반도체

비중확대(유지)

종목	투자의견	목표주가
삼성전자	매수	95,000원
SK하이닉스	매수	410,000원

Rubin CPX 발표의 의미

엔비디아, Rubin CPX GPU 공개

엔비디아는 AI Infra Summit 2025에서 대규모 컨텍스트 추론을 위해 설계된 새로운 GPU인 Rubin CPX를 공개했다. Rubin CPX는 추론의 워크로드 두 단계 중 Prefill 단계를 더 효율적으로 처리하기 위한 전용칩이다. 이는 고부가가치의 긴 컨텍스트 추론 워크로드를 위한 높은 처리량 성능을 제공하는 동시에, 분산 인프라에 원활히 통합되도록 설계된 특별한 솔루션이다. Rubin CPX는 CUDA, NVLink, Rack scales Server 등과 더불어 엔비디아의 기술적 해자를 더욱 넓고 깊게 만들어 줄 제품이 될 것이다.

엔비디아는 추론 시장까지 지배할 것

SemiAnalysis의 분석에 따르면, Rubin CPX는 연산량은 Rubin GPU의 약 60%, 가격은 약 25%로 추정된다. 즉, Rubin GPU와 동일한 연산량을 더 효율적으로 처리하면서도 약 40% 수준의 낮은 가격으로 이용할 수 있는 셈이다. 데이터센터 스케일로 확장할 경우 TCO 절감 및 리소스 효율화 수준은 칩 단위에서 계산한 것보다 훨씬 크게 나타날 것이다. 또한 Rubin CPX는 긴 컨텍스트 처리를 목적으로 개발한 전용칩이기 때문에, 향후 입력과 출력의 인터페이스가 멀티 모달로 진화함에 따라 응용처가 더욱 확장될 전망이다. Rubin CPX는 하나의 VR 플랫폼 기반 서버 Rack에 통합된 VR NVL144 CPX로, 또 기존 VR NVL144 서버를 활용하고자 하는 고객을 위해 Rubin CPX 단독으로도 제공될 예정이다.

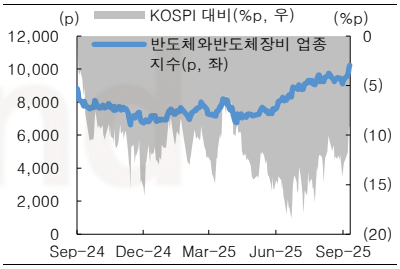
메모리는 HBM에 "추가로" GDDR7까지 사용

엔비디아 슈퍼칩은 CPU에 LPDDR5x를, GPU에 HBM을 탑재해 왔으며, Rubin CPX는 여기에 추가로 GDDR7 메모리를 도입할 예정이다. LPDDR은 원래 모바일 기기에 최적화된 저전력 메모리였고, GDDR은 그래픽 카드에 주로 사용된 메모리였다. 그러나 이런 범용 메모리가 AI 서버 등 고부가가치 응용처에 채택되기 시작하면서, 메모리 업체에 새로운 성장 기회를 제공하고 있다. 향후 AI 서버 시장에서는 HBM 중심의 프리미엄 메모리 수요가 확대됨과 동시에 GDDR, LPDDR의 병행 채택이 늘어나면서 AI 메모리 수요의 수혜가 HBM과 같은 특정 제품군에 집중되지 않고 메모리 전반으로 폭넓게 확산될 것이다. 특히 26년은 메모리 업체들의 공급 능력 확장이 제한된 상황에서 AI 서버 관련 메모리 수요가 급증함에 따라, 전반적인 메모리 쇼티지가 심화될 것으로 추정한다.

주요 종목 valuation

		2024A	2025F	2026F
SK하이닉스 (000660)	PER(x)	6.1	5.7	4.3
	PBR(x)	1.7	2.0	1.4
	EV/EBITDA(x)	3.8	4.0	2.7
	DY(%)	1.3%	0.5%	0.5%
삼성전자 (005930)	PER(x)	10.7	16.5	11.8
	PBR(x)	0.9	1.1	1.0
	EV/EBITDA(x)	3.6	5.1	3.6
	DY(%)	2.7%	2.0%	2.0%

12개월 업종 수익률



자료: FnGuide - WICS

채민숙
minsook@koreainvestment.com

황준태
jt.hwang@koreainvestment.com

한국투자증권은 당사 고객 및 별도의 서비스 계약을 맺은 법인에게만 리서치 리포트를 공개하고 있습니다. 한국투자증권의 사전 승인 없이 리포트를 어떤 형태로든 복제, 배포, 전송, 변형 및 판매하는 행위는 저작권법 위반으로 법적 처벌의 대상이 될 수 있음을 알려드립니다.

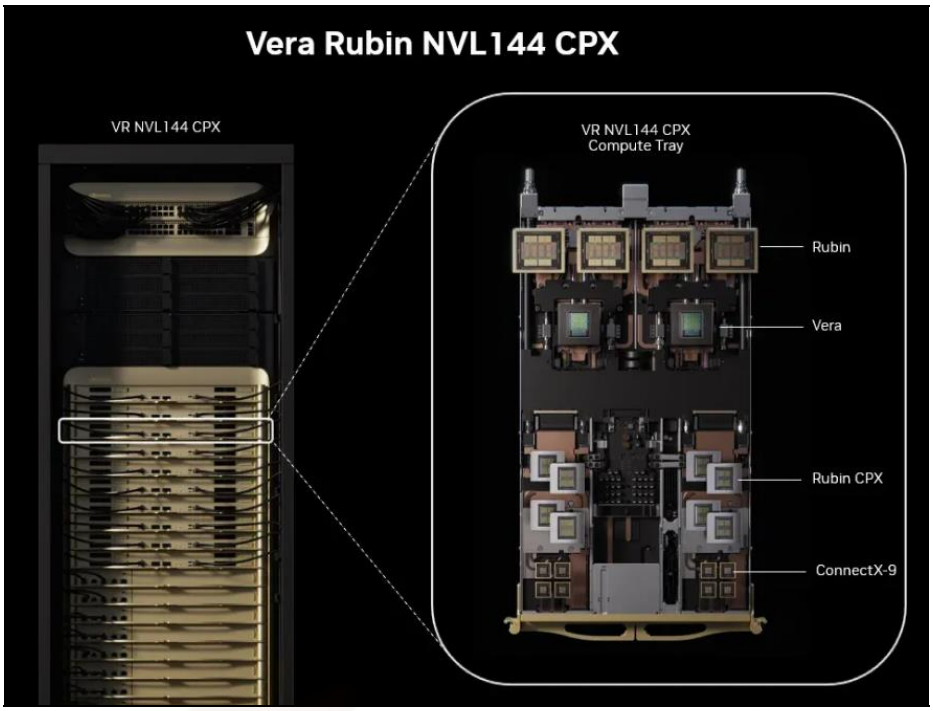
Rubin CPX 개발 배경

LLM의 추론(Inference) 과정은 크게 두 단계로 구분할 수 있다. 첫 번째는 Prefill 단계(Context phase)로, 입력(prompt) 전체를 읽어 KV캐시(Key, Value Cache)를 생성하고 첫 번째 토큰을 출력하는 과정이다. 두 번째는 Decode 단계(Generation phase)로, 이 단계에서는 저장된 KV캐시를 활용해 이후 토큰들을 하나씩 순차적으로 생성한다. Prefill 단계에서는 입력 문장을 처음부터 끝까지 처리하면서 각 단어(토큰)에 대한 특징 벡터를 계산하고, 이를 Key와 Value 형태로 저장한다. 이렇게 생성된 KV캐시는 이후 새로운 토큰을 생성할 때마다 불러와 재활용된다.

Prefill 단계는 대량의 입력 데이터를 수집하고 분석하여 첫 번째 토큰 출력 결과를 생성해야 하므로 높은 연산 처리량이 필요하다. 이 연산의 패턴은 엄청난 수의 행렬의 곱의 형태이기 때문에, Prefill 단계에서는 연산량(FLOPS)이 병목으로 작용한다. 반면 Decode 단계는 Prefill에서 출력한 첫 번째 토큰을 가지고 이어서 차례로 다음 토큰을 생성해야 한다. 매 토큰을 생성할 때마다 모델은 지금까지 생성된 모든 토큰의 KV캐시를 참조해야 한다. 즉, 새로운 연산보다는 기존에 저장된 KV캐시를 반복적으로 불러오는 과정이 지배적이므로, Decode 단계에서는 메모리 대역폭이 병목으로 작용한다. 이 두 가지 추론의 과정을 분리해, Prefill 단계는 전용칩인 Rubin CPX으로 처리함으로써 컴퓨팅 및 메모리 리소스를 목표에 맞춰 최적화할 수 있다. 이러한 아키텍처의 변화는 처리량을 향상시키고, 지연 시간(latency)을 줄이며, 전반적인 리소스 활용도를 향상시킬 수 있다. Decode는 기존대로 Rubin GPU를 통해 처리하고 있으나, 추후 Rubin CPX와 같이 Decode 전용칩을 출시할 가능성이 있다. Rubin CPX와 반대로 Decode 전용칩은 연산 능력을 낮추는 대신 메모리 밴드위스 확보를 위해 추가적으로 HBM 채용량을 증가시키는 형태로 개발될 것이다.

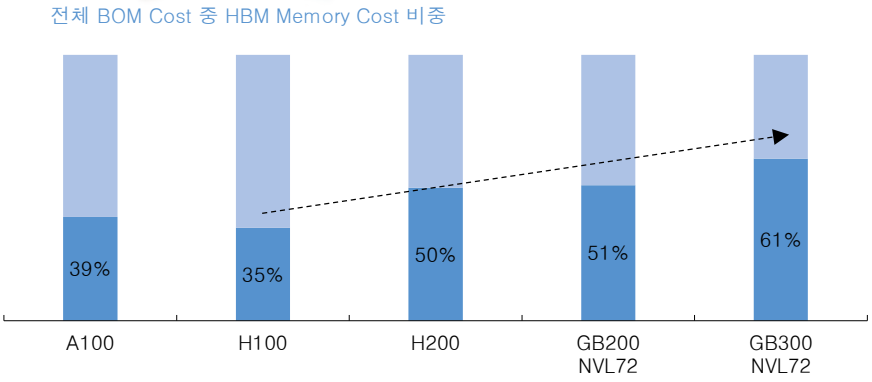
LLM의 규모는 계속 커지고, LLM이 텍스트 기반에서 음성, 영상 등 멀티 모달 입력과 출력으로 확장되면서, 처리해야 할 데이터의 양이 기하급수적으로 증가하고 있다. 이에 따라 학습과 추론 모두에서 메모리 수요가 급격히 증가하고 있으며, 이는 HBM 채용량 확대를 이끌고 있다. 그러나 Rubin GPU 기준으로 HBM은 이미 BoM cost의 약 60%를 차지하고 있어, 이후 추가적으로 필요한 메모리를 HBM으로 확장하는 것은 고객사 입장에서 비용적으로 큰 부담이다. 게다가 HBM은 칩 적층 공정의 복잡성과 인터포저 기반 패키징 때문에, LLM이 요구하는 수준으로 무한정 메모리 용량을 늘리기도 어렵다. 이에 엔비디아는 추론 워크로드를 Prefill과 Decode로 구분하고, 상대적으로 대역폭 의존도가 낮은 Prefill 단계에는 GDDR7을, 메모리 대역폭이 핵심인 Decode 단계에서는 계속해서 HBM을 사용함으로써 BoM cost 절감과 메모리 확장성이라는 두 가지 과제를 동시에 해결하고자 하는 것이다. 이러한 변화는 엔비디아와 메모리 업체 모두에게 기회 요인이 될 수 있다. 엔비디아 입장에서는 HBM에만 의존하지 않고 GDDR7까지 활용함으로써 TCO를 낮출 수 있다. 동시에 메모리 업체는 기존 HBM 수요는 확대되는 가운데 GDDR7과 LPDDR5x 등 범용 메모리 채택이 병행됨에 따라 메모리 전반으로 수혜가 확산될 수 있을 것이다.

[그림 1] Vera Rubin NVL144 CPX 구조



자료: 엔비디아, 한국투자증권

[그림 2] 서버 원가에서 HBM이 차지하는 비중이 계속 증가



자료: Semianalysis, 한국투자증권

<표 1> 엔비디아 GPU 로드맵별 상세 아키텍처 및 시스템 구성

	2022	2023	2024	2025	2026	2026	2027
	Hopper		Blackwell		Rubin		
가속기	H100	H200	B200/GB200	GB300	VR200	CPX	VR300
GPU TDP (W)	700	700	700/1,200	1,400	2,300	800	4,000+
Foundry Node	4N	4N	4NP	4NP	N3P	N3P	N3P
Logic Die Configuration	1x Reticle Sized GPU		2x Reticle Sized GPU		2x Reticle Sized GPU, 2x I/O chiplet	1x Reticle Sized GPU	4x Reticle Sized GPU, 4x I/O chiplet
FP4 PFLOPs – Dense (per Package)	4	4	10	15	33.3	20	66.7
Memory	80GB HBM3	141GB HBM3E		288GB HBM3E	288GB HBM4	128GB GDDR7	1024GB HBM4E
HBM Stacks	5	6	8	8	8	N/A	16
Memory Bandwidth	3.35TB/s	4.8TB/s	8TB/s	8TB/s	20.5TB/s	2TB/s	53TB/s
Packaging	CoWoS-S	CoWoS-S	CoWoS-L	CoWoS-L	CoWoS-L	FC-BGA	CoWoS-L
SerDes speed (Gb/s uni-di)	112G	112G	224G	224G	224G	64G(PCIe Gen6)	224G
Nvidia CPU	Grace	Grace	Grace	Grace	Vera	Vera	Vera
Maximum system density	NVL8	NVL8	NVL72 144 compute chiplets 72 Logical GPUs		NVL144 144 compute chiplets 72 Logical GPUs	144 CPX Chips	NVL576 576 compute chiplets 144 Logical GPUs
Form Factor Supported	HGX	HGX	HGX, Oberon	HGX, Oberon	HGX, Oberon	VR CPX	Kyber
# of GPU Packages	8	8	72	72	72	144	144
# of GPU dies	8	8	144	144	144	144	144
Scale up links	UBB (PCB)	UBB (PCB)	Copper Backplane	Copper Backplane	Copper Backplane	None	PCB Backplane
Aggregate FP4 PFLOPs (Dense)	32	32	720	1,080	2,398	2,877	9,605
Aggregate Memory capacity	14TB	14TB	14TB	21TB	21TB	18TB	147TB
Aggregate Memory bandwidth	27TB/s	38TB/s	576TB/s	576TB/s	1,476TB/s	288TB/s	7,668TB/s

자료: Semianalysis, 한국투자증권

<표 2> 엔비디아 VR Rack SKU별 Compute 및 Networking 칩 구성 비교

	VR200 NVL144	VR200 NVL 144 CPX	Vera Rubin CPX Only
Content Per Tray			
Rubin R200 GPUs	4	4	–
Vera CPUs	2	2	2
Rubin CPX GPUs	–	8	8
CX-9s	8	8	8
Tray 내 칩 개수	14	22	18
Content Per Rack			
Rubin R200 GPUs	72	72	–
Vera CPUs	36	36	36
Rubin CPX GPUs	–	144	144
CX-9s	144	144	144
Rack 내 칩 개수	252	396	324

자료: Semianlysis, 한국투자증권

<표 3> 커버리지 valuation

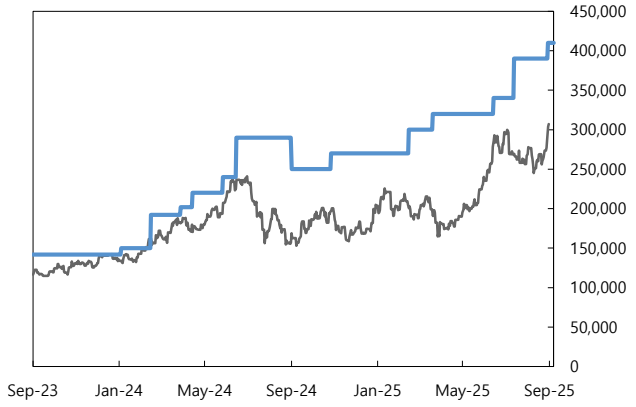
투자의견 및 목표주가			실적 및 Valuation									
종목				매출액 (십억원)	영업이익 (십억원)	순이익 (십억원)	EPS (원)	BPS (원)	PER (배)	PBR (배)	ROE (%)	EV/EBITDA (배)	DY (%)
삼성전자	투자의견	매수	2023A	258,935	6,567	14,473	2,131	52,002	36.8	1.51	4.1	10.0	1.8
(005930)	목표주가(원)	95,000	2024A	300,871	32,726	33,621	4,950	57,930	10.7	0.92	9.0	3.6	2.7
	현재가 (9/11,원)	73,400	2025F	310,970	29,755	29,958	4,444	63,854	16.5	1.15	7.3	5.1	2.0
	시가총액(십억원)	434,501	2026F	339,902	53,748	41,738	6,217	71,036	11.8	1.03	9.2	3.6	2.0
			2027F	356,897	49,772	38,935	5,801	77,803	12.7	0.94	7.8	3.4	2.0
SK하이닉스	투자의견	매수	2023A	32,766	(7,730)	(9,112)	(13,244)	76,616	NM	1.85	(15.6)	21.1	0.8
(000660)	목표주가(원)	410,000	2024A	66,193	23,467	19,789	28,732	104,567	6.1	1.66	31.1	3.8	1.3
	현재가 (9/11,원)	307,000	2025F	89,553	39,540	37,155	53,829	156,193	5.7	1.97	40.1	4.0	0.5
	시가총액(십억원)	223,496	2026F	110,987	54,848	48,956	70,904	224,031	4.3	1.37	35.9	2.7	0.5
			2027F	105,438	39,622	36,360	52,661	274,566	5.8	1.12	20.3	2.9	0.5

자료: 각 사, 한국투자증권

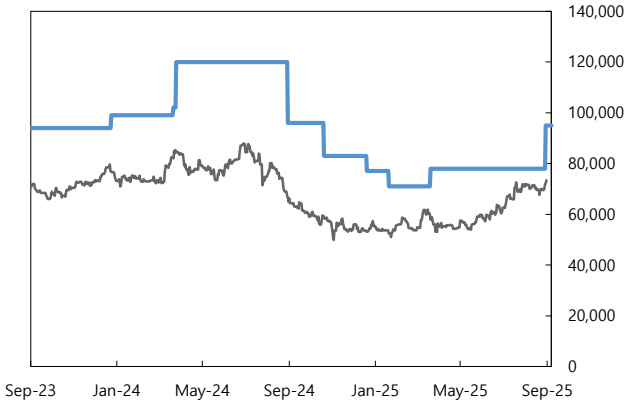
투자의견 및 목표주가 변경내역

종목(코드번호)	제시일자	투자의견	목표주가	괴리율		종목(코드번호)	제시일자	투자의견	목표주가	괴리율	
				평균주가 대비	최고(최저) 주가 대비					평균주가 대비	최고(최저) 주가 대비
SK하이닉스 (000660)	2023.07.04	매수	142,000원	-12.5	0.3		2025.07.24	매수	390,000원	-32.2	-26.2
	2024.01.15	매수	150,000원	-5.5	7.6		2025.09.10	매수	410,000원	-	-
	2024.02.26	매수	192,000원	-11.6	-2.1	삼성전자 (005930)	2023.07.04	매수	94,000원	-25.1	-15.3
	2024.04.08	매수	202,000원	-10.8	-6.7		2024.01.04	매수	99,000원	-24.8	-16.8
	2024.04.25	매수	220,000원	-14.5	-8.0		2024.04.01	매수	102,000원	-16.9	-16.4
	2024.06.07	매수	240,000원	-6.6	-1.0		2024.04.05	매수	120,000원	-34.2	-26.8
	2024.06.26	매수	290,000원	-32.5	-16.9		2024.09.10	매수	96,000원	-36.6	-30.9
	2024.09.12	매수	250,000원	-26.8	-19.6		2024.10.31	매수	83,000원	-33.8	-29.3
	2024.11.07	매수	270,000원	-30.2	-16.5		2024.12.31	매수	77,000원	-29.3	-25.6
	2025.02.25	매수	300,000원	-33.1	-28.2		2025.01.31	매수	71,000원	-20.5	-13.0
	2025.03.31	매수	320,000원	-36.7	-13.0		2025.03.31	매수	78,000원	-20.7	-6.9
	2025.06.25	매수	340,000원	-16.9	-11.8		2025.09.10	매수	95,000원	-	-

SK하이닉스(000660)



삼성전자(005930)



■ Compliance notice

- 당사는 2025년 9월 12일 현재 SK하이닉스, 삼성전자 종목의 발행주식을 1%이상 보유하고 있지 않습니다.
- 당사는 동 리포트의 내용 일부를 기관투자가 또는 제3자에게 사전에 제공한 사실이 없습니다.
- 동 리포트의 금융투자분석사와 배우자는 상기 발행주식을 보유하고 있지 않습니다.
- 당사는 SK하이닉스, 삼성전자 발행주식을 기초자산으로 하는 ELW(주식워런트증권)를 발행 중이며, 당해 ELW에 대한 유동성공급자(LP)입니다.

■ 기업 투자의견은 향후 12개월간 시장 지수 대비 주가등락 기준임

- 매　　수 : 시장 지수 대비 15%p 이상의 주가 상승 예상
- 중　　립 : 시장 지수 대비 -15~15%p의 주가 등락 예상
- 비중축소 : 시장 지수 대비 15%p 이상의 주가 하락 예상
- 중립 및 비중축소 의견은 목표가 미제시

■ 투자등급 비율 (2025. 6. 30 기준)

매수	중립	비중축소(매도)
77.9%	22.1%	0%

※ 최근 1년간 공표한 유니버스 종목 기준

■ 업종 투자의견은 향후 12개월간 해당 업종의 유가증권시장(코스닥) 시가총액 비중 대비 포트폴리오 구성 비중에 대한 의견임

- 비중확대 : 해당업종의 포트폴리오 구성비중을 유가증권시장(코스닥)시가총액 비중보다 높이 가져갈 것을 권함
- 중　　립 : 해당업종의 포트폴리오 구성비중을 유가증권시장(코스닥)시가총액 비중과 같게 가져갈 것을 권함
- 비중축소 : 해당업종의 포트폴리오 구성비중을 유가증권시장(코스닥)시가총액 비중보다 낮게 가져갈 것을 권함

■ 본 리포트는 고객의 증권투자를 돕기 위하여 작성된 당사의 저작물로서 모든 저작권은 당사에게 있으며, 당사의 동의 없이 어떤 형태로든 복제, 배포, 전송, 변형할 수 없습니다.

■ 본 리포트는 당사 리서치본부에서 수집한 자료 및 정보를 기초로 작성된 것이나 당사가 그 자료 및 정보의 정확성이나 완전성을 보장할 수는 없으므로 당사는 본 리포트로써 고객의 투자 결과에 대한 어떠한 보장도 행하는 것이 아닙니다. 최종적 투자 결정은 고객의 판단에 기초한 것이며 본 리포트는 투자 결과와 관련한 법적 분쟁에서 증거로 사용될 수 없습니다.

■ 본 리포트에 제시된 종목들은 리서치본부에서 수집한 자료 및 정보 또는 계량화된 모델을 기초로 작성된 것이나, 당사의 공식적인 의견과는 다를 수 있습니다.

■ 이 리포트에 게재된 내용들은 작성자의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.