3E	24	5	115	114	
		OpenAI 등	265	HBM4	32	12	384	102	
		Total	4,239					806	115%
	기타	ASIC/FPGA	3,091	HBM3E/4	24	6	151	467	
	Total		18,037				271	4,889	126%

자료 : 대만 서버 ODM, 각사 자료, iM증권 리서치본부

HBM4 현황

- HBM4의 일부 Revision에 들어갔던 것으로 알려진 SK하이닉스는 1월에 최종 샘플을 제출한 것으로 보임. 인증 통과 후 2Q26초경 1B 나노 베이스 HBM4 초도 공급이 가능할 듯
- 입출력 속도 문제가 있는 Micron은 1H26에 Nvidia향 HBM4 공급은 어려워 보이고 2H26에 주로 타 고객향으로 10억 Gb 수준의 HBM4를 판매할 것으로 예상됨
- 1Q26말에 HBM4 공급을 개시할 전망인 삼성전자는 Core DRAM에 1C 나노, Logic Die에 4 나노를 채택해 속도 측면에서도 유리
- CY26 HBM4 시장 점유율은 SK하이닉스 68%, 삼성전자 26%, Micron 6%으로 예상됨. 단 1H26에는 업계 전반적으로 HBM3E 위주의 공급이 진행되고 HBM4의 공급은 주로 2H26에 집중될 전망. CY26에 SK하이닉스는 Nvidia가, 삼성전자는 Broadcom이 최대 고객이 될 전망
- 일부 외신에서 Nvidia가 요구 입출력 속도를 11.7Gbps에서 하향 조정했다는 보도가 사실이라면, 이는 시장 점유율이 높은 업체들의 입출력 속도가 동 기준에 부합하지 못함에 따른 어쩔 수 없는 선택이었을 가능성이 큰 듯

<표4> Nvidia와 AMD GPU 별 채용 HBM과 업체별 HBM 양산 시기

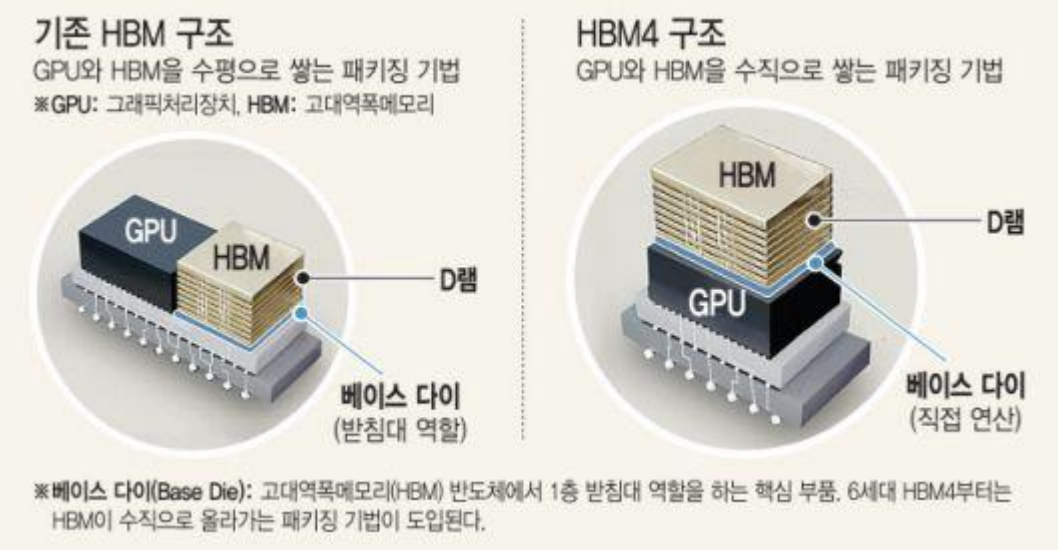


자료 : Trendforce, 각사 자료

HBM 가격 전망: 당초 우려보다 양호한 가격 예상

- 최근 HBM 계약은 고객의 연간 최대 구매 물량만 정하고 실구매량에는 제약이 없는 것으로 보임. 또한 연간 가격을 일정 구간 대로 설정하고 분기별로 구간 내에서 가격을 결정하는 방식인 것으로 추정됨
- 최대 고객은 CY25에 HBM 공급 경쟁 심화를 유도하고 수요 증가율 둔화를 이용해 HBM 가격을 분기 단위로 인하해 왔음
- 단 최근 전반적인 DRAM 공급 부족 상황과 HBM 수요 전망치 상향에 따라, 가격 협상에서 다른 DRAM 제품과 HBM을 패키지로 공급하는 조건으로 가격을 당초 고객 제시치보다 인상하려는 Memory 반도체 업체들의 시도가 있는 것으로 보임
- CY26의 연간 HBM 가격은 아직 확정되지 않은 듯. 단 HBM3E 12단 가격의 경우 여전히 YoY로는 하락할 가능성이 높으나 당초 우려되었던 것보다 훨씬 높은 수준으로 결정될 전망임

<그림4> HBM4의 구조 변화



자료 : 이데일리

빅테크 업체들의 막대한 AI 투자 지속 여부에 대한 의심

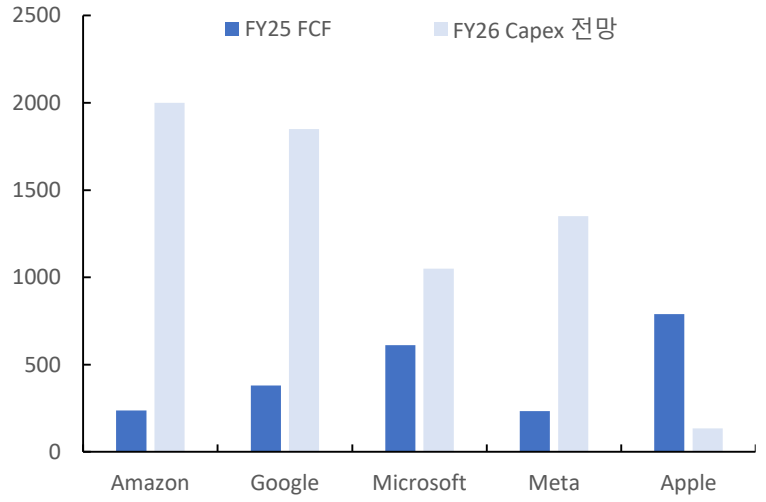
- Nvidia에 대한 Risk 요인 중 하나는 빅테크 업체들이 FY27 이후에도 현재와 같은 막대한 투자를 지속할 수 있을지에 대한 의심임. Apple을 제외한 모든 주요 빅테크 업체들의 FY26 Capex 전망치는 FY25 FCF의 1.7배 (MS) ~ 8.4배 (Amazon)에 달함. 따라서 일부 업체의 경우 대규모의 차입이 불가피할 전망임
- 빅테크 기업들의 천문학적인 Capex는 FY26 이후 감가상각비의 급격한 상승으로 이어지며 이익률에 본격적인 압박을 가할 것으로 예상됨
- 빅테크 업체들의 매출 대비 감가상각비 비중은 FY25의 한자릿수대 후반에서 FY26에는 20% 수준까지 치솟을 것으로 추정
- 골드만삭스에 따르면 CY25 가동을 시작한 데이터센터들은 연간 400억 달러의 감가상각비를 발생시키나, 현재 이용률 기준 창출 매출은 150억 ~ 200억 달러에 불과해 단기적으로 이익률을 저해할 가능성이 높음
- 이익률 하락을 방어하기 위해 빅테크 업체들은 서버 및 네트워크 장비의 내용연수 연장 조치를 취하고 있음. Meta는 서버 내용연수를 기존 4 ~ 5년에서 5.5년으로 연장. 그러나 실제 교체 주기가 회계상 내용연수보다 짧아질 경우, 향후 한번에 막대한 자산 손상 차손이 발생할 위험 상존

<표5> 빅테크 업체들의 FY25 FCF와 FY26 Capex 전망치

(억$)	FY25 FCF	FY26 Capex 전망
Amazon	238	2,000
Google	381	1,850
Microsoft	611	1,050
Meta	234	1,350
Apple	789	135

자료 : iM증권 리서치본부

<그림5> 빅테크 업체들의 FY25 FCF와 FY26 Capex 전망치

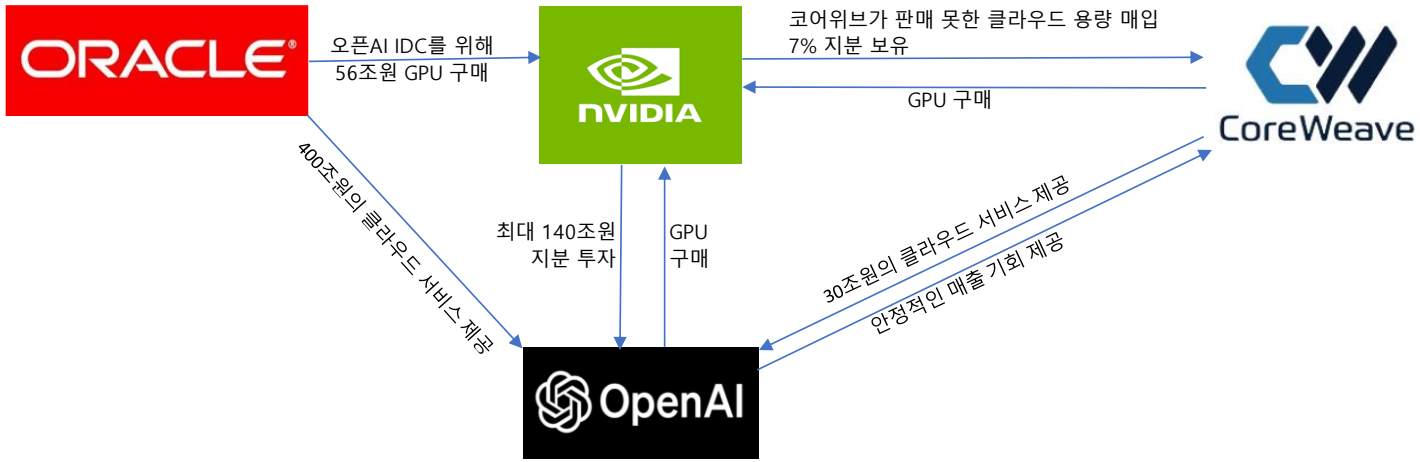


자료 : iM증권 리서치본부

OpenAI를 중심으로 한 Nvidia 주도의 순환 투자 구조

- Nvidia는 OpenAI에 대규모 투자를 하고 OpenAI는 이 투자액을 재원으로 Nvidia의 GPU를 구매하며, Oracle과 CoreWeave에게 클라우드 서비스를 제공받는 구조. 또한 Oracle과 CoreWeave는 OpenAI를 위한 데이터 센터 건설을 위해 Nvidia의 GPU를 구매
- Nvidia와 OpenAI: Nvidia는 OpenAI에 최대 1,000억 달러 (약 140조 원) 규모의 지분 투자를 단계적으로 실행할 것으로 언급. 이 투자는 OpenAI의 AI 데이터센터 구축을 위한 자금으로 사용되며, OpenAI는 이 자금을 주로 Nvidia의 GPU 구매에 사용
- Nvidia와 CoreWeave: Nvidia는 CoreWeave의 초기 투자자로서 7%의 지분을 보유 중이며, CoreWeave가 판매하지 못한 클라우드 용량을 매입하는 계약 체결. Nvidia는 CoreWeave에 GPU를 공급하고, 투자자로서 자금 조달을 지원
- CoreWeave와 OpenAI: AI 클라우드 전문 기업인 CoreWeave는 OpenAI와 총 224억 달러 (약 30조 원)에 달하는 클라우드 서비스 계약을 체결. 이 계약을 통해 OpenAI는 GPU 클라우드 자원을 확보하고, 적자 기업인 CoreWeave는 안정적인 매출을 확보
- Oracle과 Nvidia: Oracle은 OpenAI용 데이터센터 구축을 위해 Nvidia로부터 400억 달러 (약 56조 원) 규모의 GPU 구매 계약 체결
- Oracle과 OpenAI: Oracle과 OpenAI는 5년간 3,000억 달러 (약 400조원)가 넘는 규모의 클라우드 서비스 공급 계약 체결

<그림6> OpenAI를 중심으로 한 Nvidia 주도의 AI 순환 투자 구조



자료 : iM증권 리서치본부

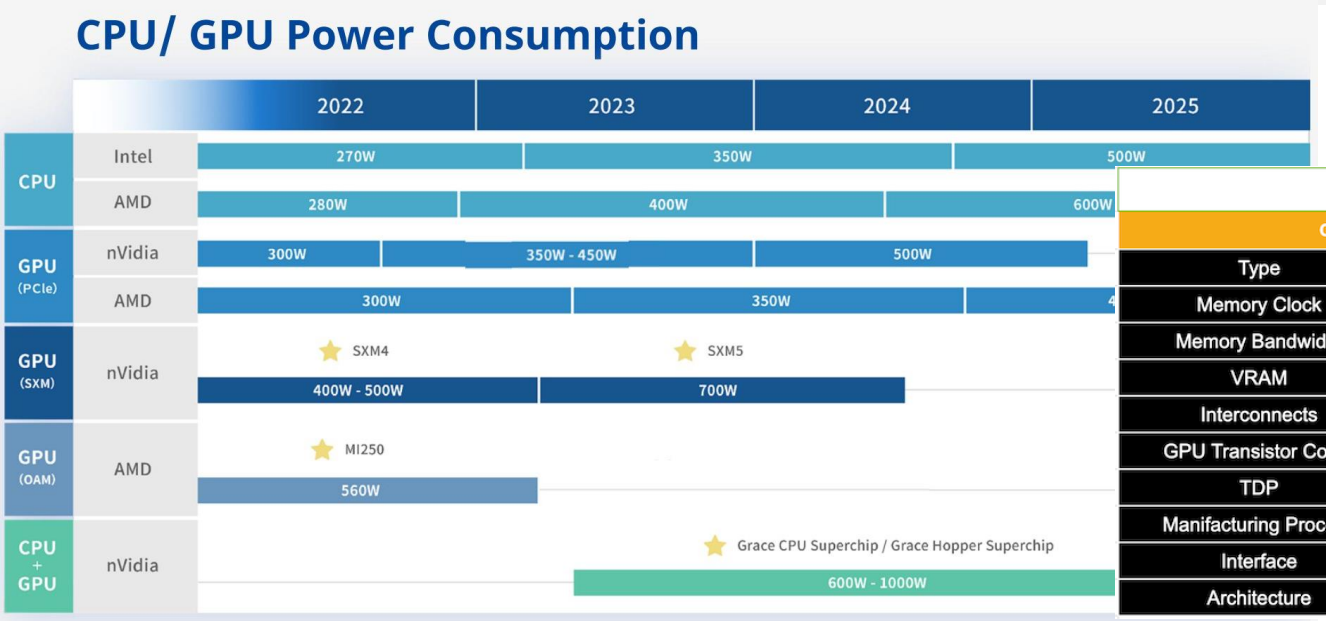
순환 투자 구조의 배경

- 2H24에 GB200을 탑재한 NVL72에서 과열, 냉각수 누수, 칩 간 연결 오류 등의 기술적 문제가 제기되었고 이에 따라 MS, Amazon, Google, Meta 등 주요 고객사들은 주문을 연기하거나, 이전 세대 칩인 Hopper로 일부 주문을 전환. CY25 5월에 GB200 서버 생산이 본격화되었으나 일부에서는 여전히 액체 냉각 시스템의 Quick Connect Fitting에서 누수 문제가 발생하는 등 지속적인 문제가 있다고 지적

- GB300에서는 각 칩에 전용 입/출력 액체 Cold Plate를 장착하고 내부 케이블링 문제를 해결하기 위해 케이블 없는 디자인을 채택하는 등 개선이 있었음. 그럼에도 불구하고, Rack당 135 ~ 140kW에 달하는 높은 전력 소모는 여전히 냉각 시스템에 큰 부담. 또한 액체 냉각 시스템은 인프라 의존성이 높아 기존 데이터센터에 도입하기 어려운 문제가 있음

- 발열 문제에 따라 Blackwell GPU에 대한 빅테크 업체들의 구매가 예상보다 저조했고 빅테크 업체들은 자체 ASIC 기반 가속기 반도체의 사용량을 늘리려고 노력 중임. Nvidia는 지속 성장을 위해 이러한 도전에 대응할 필요가 있으며 OpenAI를 중심으로 한 Nvidia 주도의 순환 투자 구조는 이러한 배경에서 나타난 것으로 판단됨

<표6> 각 Processor 별 TDP와 GB200의 Configuration



GB200	
Configuration	
Type	Grace Blackwell Superchip
Memory Clock	8Gbps HBM3E
Memory Bandwidth	2x8TB/sec
VRAM	384GB (2x2x96GB)
Interconnects	2x NVLink5 2x PCIe 6.0
GPU Transistor Count	416B (2x2x104B)
TDP	2700W
Manifacturing Process	TSMC 4NP
Interface	Superchip
Architecture	Grace + Blackwell

자료 : Gigabyte, Nvidia, iM증권 리서치본부

OpenAI의 조기 이익 발생 여부가 관건

- 이러한 순환 투자 구조가 성공하기 위해서는 Nvidia의 GPU를 자체적으로, 또한 Oracle, CoreWeave를 통해 간접적으로 구매할 OpenAI의 조기 이익 발생이 필수적임
- 향후 10년 간 1.4조 달러를 지출할 것으로 언급한 OpenAI가 현재까지 Oracle, CoreWeave, AWS 등과의 클라우드 컴퓨팅 설비 임차 계약 및 AMD, Broadcom, Nvidia와의 AI 반도체 구매 계약을 모두 합하면 향후 7년간 총 7,100억 달러의 비용이 발생
- OpenAI의 FY24 매출은 55억 달러, CY25 매출은 131억 달러를 기록했음. 단 동사는 FY24에 50억 달러, 1H25에 135억 달러의 순손실을 기록. The Information은 OpenAI가 연 매출 1,000억 달러에 도달하는 CY29 이후에야 흑자를 낼 수 있을 것으로 예상
- OpenAI의 적자가 심화되면 Nvidia의 투자금 회수가 어려워지고, 보유 지분 가치가 하락해 재무 안정성에 타격을 입을 수 있음. OpenAI의 재정난이 심화되면 GPU 구매량을 줄일 가능성이 높고, 이는 Nvidia의 매출 감소와 실적 악화로 이어질 수 있음. 또한 Nvidia는 CoreWeave가 판매하지 못한 클라우드 용량을 매입하게 되어 있으므로 CoreWeave가 어려움에 처하면 Nvidia가 막대한 비용을 떠안게 됨
- Oracle은 OpenAI와 3,000억 달러 규모의 초대형 클라우드 컴퓨팅 계약을 체결해 동사의 향후 성장은 OpenAI에 크게 의존. OpenAI가 재정적으로 어려움을 겪으면 이 계약을 이행하지 못할 가능성이 커짐. CY25 11월 현재 Oracle의 차입금 규모는 1,244억 달러, 차입금 비율은 415%에 달함

<표7> 오픈AI가 현재 계약상 향후 7년간 지불해야할 금액

Partner	예상 지불 총액 (가정/추정치 포함)	계약 내용	계약 기간
Oracle	3,000억 달러	클라우드 컴퓨팅 설비 임차	5년 (향후 2029년까지)
Broadcom	1,500억 달러 (가정)	AI 칩 구매	수년 간 (2026년 후반부터 배포)
AMD	900억 달러 (가정)	AI 칩 구매	수년 간 (4년 이상 예상)
Amazon AWS	380억 달러	클라우드 컴퓨팅 설비 임차	7년
CoreWeave	224억 달러	클라우드 컴퓨팅 설비 임차	수년 간 (현재 진행 중)
NVIDIA	1,000억 달러 (NVIDIA의 투자 금액으로 충당)	AI 칩 구매	파트너십
총합계	약 6,100억 달러 (900조원) 이상		최대 7년

자료 : iM증권 리서치본부

OpenAI의 위기: Funding에도 불구하고 여전히 IPO 성공 필요

- 향후 10년 간 1.4조 달러를 지출할 것으로 언급한 OpenAI는 최근 1,100억 달러 규모의 Funding 성공. 단 당초 목표치였던 8,400억 달러보다는 낮은 7,300억 달러의 기업 가치를 인정 받았고, 이번 Funding에서 Nvidia의 투자 규모 축소와 Amazon의 단계별 지급 조건 등 다소 순탄치 못한 모습을 보였음

- 이번 Funding에서 OpenAI가 향후 2년 내에 완전히 영리 법인으로 전환하지 못할 경우, 투자사들은 투자금 회수를 요구할 수 있는 권리를 가짐. 또한 Amazon은 자사 클라우드 사용량을 대폭 늘리는 조건을, Nvidia는 자사 칩 우선 구매 및 기술 협력을 조건으로 걸어 순환 투자의 성격이 짙음

- 이번 Funding 성공에도 불구하고 OpenAI는 1.4조 달러의 투자 약속을 지키기 위해, 당초 계획한 바와 같이 CY26 내 1조 달러 규모의 IPO 성공이 필수적임. 만약 IPO에 실패할 경우 OpenAI가 약속한 대규모 투자와 Nvidia GPU 구매가 물거품이 될 수 있음

- 동사 IPO 성공에 대한 우려 요인은 1) 막대한 적자 규모, 2) AI 버블 논란, 3) 창업자 중 한 명인 일론 머스크가 제기한 손해 배상 소송, 4) IPO로 인해 지배권이 희석되길 원치 않는 Microsoft의 반대 등임

- 동사의 IPO 성공 여부는 장기적인 관점에서 AI 투자 강도와 AI 주식에 대한 투자 심리에 큰 영향을 줄 대형 이벤트가 될 전망이며, 최근 동사 IPO 성공 여부와 관련해 일론 머스크와의 소송 Risk가 강하게 부각 중

<표8> 오픈AI의 최근 Funding에서 투자자별 투자 금액 및 세부 조건

투자자	투자액	세부 조건
Amazon	$500억	- 최대 투자자이나, 이 중 $150억만 즉시 지급. 나머지 $350억은 OpenAI가 AGI (범용 인공지능)를 달성하거나 IPO를 발표할 때 지급되는 조건부 계약 - OpenAI가 AWS의 Cloud 사용량을 대폭 늘리는 조건 포함
Nvidia	$300억	- 당초 젠슨 황 CEO는 $1,000억 투자를 시사했으나, 오픈AI의 방만한 경영과 경쟁사 (구글, 앤스로픽)의 추격을 우려해 실제 투자액을 1/3 수준으로 대폭 축소 - OpenAI가 Nvidia의 GPU를 우선 구매하고 양사 간 기술 협력을 실시하는 조건
SoftBank	$300억	- 전액 현금 투자이나 리스크 분산을 위해 CY26 4월, 7월, 10월에 걸쳐 3회 분할 집행
Total	$1,100억	- OpenAI가 향후 2년 내 완전 영리 법인으로 전환에 실패할 경우 투자금 회수 가능

자료 : iM증권 리서치본부

OpenAI의 위기: 일론 머스크와의 소송 Risk 심화

- 일론 머스크와의 소송이 OpenAI에게 치명적인 Risk로 작용할 가능성이 존재함. 일론 머스크와의 소송에서 패소할 경우 OpenAI는 존립이 어려워지고 올해 예상되는 IPO에 실패할 가능성이 있는 것으로 판단됨
- 일론 머스크는 OpenAI의 설립 초기에 560억원 (설립 당시 지분의 60%) 규모의 지분 투자를 원했으나 OpenAI는 비영리 연구소라는 이유로 동 금액을 기부금으로 받았음
- 그러나 OpenAI는 일론 머스크의 이사회 퇴임 후 3개월만에 영리 법인으로 전환했으며, 최근 창업자 중 한명인 그레그 브록먼이 기부금을 받기 이전에 이미 영리법인으로의 전환을 샘알트먼과 논의했고, 일론 머스크를 이사회에 퇴출시킬 계획이 있었음이 적혀 있는 일기가 증거로 제출되었음
- 이에 따라 미국 연방법원은 OpenAI와 MS가 요청한 머스크의 소송에 대한 기각 신청을 거부하고 CY26 4월 27일부터 배심원 재판으로 1심을 진행할 예정임
- 머스크는 최대 200조원의 손해 배상을 청구한 상황이며 OpenAI는 패소 시 이를 감당할 현금이 없는 상황이므로 머스크는 OpenAI의 지적재산권에 대한 강제 압류를 신청할 전망
- 또한 1심에서 어떤 결과가 나오든 패소한 쪽은 항소할 가능성이 높으며, 회사의 존립에 영향을 줄 막대한 규모의 소송이 진행 중인 상황에서 SEC는 OpenAI의 IPO를 승인하지 않을 가능성이 있는 것으로 판단됨. 상장 이후 OpenAI의 패소 시 일반 주주들의 돈이 머스크에게 배상금으로 지불될 수 있기 때문

CUDA 아성의 균열 조짐?

- AMD의 HIP (Heterogeneous-computing Interface for Portability)와 Google의 PyTorch에 대한 AOT (Ahead-of-Time) Autograd를 통한 지원 강화는 향후 CUDA의 아성에 균열을 일으킬 가능성이 있음. HIP는 CUDA에 대한 오픈 소스 대안으로, 개발자들이 기존 CUDA 코드를 AMD GPU에서도 쉽게 실행할 수 있도록 설계된 이식성 인터페이스. 개발자들이 최소한의 수정만으로 CUDA 기반 코드를 AMD GPU로 옮길 수 있어, Nvidia 생태계에서 벗어나는 전환 비용을 크게 낮춤

- Google은 XLA (Accelerated Linear Algebra)를 통해 PyTorch 코드가 TPU에서도 자연스럽게 실행되도록 지원을 대폭 강화. 기존에는 PyTorch가 코드를 한줄씩 실행하며 실시간으로 계산 순서를 결정했으나, AOT (Ahead-of-Time) Autograd는 모델 학습이 시작되기 전에 전체 계산 과정을 분석하여 최적의 실행 계획표를 미리 작성. XLA는 동적인 연산보다는 전체 연산 흐름이 고정된 정적 그래프를 입력 받아 컴파일하고 최적화하는 데 특화되어 있음. AOT Autograd는 PyTorch의 동적 특성을 유지하면서도 XLA가 요구하는 정적 그래프를 제공

- AOT Autograd는 PyTorch 코드와 Google TPU 사이의 호환 문제를 제거하는 핵심 기술로 이를 통해 Meta와 같은 기업들이 기존의 방대한 PyTorch 코드 베이스를 거의 수정 없이 TPU에서 고성능으로 실행할 수 있게 되었음. 이러한 노력을 통해 고객사들이 Nvidia의 CUDA에서 벗어나게 된다면 고객사들은 Nvidia의 GPU보다 훨씬 저렴한 AMD의 GPU 또는 Google의 TPU를 보다 많이 사용하게 될 전망

- 내년 주력 TPU인 Ironwood에는 HBM3E 12단 192GB, Rubin GPU에는 HBM4 12단 288GB가 탑재되므로 고객들이 투자 효율화를 위해 TPU의 사용량을 늘리고 Nvidia GPU 사용량을 줄인다면 장기 HBM 수요에 부정적인 영향을 줄 수 있음

<표9> Nvidia와 구글의 AI 생태계 비교

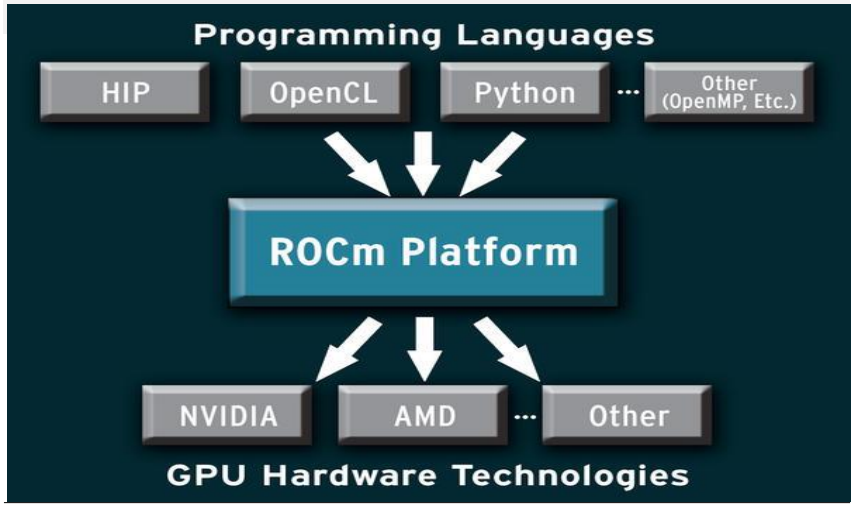
구분	Nvidia	Google
최상위 S/W	PyTorch	Jax
자동 미분 엔진	Autograd	AOT Autograd
하위 H/W Interface	CUDA	XLA
H/W	GPU	TPU

자료 : iM증권 리서치본부

AMD를 위한 환경 개선 중

- AMD AI GPU의 성능은 가격이 30 ~ 50%나 저렴한데도 발표 스펙을 기준으로 하면 Nvidia의 제품과 유사하거나 또는 상회. 다만 실제 적용 시 동사 소프트웨어 (ROCm)의 경쟁력 부족으로 CUDA를 보유한 Nvidia의 GPU 대비 떨어지는 성능과 비효율성을 보여 왔음
- 지금까지 AMD가 Rack 단위로 GPU를 공급하지 못했던 것은 Infinity Fabric, UALink 등에서 일부 비효율성이 발생하였기 때문이며, AMD의 ROCm이 CUDA 대비 호환성이 부족함에 따라 이론상 최대 성능을 구현하지 못한 데 원인이 있었음
- 그럼에도 불구하고 OpenAI, Meta, Oracle 등이 최근 AMD와 대규모 계약을 체결한 것은 단일 Supplier Risk를 회피하고자 하는 점 뿐 아니라 점차 AMD와 Nvidia간 격차가 좁혀질 수 있는 환경이 조성되고 있기 때문
- AI 코딩 기술의 발전은 GPU 구매자들이 보다 간단히 오픈 소스인 AMD의 ROCm을 직접 수정할 수 있도록 변화시키고 있음. 또한 AMD는 Nvidia와 자사 GPU 모두에서 실행할 수 있는 애플리케이션을 만들 수 있도록 HIP라는 프로그래밍 인터페이스를 개발
- AMD는 지난 6월 72개의 GPU가 탑재된 Rack 단위 공급을 시작하겠다고 밝혔으며 OpenAI는 MI450 시리즈 GPU 및 Rack 스케일 솔루션 설계 요건에도 핵심적인 기여를 한 것으로 추정

<그림7> AMD의 ROCm Platform



자료 : AMD, iM증권 리서치본부

중국의 자체 AI 칩 사용 증가

- 미국의 대중국 반도체 수출 규제가 심화됨에 따라 CY25에 Baidu, Alibaba, Tencent 등 중국 빅테크 기업들이 Nvidia GPU의 대안으로 Huawei의 Ascend 시리즈 등을 대거 채택. 이에 따라 CY25 중국 AI 서버칩 시장에서 중국 업체들이 40%의 시장 점유율 획득

- 최근 미국 정부가 전면 금지에서 조건부 허용으로 전환했으나 여전히 제한 조건 존재. CY26에는 Nvidia의 H200 및 AMD의 MI325X급 칩의 출하가 허용되나 최첨단 제품인 Blackwell과 Rubin은 여전히 수출 금지 대상. 그리고 중국에 판매되는 칩 가격의 25%를 미국 정부에 지불해야 하며 중국 수출량은 해당 제품의 미국 내 전체 판매량의 50%를 넘을 수 없음. 또한 국가 안보에 위협이 안되는 승인된 중국 고객사에만 수출 허가

- 미국의 대 중국 수출에 대한 제한은 언제든지 강화될 수 있으며 이에 대응하기 위해 중국의 자체 AI 가속기 반도체 개발은 더욱 가속화되고 있음. 이는 장기적으로 Nvidia의 성장에 부정적인 요인이 될 수 있음

- 다만 이러한 미국의 중국향 수출 제한은 한국 Memory 반도체에 대형 호재 역할을 하고 있음. Ascend 910C는 H100 성능의 약 60 ~ 70% 수준으로 부족한 개별 칩의 성능을 더 많은 수의 Huawei GPU를 연결하여 보완. Nvidia의 GPU 1개가 처리할 연산을 Huawei GPU 1.5 ~ 2개가 처리

- 이를 지원하기 위한 시스템 메인 메모리 (DDR5)와 HBM 수요가 중국에서 급증. 이는 4Q25 이후 공급 부족의 트리거 역할을 한 것으로 판단됨

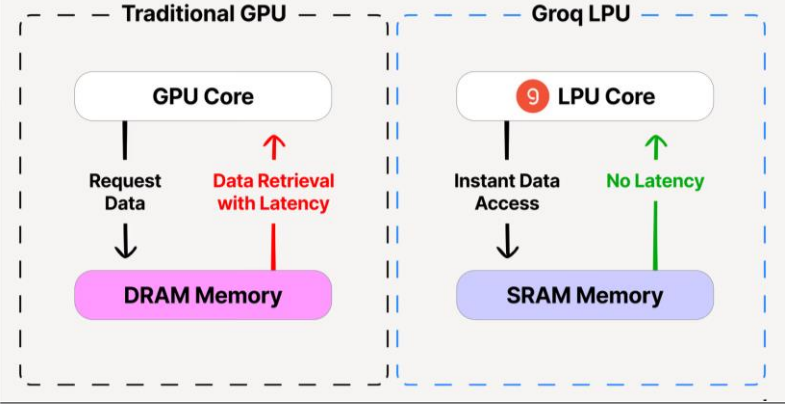
<그림8> CY25 중국 AI 서버칩 시장 점유율

자료: IDC

Groq 인수와 SRAM 활용

- 추론 시대를 맞이하여 AMD, Google 등의 도전에 대한 Nvidia의 대응은 지난 12월 24일에 있었던 200억 달러 규모의 Groq 인수에서 잘 나타남
- Nvidia가 Groq를 인수한 이유는, 1. LPU 기술 확보: Groq의 LPU (언어 처리 장치)는 기존 GPU보다 압도적으로 빠른 초저지연 (Low-latency) 추론 성능 보유. 2. 미래 경쟁자 제거: Groq은 구글 TPU를 만든 핵심 인재들이 설립한 회사로, 엔비디아 GPU의 비싼 가격과 전력 소모 문제를 해결할 가장 강력한 대항마로 꼽혀 왔음. 이번 딜로 잠재적 라이벌을 우군으로 흡수. 3. 추론 시장 주도권: Nvidia는 Groq의 기술을 통해 저비용, 고효율 추론 솔루션을 강화함으로써 빅테크들의 탈 Nvidia 움직임을 차단
- LPU의 차별점은 1. HBM 병목 현상 제거: GPU는 HBM에서 데이터를 읽어오는 과정에서 지연 시간 발생. 반면 LPU는 데이터 이동 없이 칩 안에서 즉각 연산이 가능한 On-Chip-SRAM 구조를 사용하여 초저지연 실현. 2. 결정론적 스케줄링: 데이터 처리의 흐름을 미리 예측 가능하게 설계하여, 여러 칩을 연결해도 효율 저하 없이 하나의 거대한 코어처럼 작동. 3. 실시간 처리 능력: 기존 GPU 대비 약 10배 이상 빠른 응답 속도와 90% 이상 높은 에너지 효율을 제공하여, 실시간 대화형 AI나 자율주행 등 즉각적인 반응이 필요한 서비스에 최적화
- Nvidia는 고사양 모델 학습에는 기존의 GPU를 사용하고, 실시간 추론 단계에는 LPU의 저지연 설계를 도입하는 하이브리드 솔루션을 제공할 계획. CY26 출시될 Rubin 및 그 후속 모델에 LPU 설계 아이디어를 통합하여, Nvidia 생태계 내에서 최고의 추론 성능을 낼 수 있도록 할 예정. AI 에이전트와 로보틱스 시장에서 요구하는 즉각적인 의사결정 인프라를 선점하기 위해 LPU 기술을 핵심 엔진으로 활용할 계획
- Google이나 AMD로 눈을 돌리려던 고객들에게 Nvidia가 학습뿐 아니라 추론에서도 가장 빠르고 저렴하다는 인식을 각인시켜 이탈을 방지

<그림9> LPU의 Low Latency



자료 : eesel AI

Groq 인수와 SRAM 활용

- LPU의 핵심은 Ultra Low Latency임. 기존 GPU는 연산 장치와 HBM이 분리되어 있음. GPU가 계산을 하려면 외부 HBM에서 데이터를 가져와야 하는데, 이 이동 속도가 계산 속도를 따라가지 못하므로 이 과정에서 지연 시간 (Latency) 발생

- 반면 Groq의 LPU는 모든 모델 데이터를 LPU 내부의 초고속 메모리인 SRAM에 직접 올림. 데이터 이동 없이 LPU 내에서 바로 연산이 이루어지므로 초 저지연이 가능. 또한 데이터를 먼 거리 (HBM)까지 이동시킬 필요가 없어 전력 소모가 대폭 감소. Nvidia는 이 기술을 확보함으로써 고용량 데이터는 HBM이 담당하고 실시간 고속 연산은 SRAM 기반의 LPU 기술이 담당하는 하이브리드 메모리 아키텍처 구축 가능

- SRAM은 기존 GPU들에 이미 탑재되어 있었음. H100, B200의 내부에는 L1/L2 캐시라고 불리는 SRAM이 이미 존재. 다만, 그 용량이 80 ~ 144MB 수준으로 매우 작아 모델 전체를 담지 못하고 데이터가 잠시 거쳐 가는 용도. Rubin 설계 단계에서도 이미 SRAM의 용량을 기존보다 대폭 키우고, 연산 유닛과의 대역폭을 극대화하는 설계를 반영. 이는 HBM의 속도 한계를 극복하려는 반도체 업계의 공통된 흐름임

- 기존의 Nvidia는 SRAM을 보조적인 Cache Memory로 썼으며 데이터가 SRAM에 있는지 없는지 체크하는 과정에서 시간 필요. 반면 LPU는 SRAM을 Main Memory처럼 사용하므로 체크 과정 없이 데이터의 위치를 미리 알고 연산. Nvidia는 Rubin 설계 당시에 이미 SRAM 비중을 높여 놨는데, 이 SRAM을 가장 잘 다루는 소프트웨어 알고리즘이 Groq에 있었던 것

- 2H26 출시 예정인 Rubin의 하드웨어 구조는 이미 정해져 있으나, Groq의 핵심 인재들이 Rubin의 소프트웨어를 최적화하여, Rubin 내부에 이미 커진 SRAM을 LPU처럼 효율적으로 제어하도록 만드는 작업을 수행할 전망. Groq의 SRAM 내 데이터 흐름을 완벽히 제어하는 하드웨어 설계 IP가 직접적으로 칩 설계 도면에 적용되는 시점은 Rubin Ultra가 될 듯

- Rubin에서 SRAM 용량이 600MB 이상으로 대폭 증가하더라도, SRAM은 모델 전체를 담는 용도가 아니라 HBM에서 가져온 데이터를 한 번에 더 많이 펼쳐 놓고 작업하는 넓은 책상의 역할임. SRAM이 커질수록 HBM에서 데이터를 다시 불러오는 횟수가 줄어들어 추론 속도는 빠르게 상승하지만, 모델 전체를 저장해야 하는 HBM의 절대 용량은 모델의 크기가 커짐에 따라 여전히 완만하게 증가할 전망임

- 가속기 반도체에 사용되는 SRAM은 Memory 반도체 업체가 만드는 단품 칩이 아니라, GPU 연산부 내부에 함께 설계되어 찍혀 나오는 Embedded Memory로 TSMC가 SRAM 생산의 주체임. 공정이 미세화 될수록 연산 코어는 작아지나, SRAM은 크기가 거의 줄어들지 않음에 따라 TSMC는 3D 패키징 기술을 사용하여 GPU 다이 위에 SRAM 전용 다이를 수직으로 쌓아 올리는 방식으로 제조

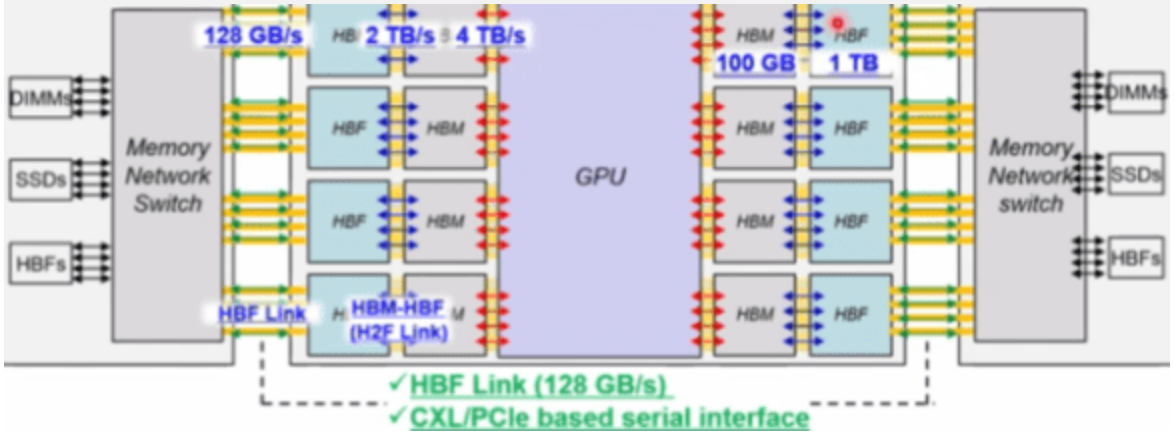
NAND / HBM / SRAM을 모두 활용한 AI Memory 전략

- Nvidia는 Rubin 플랫폼에 BlueField-4 DPU를 활용한 새로운 스토리지 레이어인 ICMS (Inference Context Memory Storage)를 도입할 것으로 CES에서 발표. ICMS는 LLM이 추론 과정에서 생성하는 방대한 KV Cache 데이터를 저장하기 위한 공간. 기존에는 이 데이터를 비싼 HBM에 두었으나, 모델의 문맥 (Context)이 길어지면서 HBM 용량만으로는 한계에 부딪혔고, 이를 보완하기 위해 대용량 NAND 플래시 기반의 eSSD를 활용
- Rubin GPU 1개당 약 16TB의 NAND가 할당되는 구조. CY26 전 세계 NAND 수요의 2.7%를 차지할 듯
- Rubin 아키텍처 이후로 Nvidia는 SRAM, HBM, NAND 세 가지를 하나로 묶는 계층형 AI Memory 전략을 완성하고자 하는 듯. SRAM은 책상, HBM은 책꽂이, NAND는 대형 창고의 역할을 수행
- SRAM: 연산에 즉각 필요한 현재의 단어 (토큰)를 처리. CY25 말 Groq 인수를 통해 확보한 기술을 바탕으로, On-Chip-SRAM의 활용 효율을 극대화해 칩 내부 SRAM에서 즉시 처리 함으로써, LLM이 답변을 시작하는 속도를 단축
- HBM: AI 모델의 매개변수와 현재 진행 중인 대화의 핵심 데이터를 보관. SRAM에 담기엔 크고 NAND에 두기엔 빠른 처리가 필요한 데이터들이 머무는 중간 정류장
- NAND: 방대한 KV Cache (과거 대화 내역 전체)를 저장. 사용자가 며칠 전 대화했던 내용이나 수만 페이지 분량의 PDF 문서를 AI가 읽어야 할 때, 이 방대한 데이터를 저렴하고 용량이 큰 NAND에 보관. 필요할 때 BlueField-4 DPU가 NAND에서 필요한 문맥만 선별해 HBM/SRAM으로 이동
- 만약 AI에게 "지난달에 분석했던 보고서 내용이랑 오늘 뉴스를 비교해줘"라고 시킨다면, NAND는 보관 중이던 수천 페이지의 지난달 보고서 데이터를 꺼내고, HBM은 오늘 뉴스의 내용과 모델의 지식을 로딩하며, SRAM은 GPU가 두 데이터를 빠르게 비교 연산하도록 돕는 구조
- 이러한 방향으로 진행될 경우 Rubin Ultra 이후 AI GPU의 HBM (기존에 KV Cache를 저장했던) 탑재량은 기존에 알려졌던 것 대비 축소되고, 반대로 NAND의 수요량이 증가할 것으로 판단됨. 또한 Rubin 아키텍처가 지향하는 SRAM-HBM-NAND 중심의 계층 구조는 기존 AI 서버에서 범용 DDR5 DRAM이 수행하던 역할을 상당 부분 대체하거나 축소시키는 방향으로 작용할 가능성이 큰 듯
- 과거에는 GPU의 HBM 용량이 부족해 남는 데이터를 CPU와 연결된 DDR5로 보내 저장. 그러나 Rubin은 이 역할을 NAND (ICMS)가 직접 가져가도록 설계. 이 과정에서 DDR5를 거쳐야 했던 경로가 생략되면서, AI 연산만을 위한 DDR5의 중요도와 수요는 상대적으로 낮아질 듯
- Nvidia가 Groq의 기술을 흡수하고 NAND를 GPU 근처에 붙이는 행보는, 결국 느린 CPU-DDR5 경로를 거치지 않고 직접 모든 것을 처리하겠다는 의도인 듯. 이는 Memory 반도체 업체들에게 DDR5보다는 HBM, CXL (공유 Memory), eSSD에 더 집중해야 한다는 의미로 판단됨

NAND 수요 급성장 예상

- Nvidia의 현재 ICMS 전략은 일반 eSSD를 그대로 사용한다는 것임. 이는 HBF로 간다 해도 NAND 자체의 느린 Latency를 극복하기 어렵고, 고대역폭은 eSSD로도 가능하다고 보는 견해임. 반면 Google은 HBF를 적극적으로 추진하고 있는 상황
- HBF (High Bandwidth Flash): NAND를 HBM처럼 수직으로 쌓아 대역폭을 넓힌 초대용량 고속 메모리. HBF는 HBM 대비 속도는 약 80 ~ 90% 수준이나, 용량은 8 ~ 16배 크고 소비 전력은 40% 가량 낮아 AI 추론 서버의 중간 계층 메모리에 최적화
- HBF의 기술적 난관: 초고층 적층으로 인한 고열, 하단의 고성능 로직 칩 설계, 수만 개의 구멍을 뚫는 TSV 공정, 신뢰성 및 수명 확보, 메탈 배선 공정에서 몰리브덴으로의 전환
- 각 업체들의 HBF 대응: 삼성전자는 파운드리 강점을 살린 로직 칩 통합에, SK하이닉스는 SanDisk와의 협력을 통한 글로벌 표준화에 노력 중임. 현재 삼성전자보다는 SK하이닉스가 HBF에 좀더 적극적인 것으로 보임. 2H26에 삼성전자와 SK하이닉스가 시제품 공개 및 고객사 검증 개시 예정
- Nvidia의 견해대로 ICMS를 위해 eSSD를 사용하건, Google의 방향대로 HBF를 적극 도입하건 향후 NAND 수요 성장률이 매우 빠를 것이라는 것은 변함이 없는 듯

<그림10> AI 서버 내 HBM – HBF – Storage 구성도

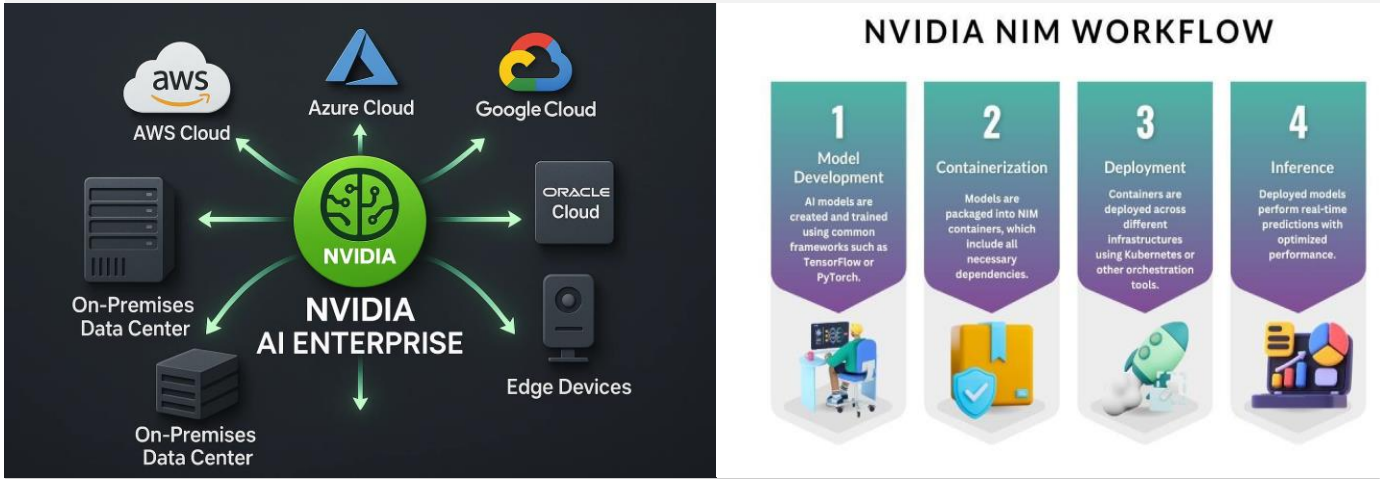


자료 : EBN

소프트웨어와 서비스 부문 강화

- Nvidia는 고객사가 동사 칩을 구매한 뒤 실제 서비스를 구현하는 과정에서 겪는 기술적 난관을 해결해 주며 매출을 확대하는 전략을 최근 강화 중
- NVIDIA AI Enterprise (기업용 AI 운영체제): 기업이 AI를 안전하고 안정적으로 운영할 수 있게 돕는 유료 구독형 플랫폼. 기업이 오픈소스 AI 모델을 가져와 실제 서비스에 적용하려면 보안, 호환성, 업데이트 등 관리할 요소가 매우 많음. Nvidia는 이를 검증하고 최적화하여 패키지 형태로 제공. GPU당 연간 약 4,500달러 수준의 구독료 발생. 문제 발생 시 Nvidia가 직접 기술 지원을 해준다는 점이 고객들에게 가장 큰 매력 요소
- NIM (NVIDIA Inference Microservices): 최신 AI 모델을 서버에 올리려면 설정에만 수주가 걸리나 NIM은 이를 단 몇 분 만에 실행할 수 있게 모든 설정이 완료된 상태로 제공됨. Nvidia의 최신 칩 성능을 100% 끌어낼 수 있도록 내부 코드가 짜여 있으나 타사 칩이나 일반 서버에서는 이만큼의 속도가 나오지 않게 설계되어 있음. 기업 AI 개발자들이 NIM의 편리함에 익숙해지면, 인프라를 AMD나 자체 칩으로 바꾸려 할 때 소프트웨어를 통째로 다시 짜야 하는 부담을 가지며 결국 소프트웨어가 편해서 계속 Nvidia 칩을 계속 사용하게 만드는 전략
- AI Enterprise는 기업의 관리 부담을 덜어주며 매년 안정적인 매출을 발생시키고 NIM은 개발자의 배포 편의성을 극대화하여 Nvidia 생태계 밖으로 나가지 못하게 묶어 두는 전략. 이 두 서비스를 결합해 동사는 AI 인프라 전체를 지배하는 플랫폼 사업자의 위치를 유지하려고 노력 중

<그림11> Nvidia AI Enterprise의 고객들과 NIM의 Workflow



자료 : Uvation, Deepchecks