(CBRS.US)

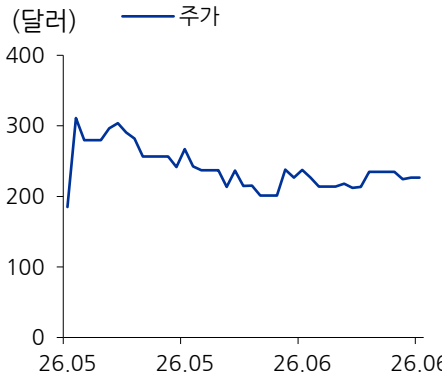
Cerebras Systems

2026.06.24

글로벌 테크 박재환
6158 / jaehwan124@eugenefn.com

GPU 킬러 (X), Decode 킬러 (O)

투자의견	NA
현재주가	226.72 USD(6/23)
시가총액	50(십억 달러)/77(조 원)



❝ 세레브라스는 WSE(Wafer Scale Engine)을 설계하는 팹리스 업체. 일반적인 반도체 칩이 300mm 웨이퍼 위에 다수의 다이(Die)를 형성한 후 이를 잘라내어 개별적으로 패키징 하는 것과 다르게, WSE는 웨이퍼 전체를 하나의 거대한 컴퓨팅 패브릭으로 활용. 수율 문제는 양품 코어만 활성화하고 결함 코어를 우회하는 라우팅 방식으로 극복. TSMC의 N5 공정으로 제조되며, **HBM을 사용하는 대신 칩 면적의 50% 수준을 SRAM으로 배정해 칩 당 44GB의 SRAM 용량과 21PB/s의 대역폭을 제공**. 또한 WSE-3을 전력, 냉각 장비와 결합한 CS-3 랙시스템의 형태로 판매.

❝ 코어 기준 1분기 매출 1.91억달러(+92% yoy), GPM 46.5%, OPM -2%를 기록하며 컨센을 상회. 하드웨어 매출은 1.12억달러(+60% yoy), 클라우드/서비스 매출은 7,980만달러(+167% yoy)를 기록. 가이던스는 2분기 코어 매출액 1.94억달러(+88% yoy), 연간 코어 매출액 8.6억달러를 제시.

❝ 오픈AI와 200억달러(750MW) 규모의 컴퓨팅 다년 공급 계약을 체결했으며, 상용 배포가 진행 중. 다만 이 중 일부는 "Pass-Through" 형태의 계약으로 마진은 3% 이하 수준에 불과. 이는 세레브라스의 전체 마진 대비 현저히 낮기 때문에, 코어 지표에서는 제외됨. AWS와의 계약은 Trainium3와 CS-3의 분산 추론 형태이며 Trainium은 추론의 Prefill을, CS-3는 Decode를 담당하는 구조. 계약 규모는 공개되지 않았으며, 본격적인 출하는 2027년부터 시작될 것으로 언급.

❝ 또한 급증하는 세레브라스 자체 클라우드 수요로 인해 이미 하드웨어를 판매한 고객에게 시스템을 다시 빌려오는 "Rent Back" 형태의 계약을 추진 중. 이는 단기적으로 클라우드 마진에 10%p 이상의 부담으로 작용할 전망이며, 수익성 악화를 감안하더라도 시장 점유율 확대를 우선으로 하는 전략으로 판단. 이로 인해 코어 기준 2분기 GPM과 OPM 가이던스는 각각 전분기 대비 악화된 37%, -31%로 제시.

❝ **세레브라스는 일반적인 AI 가속기와 달리 SRAM 용량 한계로 인해 모델 파라미터를 MemoryX라는 외부 메모리 계층에 저장하는 방식을 채택.** 사측은 WSE/CS-3을 학습과 추론이 모두 가능한 웨이퍼 스케일 시스템으로 포지셔닝 중이나, 외부 메모리 계층에 의존하는 데에 반해 칩 외부 시스템 I/O 대역폭은 CS-3 1대 기준 150GB/s 수준으로 **아직까지 대형 모델이나 대규모 컨텍스트 향으로 사용되기에는 시스템적 한계가 존재.** AI Agent 확산으로 일부 에이전트 루프에서 반복적인 Decode 호출과 메모리 접근 지연시간의 중요성이 커질수록 WSE의 활용처는 확대될 수 있음. 다만 **WSE가 AI 인프라 시장 전반으로 유의미하게 확산되기 위해서는 외부 메모리·시스템 I/O 병목 완화와 소프트웨어 생태계 확장이 필요하다고 판단.**

현지명	Cerebras Systems Inc
한글명	세레브라스 시스템스
시가총액(십억 달러/조원)	50/77
설립연도	2015 년
설립자	Andrew Feldman
본사 위치	미국 캘리포니아
현 CEO	Andrew Feldman
52 주 최고/최저(달러)	386/197
배당수익률(26F, %)	0.0
주요주주 지분율(%)	
이클립스 벤처	39.0
FMR	27.1
JP 모건	10.1

주가상승률(%)	1M	6M	YTD
	-11.7	-	-

(Non GAAP)	FY2025	FY2026F	FY2027F
매출액(백만 달러)	-	838	2,924
영업이익(백만 달러)	-	-362	270
당기순이익(백만 달러)	-	-327	173
EPS(달러)	-	-1.74	0.59
증감률(%)	-	적전	흑전
PER(배)	-	-	386.9
ROE(%)	-	-10.2	4.1
PBR(배)	-	6.5	6.6
EV/EBITDA(배)	-	-	45.4

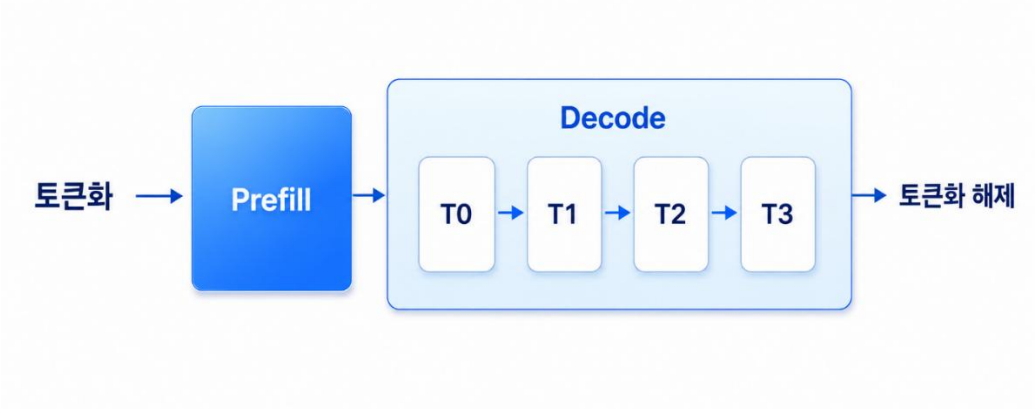
유진투자증권

추론의 두 단계: Prefill 과 Decode

LLM 추론은 크게 Prefill 과 Decode 의 두 단계로 구분된다. Prefill 은 사용자가 입력한 프롬프트를 한 번에 읽고 문맥을 해석하는 구간이다. 입력 토큰들을 병렬적으로 처리한다는 점에서 학습과 유사한 성격을 가지며, 상대적으로 병렬 연산 집약적인 워크로드에 해당한다. 따라서 Prefill 단계에서는 GPU 의 연산 성능과 HBM 대역폭이 핵심 자원으로 작용한다.

반면 Decode 는 Prefill 결과를 바탕으로 다음 토큰을 순차적으로 생성하는 단계다. 다음 토큰은 이전 토큰의 결과에 의존하기 때문에 병렬화 강도가 낮고, **매 토큰 생성 과정에서 모델 파라미터와 KV Cache 에 반복적으로 접근**해야 한다. 이에 따라 Decode 단계에서는 **단순 연산 성능보다 메모리 접근 지연시간과 시스템 계층의 데이터 이동 효율이 핵심 병목으로 작용**한다.

추론의 과정



자료: 유진투자증권

Prefill 과 Decode 의 차이

구분	Prefill	Decode
역할	입력 프롬프트를 한번에 읽고 문맥을 해석하는 구간	생성된 문맥을 바탕으로 토큰을 생성하는 구간
처리 방식	다수의 입력 토큰을 동시 처리	토큰을 순차적으로 생성
병렬화 강도	높음	낮음
성격	연산 집약적	메모리·레이턴시 집약적
핵심 자원	GPU 연산 자원	메모리·시스템 계층

자료: 유진투자증권

WSE: 웨이퍼 전체를 하나의 패브릭으로

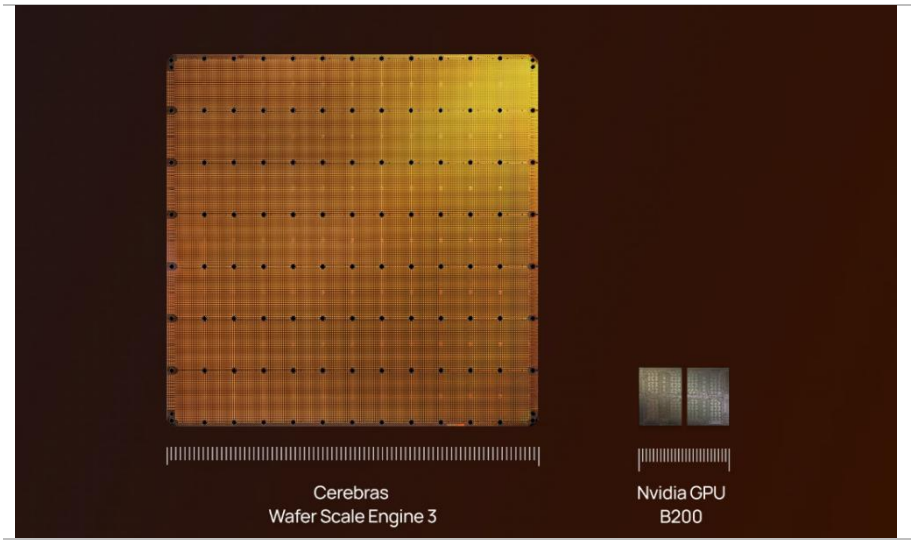
일반적인 반도체 칩은 300mm 웨이퍼 위에 다수의 다이(Die)를 형성한 뒤, 이를 절단해 개별 GPU·CPU로 패키징한다. 이후 개별 칩은 NVLink와 같은 외부 인터커넥트를 통해 클러스터 단위로 연결된다. 반면 세레브라스의 WSE(Wafer-Scale Engine)는 **웨이퍼 전체를 절단하지 않고 하나의 거대한 칩으로 사용하는 구조**다.

WSE-3는 세레브라스의 선단 프로세서로, TSMC N5 공정으로 제조된다. WSE-3는 총 84개(12x7)의 다이가 하나의 실리콘 위에서 연결되어 있으며, 약 97만개의 물리 코어가 집적되어 있다. 이 중 약 90만개 수준의 양품 코어만 실제로 활성화된다. 웨이퍼 전체를 하나의 칩으로 사용할 경우 수율 문제가 가장 큰 제약으로 작용하지만, 세레브라스는 여유 코어를 배치하고 결함 코어를 우회하는 라우팅 구조를 통해 이를 보완했다. 또한 AI XPU 대비 상대적으로 성숙한 공정을 사용함으로써 제조 안정성을 높인 것으로 판단한다.

WSE의 가장 특징적인 부분은 HBM을 사용하지 않는다는 점이다. 대신 웨이퍼 면적의 50% 수준을 SRAM에 할당해 총 44GB의 SRAM 온칩 메모리와 21PB/s의 대역폭을 제공한다. 이는 메모리에 빠르게 접근해 토큰을 순차적으로 빠르게 생성해야 하는 **Decode 워크로드에 특화된 설계라고 판단**한다.

WSE는 B200 대비 29배의 크기, 19배 많은 트랜지스터가 집적

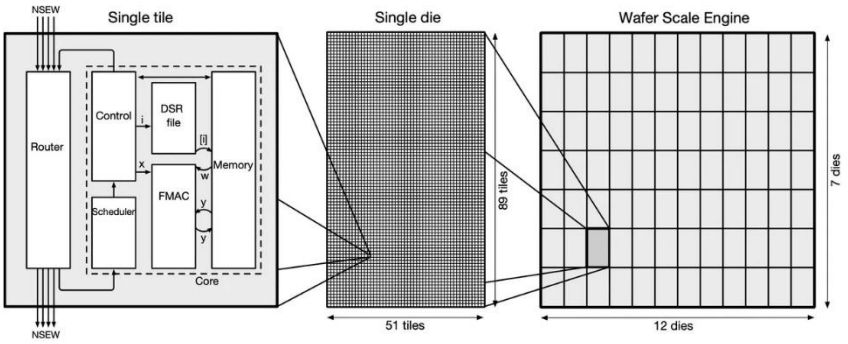
세레브라스 WSE와 블랙웰 GPU 다이 크기 비교



자료: Cerebras, 유진투자증권

총 97 만개의 코어,
84 개의 다이가 하나의
웨이퍼 위에서 연결

세레브라스 WSE 구조



자료: Cerebras, 유진투자증권

WSE-3 기반 랙시스템 CS-3

CS-3 는 WSE-3 를 탑재한 세레브라스의 데이터센터용 랙 시스템이다. 세레브라스는 WSE 를 엔진 블록(Engine Block)의 형태로 패키징하고, CPU, 전력 인입부, 액체 냉각 장비를 결합해 하나의 시스템으로 구성한다.

CS-3 의 시스템 확장은 SwarmX 와 MemoryX 를 통해 구현된다. SwarmX 는 다수의 CS-3 을 연결하는 400G/800G 스케일아웃 네트워크로 통해 최대 2,048 대 규모의 클러스터 구성을 지원한다. 다만 **WSE/CS-3 은 SRAM 만을 메모리로 사용하는 구조의 특성상 용량 한계가 존재하기 때문에, 대규모 모델 파라미터 전체를 칩에 저장하기 힘들다.** 따라서 세레브라스는 DRAM/NAND 기반 **중앙 메모리 시스템인 MemoryX 에 모델 파라미터를 별도로 저장**하고, 연산에 필요한 파라미터를 WSE 로 순차적으로 공급하는 **파라미터 스트리밍 방식**이 적용한다.

세레브라스 엔진 블록(Engine Block)

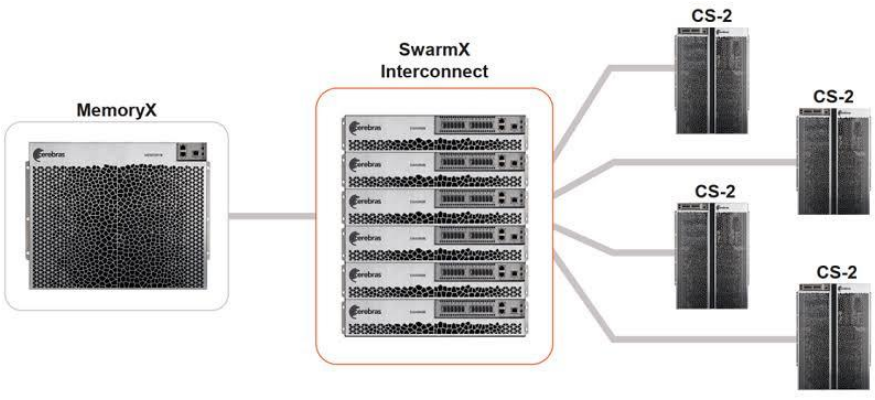
자료: Cerebras, 유진투자증권

세레브라스 CS-3

자료: Cerebras, 유진투자증권

MemoryX, SwarmX 와 연결되어 클러스터 시스템을 형성

세레브라스 클러스터 시스템



자료: Cerebras, 유진투자증권

엔비디아를 대체하긴 어렵다

WSE-3 는 44GB 의 SRAM 온칩 메모리와 21PB/s 의 온칩 대역폭을 기반으로 **메모리에 빠르게 접근해 토큰을 순차적으로 빠르게 생성해야 하는 Decode 워크로드에 특화되었다고 판단**한다. 따라서 **직접적인 비교 대상은 엔비디아의 Vera Rubin NVL72 라기보다 Groq 3 LPU 와 같은 저지연 추론 시스템**에 가깝다. Groq 3 LPU 역시 프로세서당 500MB 의 SRAM 을 탑재하고, LPX 랙 시스템당 256 개의 LPU 를 통해 128GB 의 SRAM 용량과 640TB/s 의 대역폭을 제공한다.

Vera Rubin NVL72 는 대규모 HBM 용량, GPU 연산 성능, NVLink 기반 스케일업 구조를 바탕으로 학습, Prefill, 긴 컨텍스트, 범용 추론 전반에서 강점을 가진다. 반면 WSE 와 LPU 는 토큰을 빠르게 순차 생성해야 하는 Decode, 짧은 응답 지연시간이 중요한 일부 실시간 추론, 반복 호출이 발생하는 일부 에이전트 루프에서 강점을 보일 수 있다.

랙시스템 별 포지션

작업	요구	VR NVL72	Groq 3 LPX	WSE/CS-3
학습	연산성능, 메모리 용량/대역폭	매우 강함	매우 약함	약함
추론 Prefill	연산성능, 메모리 용량/대역폭	매우 강함	약함	강함
추론 Decode	연산성능, 저지연 메모리	강함	매우 강함	매우 강함
긴 컨텍스트	메모리 용량/대역폭	매우 강함	약함	약함
에이전트 루프	연산성능, 저지연 메모리	강함	매우 강함	매우 강함

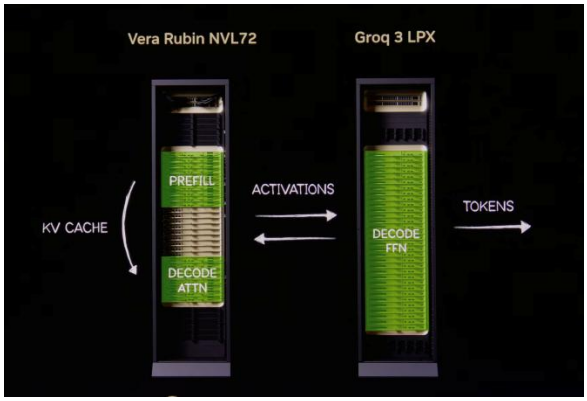
자료: Bloomberg, 유진투자증권

따라서 WSE 는 단독으로 AI 워크로드를 모두 처리하는 시스템이라기보다, **범용 프로세서와 결합되어 특정 추론 구간을 가속하는 형태로 활용**될 가능성이 높다. 엔비디아가 NVL72 와 Groq LPU 연결해 Prefill 과 Decode 를 분리하는 것과 유사하게, 세레브라스 역시 범용 AI 가속기와 WSE 를 결합하는 분리형 추론 구조에서 효용이 가장 명확해질 수 있다. 실제로 AWS 와의 협업에서도 Trainium 은 Prefill 을, CS-3 은 Decode 를 담당하는 구조를 통해 추론 워크로드를 분리 처리하는 방식을 추진 중이다.

또한 WSE 시스템 내부에서는 약 21PB/s 의 온칩 메모리 대역폭을 제공하지만, CS-3 랙(CS-3 2 대 탑재 가정)의 시스템 I/O 대역폭은 2.4Tb/s, 약 300GB/s 수준으로 추정된다. 엔비디아 NVL72 와 달리 모델 파라미터를 외부 MemoryX 에 배치하는 구조는 **대규모 모델 학습이나 긴 컨텍스트 처리에서 GPU 기반 랙 시스템 대비 구조적인 메모리 병목으로 작용할 수 있다.**

결론적으로 **세레브라스의 WSE 는 엔비디아의 범용 AI 랙 시스템을 대체하는 것이 아니라, 저지연 Decode 에 특화된 니치형 추론 시스템으로 판단**한다. AI Agent 확산으로 반복적인 Decode 호출과 짧은 응답 지연시간의 중요성이 커질수록 WSE 의 활용처는 확대될 수 있으나, **WSE 가 AI 인프라 시장 전반으로 확산되기 위해서는 외부 메모리·시스템 I/O 병목 완화와 소프트웨어 생태계 확장이 필요**하다고 판단한다.

GPU 와 LPU 의 추론 분담



자료: Nvidia, 유진투자증권

AWS Trainium 과 WSE 의 추론 분담

자료: Cerebras, 유진투자증권

프로세서 스펙 비교

구분	Rubin GPU	Groq 3 LPU	WSE-3
노드	N3	SF4X	N5
FP8 Dense (PFLOPS)	17.5	1.2	15.6
메모리	HBM4	SRAM	SRAM
메모리 용량	288GB	500MB	44GB
메모리 대역폭	20.5TB/s	150TB/s	21PB/s
스케일업 대역폭	3.6TB/s	2.5TB/s	-
스케일아웃 대역폭	400GB/s	-	150GB/s

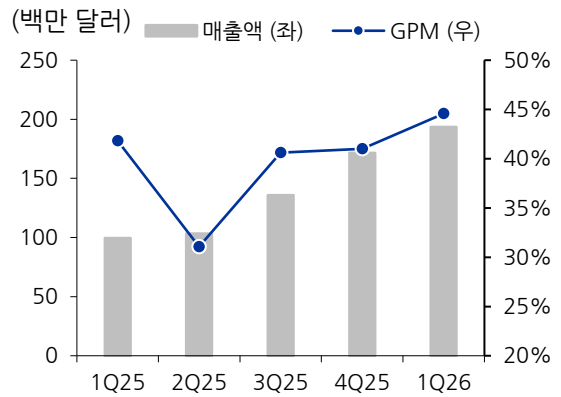
자료: 각 사, SemiAnalysis, 유진투자증권

랙시스템 스펙 비교

구분	Vera Rubin NVL72	Groq 3 LPX	CS-3 (추정)
주요 프로세서	Rubin GPU	Groq 3 LPU	WSE-3
랙 당 주요 프로세서 수	72	256	2
랙 당 CPU 수	36	32	4
주요 메모리	HBM	SRAM	SRAM
랙 당 메모리 용량	20.7TB	128GB	88GB
랙 당 메모리 대역폭	1.6PB/s	40PB/s	42PB/s
스케일업 대역폭	260TB/s	640TB/s	-
스케일아웃 대역폭	28.8TB/s	미공개	300GB/s

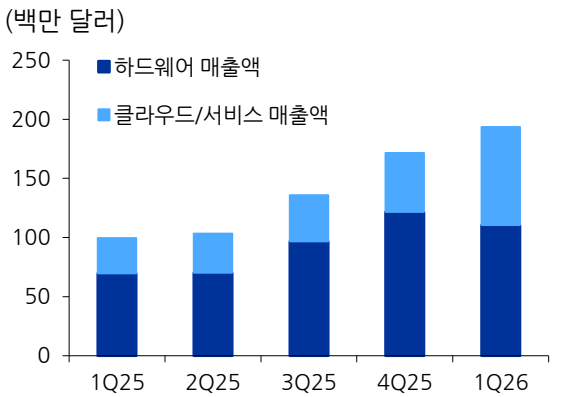
자료: 각 사, 유진투자증권

세레브라스 매출액, GPM 추이



자료: Bloomberg, 유진투자증권

세레브라스 매출액 구성



자료: Bloomberg, 유진투자증권

Compliance Notice

당사는 자료 작성일 기준으로 지난 3 개월 간 해당종목에 대해서 유가증권 발행에 참여한 적이 없습니다
당사는 본 자료 발간일을 기준으로 해당종목의 주식을 1% 이상 보유하고 있지 않습니다
당사는 동 자료를 기관투자가 또는 제 3 자에게 사전 제공한 사실이 없습니다
조사분석담당자는 자료작성일 현재 동 종목과 관련하여 재산적 이해관계가 없습니다
동 자료에 게재된 내용들은 조사분석담당자 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다

동 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다

투자기간 및 투자등급/투자의견 비율

종목추천 및 업종추천 투자기간: 12 개월 (추천기준일 종가대비 추천종목의 예상 목표수익률을 의미함)		당사 투자의견 비율(%)
· STRONG BUY(매수)	추천기준일 종가대비 +50%이상	0%
· BUY(매수)	추천기준일 종가대비 +15%이상 ~ +50%미만	95%
· HOLD(중립)	추천기준일 종가대비 -10%이상 ~ +15%미만	5%
· REDUCE(매도)	추천기준일 종가대비 -10%미만	0%

(2026.03.31 기준)