



반도체 (Positive)

Rubin의 출하 확대, HBM4 성장의 시작

▶ Analyst 박준영 jyp94@hanwha.com 3772-7481 / RA 김나우 now.kim@hanwha.com 3772-7710

NVIDIA 의 차세대 AI 플랫폼 Vera Rubin 이 초기 상용화 단계에 진입했다. 2026 년 하반기 생산 출하가 시작된 이후 주요 클라우드 사업자를 중심으로 고객 인도가 진행되고 있으며, CoreWeave 는 9 월 말 Rubin NVL72 기반 서비스를 상용화했다. AI 코딩 에이전트 개발사 Cognition 역시 Rubin 기반 인프라에서 실제 추론 워크로드를 운영하기 시작했다. 차세대 GPU 가 제품 발표와 성능 검증을 넘어 실제 고객 환경에서 활용되기 시작했다는 점은 향후 생산량 확대를 기대할 수 있는 중요한 변화다.

Rubin 의 고객 기반도 확대되고 있다. NVIDIA 는 AWS, Microsoft, Google, Oracle 등 글로벌 클라우드 사업자를 Rubin 생태계에 포함했으며, CoreWeave, Nebius, Lambda 등 네오클라우드 사업자 역시 차세대 시스템 도입을 추진하고 있다. AI 모델 개발사들도 더 높은 추론 성능과 낮은 운영비용을 확보하기 위해 차세대 인프라로의 전환을 준비하고 있다. 주요 클라우드 사업자와 AI 서비스 기업의 수요가 동시에 발생하고 있다는 점에서 Rubin은 단순한 기존 GPU의 교체를 넘어 신규 AI 인프라 투자 확대의 수혜를 받을 것으로 예상된다.

Rubin 은 GPU 연산 성능뿐 아니라 메모리와 GPU 간 통신 구조를 동시에 개선한 플랫폼이다. Rubin NVL72는 72 개의 GPU 와 36 개의 Vera CPU 를 통합했으며, GPU 당 288GB 의 HBM4 와 NVLink 6 를 채택했다. 특히 HBM 대역폭은 GPU 당 최대 22TB/s 로 확대됐으며, GPU 간 연결 대역폭도 3.6TB/s 로 증가했다. 대규모 AI 모델의 추론 과정에서 발생하는 메모리와 GPU 간 데이터 이동 병목을 완화하기 위한 설계다.

실제 성능 개선도 확인되고 있다. MLPerf Inference v6.1 에서 Rubin 은 Qwen3-VL 멀티모달 추론 기준 GB300 NVL72 대비 최대 3.7 배 높은 처리량을 기록했다. CoreWeave 의 고객사 Cognition 은 AI 코딩 에이전트 추론에서 GB200 NVL72 대비 최대 4.8 배 높은 토큰 처리량을 확인했다. 모델과 측정 조건에 따라 개선 폭은 달라지지만 Rubin 의 기술적 변화가 실제 AI 서비스의 처리 능력 향상으로 연결되고 있다는 점은 공통적이다.

특히 AI 에이전트의 확산은 고성능 GPU 의 활용 범위를 넓힐 것으로 전망한다. 기존 챗봇이 질문에 대한 답변을 생성하는 데 집중했다면, AI 에이전트는 업무 처리, 코드 작성, 분석 등 복수의 작업을 반복적으로 수행한다. 복잡한 작업일수록 더 많은 토큰 생성과 긴 문맥(Context) 처리가 필요하므로 추론 성능 개선의 가치가 높아진다. Rubin 의 처리량 증가와 추론 비용 절감은 이러한 서비스의 상용화 범위를 확대하고, 결과적으로 AI 인프라에 대한 추가적인 수요를 창출할 것으로 판단한다.

과거 NVIDIA 의 GPU 세대교체를 살펴보면 초기 고객 인도 이후 본격적인 공급 확대까지 일정한 시차가 존재했다. Hopper는 2022년 하반기 출하를 시작한 이후 2023년 생성형 AI 인프라 투자가 확대되면서 공급량이 빠르게 증가했다. Blackwell 역시 2024년 4분기 생산 출하를 시작한 이후 2025년 상반기 랙 단위 시스템 공급이 본격화됐다.

Rubin 역시 초기 고객 인도 이후 시스템 통합과 공급망 안정화를 거치며 생산량이 증가할 전망이다. 특히 Blackwell 세대에서 축적된 랙 시스템 제조 경험과 개선된 모듈형 설계를 고려하면 생산 확대 속도는 과거보다 빨라질 가능성도 있다. 당사는 2026 년 4 분기 Rubin 의 고객 인도가 확대되고, 2027 년 1 분기부터 본격적인 대량 출하 구간에 진입할 것으로 전망한다.

이에 따라 2027 년 국내 메모리 업체의 핵심 성장 동력은 HBM4 가 될 전망이다. NVIDIA 의 Rubin 이 2026 년 하반기 초기 고객 인도를 시작한 가운데, 2027 년에는 본격적인 출하량 확대가 예상된다. 특히 2027 년 상반기 Rubin 의 생산량이 증가하는 과정에서 HBM4 공급도 빠르게 확대될 가능성이 높다. HBM 은 GPU 패키징과 시스템 조립에 앞서 공급되기 때문에 최종 시스템 출하보다 메모리 수요가 선행해서 발생하는 구조다.

Rubin NVL72 는 72 개의 GPU 를 탑재하며 시스템 한 대에 약 20.7TB 의 HBM4 가 필요하다. 따라서 Rubin 의 생산량이 증가할수록 HBM4 수요도 직접적으로 확대된다. 특히 2026 년이 HBM4 의 초기 양산 및 공급이 시작된 해라면 2027 년은 Rubin 의 대량 생산과 함께 HBM4 가 주력 제품으로 자리 잡는 해가 될 것으로 판단한다.

HBM4 는 기존 HBM3E 대비 높은 메모리 대역폭을 제공한다. Rubin 의 GPU 당 HBM 대역폭은 최대 22TB/s 로 Blackwell 대비 약 2.75 배 확대됐다. 최근 AI 모델은 긴 문맥과 복잡한 추론을 처리하는 과정에서 대량의 데이터를 메모리에서 반복적으로 불러와야 한다. 이에 따라 HBM 의 성능은 GPU 의 실제 추론 처리량을 결정하는 중요한 요소로 작용한다. Rubin 의 초기 성능 검증은 이러한 메모리 성능 개선이 실제 AI 서비스의 처리량 증가로 이어질 수 있음을 보여준다.

NVIDIA GPU 외에도 주요 클라우드 사업자들이 자체 AI 가속기 개발을 확대하고 있다는 점에 주목할 필요가 있다. AI 모델의 학습과 추론에 필요한 연산량이 증가하면서 각 기업은 용도에 최적화된 AI ASIC 을 개발하고 있다. 이들 제품에서도 대용량 고대역폭 메모리의 중요성이 높아지고 있어, 향후 HBM 수요는 NVIDIA 의 제품 출하뿐 아니라 다양한 AI 가속기 플랫폼으로 확대될 가능성이 높다.

HBM4 의 본격적인 공급 확대는 국내 메모리 업체의 제품 믹스 개선으로도 연결될 전망이다. 2027 년 HBM4 의 출하 증가율은 초기 양산 단계였던 2026 년의 낮은 기저를 고려할 때 전체 HBM 시장의 성장률을 상당 폭 상회할 가능성이 높다. 출하량 증가와 고부가 제품 비중 확대가 동시에 나타난다면 국내 메모리 업체의 매출 성장에도 긍정적으로 작용할 전망이다.

이러한 흐름은 Rubin 이후에도 이어질 가능성이 높다. NVIDIA 는 2027 년 Rubin Ultra, 2028 년 Feynman 으로 이어지는 차세대 GPU 아키텍처를 준비하고 있다. AI 모델의 규모와 복잡성이 증가하면서 차세대 GPU 역시 높은 메모리 대역폭과 대용량 메모리를 요구할 것으로 예상된다. 따라서 HBM4 를 시작으로 차세대 HBM 기술의 중요성은 지속적으로 높아질 전망이다.

2027 년은 HBM4 가 국내 메모리 업체의 새로운 실적 성장 동력으로 자리 잡는 해가 될 것으로 판단한다. Rubin 의 초기 고객 인도가 시작된 가운데 과거 Hopper 와 Blackwell 의 사례를 고려하면 2027 년 상반기부터 출하량이 본격적으로 증가할 가능성이 높다. 여기에 AI 에이전트의 확산과 신규 AI 인프라 투자가 더해지면서 HBM4 수요도 빠르게 증가할 것으로 예상된다. 국내 메모리 업체는 HBM4 공급 확대와 고부가 제품 비중 상승을 바탕으로 2027 년 실적 성장을 이어갈 전망이다.

[표1] Rubin의 Blackwell 대비 성능 개선

구분	비교 기준	Rubin 성능
Qwen3-VL 추론 처리량	GB300 NVL72	최대 3.7배
DeepSeek-R1 추론 처리량	GB300 NVL72	최대 2.5배
Cognition AI 코딩 추론	GB200 NVL72	최대 4.8배
Cognition 강화학습 출력	GB200 NVL72	3.8배
GPU당 HBM 대역폭	GB200	2.75배
GPU당 NVLink 대역폭	GB200	2.0배

자료: NVIDIA, MLPerf Inference v6.1, CoreWeave, 한화투자증권 리서치센터

[표2] 과거 GPU 세대교체와 Rubin 출하 전망

구분	Hopper	Blackwell	Rubin
초기 출하	2022년 하반기	2024년 4분기	2026년 하반기
출하 확대	2023년	2025년 상반기	2027년 상반기 전망
핵심 메모리	HBM3	HBM3E	HBM4
주요 수요	생성형 AI 학습	학습 추론 확대	AI 에이전트·대규모 추론
핵심 변화	AI GPU 수요 확대	랙 스케일 시스템	추론 경제성·HBM4 확대

자료: NVIDIA, 한화투자증권 리서치센터

[Compliance Notice]

(공표일: 2026년 10월 13일)

이 자료는 조사분석 담당자 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성됐음을 확인합니다. 본인은 이 자료에서 다룬 종목과 관련해 공표일 현재 관련 법규상 알려야 할 재산적 이해관계가 없습니다. 본인은 이 자료를 기관투자자 또는 제3자에게 사전에 제공한 사실이 없습니다. (박준영, 김나우)

저희 회사는 공표일 현재 이 자료에서 다룬 종목의 발행주식을 1% 이상 보유하고 있지 않습니다.

이 자료는 투자자의 증권투자를 돕기 위해 당사 고객에 한하여 배포되는 자료로서 저작권이 당사에 있으며 불법 복제 및 배포를 금합니다. 이 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료나 정보출처로부터 얻은 것이지만, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 이 자료는 어떠한 경우에도 고객의 증권투자 결과와 관련된 법적 책임소지에 대한 증빙으로 사용될 수 없습니다.

[종목 투자등급]

당사는 개별 종목에 대해 향후 1년간 +15% 이상의 절대수익률이 기대되는 종목에 대해 Buy(매수) 의견을 제시합니다. 또한 절대수익률 -15~+15%가 예상되는 종목에 대해 Hold(보유) 의견을, -15% 이하가 예상되는 종목에 대해 Sell(매도) 의견을 제시합니다. 밸류에이션 방법 등 절대수익률 산정은 개별 종목을 커버하는 애널리스트의 추정에 따르며, 목표주가 산정이나 투자의견 변경 주기는 종목별로 다릅니다.

[산업 투자의견]

당사는 산업에 대해 향후 1년간 해당 업종의 수익률이 과거 수익률에 비해 양호한 흐름을 보일 것으로 예상되는 경우에 Positive(긍정적) 의견을 제시하고 있습니다. 또한 향후 1년간 수익률이 과거 수익률과 유사한 흐름을 보일 것으로 예상되는 경우에 Neutral(중립적) 의견을, 과거 수익률보다 부진한 흐름을 보일 것으로 예상되는 경우에 Negative(부정적) 의견을 제시하고 있습니다. 산업별 수익률 전망은 해당 산업 내 분석대상 종목들에 대한 담당 애널리스트의 분석과 판단에 따릅니다.

[당사 조사분석자료의 투자등급 부여 비중]

(기준일: 2026년 9월 30일)

투자등급	매수	중립	매도	합계
금융투자상품의 비중	96.4%	3.6%	0.0%	100.0%