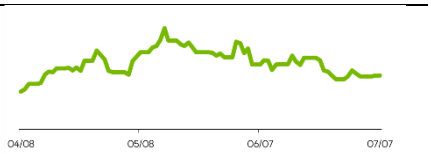
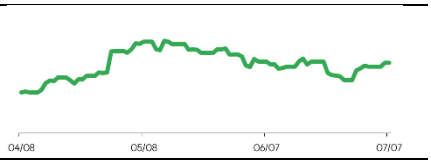
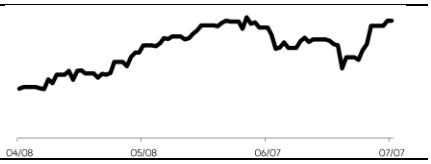
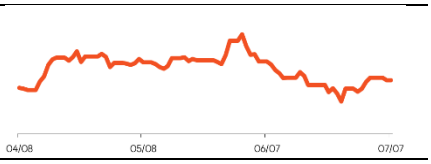
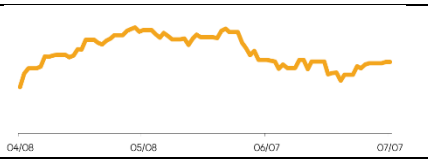
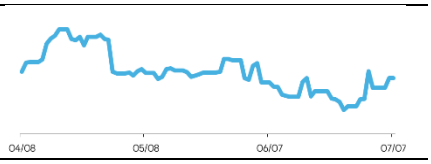
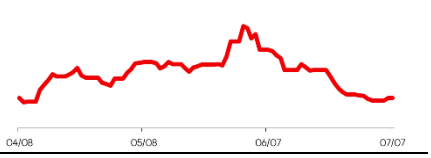
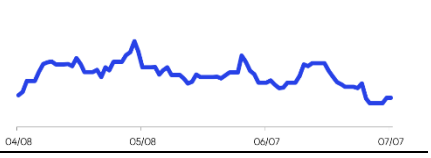
2026-07-08

박제민의 Q&A.I

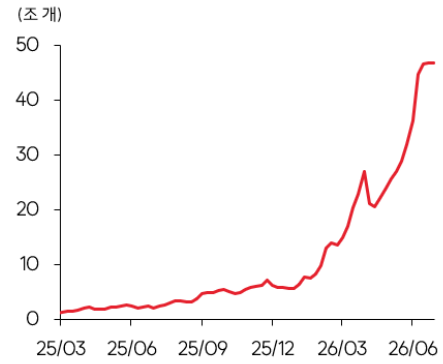
SK securities
SK증권 해외주식/AI. 박제민
jeminwa@sks.co.kr

	엔비디아	시가총액: 4,736	12MF PER: 19	
		-독보적인 GPU 기술력있으나 ASIC 으로 독점력 축소 -Rubin Ultra 기술 전환 및 CPU, LPU 주문 추이 주목		
	알파벳	시가총액: 4,463	12MF PER: 29	
		- TPU 활용 Anthropic 독주로 인프라 해자 인정받는 중 - Coding Agent 역전, 소비자용 Agent 출시 주목		
	애플	시가총액: 4,592	12MF PER: 34	
		-메모리 가격 상승이 브랜드 파워로 극복 가능한지 시험 중 -Siri 고도화로 인한 소비자용 Agent 출시 주목		
	마이크로소프트	시가총액: 2,872	12MF PER: 20	
		-OpenAI 해자 약화로 인프라 사업부 성장률 축소 -자체 모델 및 커스텀 AI 도입 컨셉 수요 및 기술력 주목		
	아마존	시가총액: 2,626	12MF PER: 29	
		-양대 AI Labs 모두에 공격적 투자를 통한 고객 확보 성공 -인프라 외 Bedrock 등 AIaaS ARR 전망에 주목		
	메타 플랫폼스	시가총액: 1,523	12MF PER: 18	
		-플랫폼 인기 여전, Muse Spark 개발로 유통과 모델 확보 -플랫폼 데이터, 글래스 등 활용한 소비자용 Agent 주목		
	오라클	시가총액: 414	12MF PER: 17	
		-부품 비용, 자본 비용 증가 추세 속 높은 수주 잔고 무색 -인력, 전기 쇼티지에도 매출 전환 성공 여부 주목		
	코어위브	시가총액: 47	12MF PER: -	
		-부품 비용, 자본 비용 증가 추세 속 높은 수주 잔고 무색 -인력, 전기 쇼티지에도 매출 전환 성공 여부 주목		

주: 시가총액 십억달러

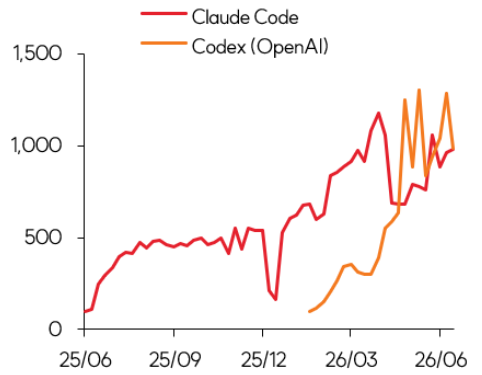
Update Chart

OpenRouter 토큰 사용량



자료: OpenRouter, SK 증권

주요 Coding Agent 활용 강도 (백만회 달성일 = 100)

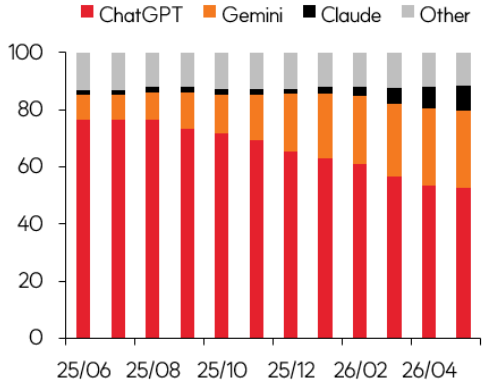


자료: NPMjs, SK 증권 / 주: 업데이트 및 다운로드 수 기준. Codex 5 월 수치 일부 조정

GPU 렌탈 가격 추이 (SiliconData)

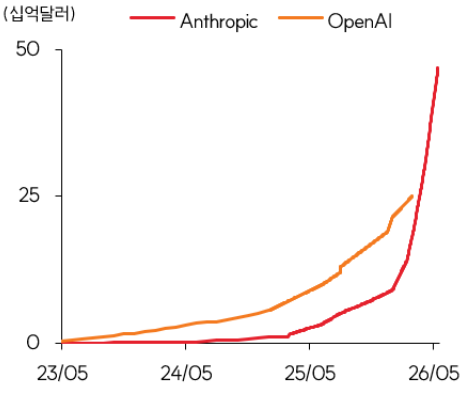
자료: Bloomberg, SK 증권

LLM 별 점유율 추이



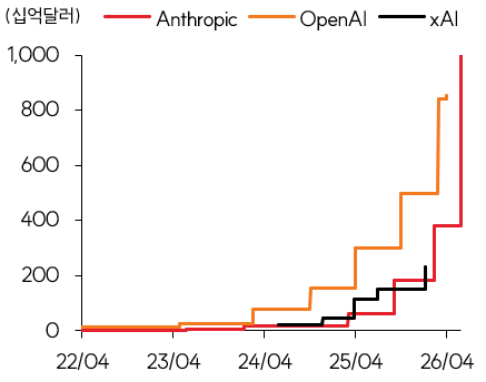
자료: SimilarWeb, SK 증권

주요 AI Labs ARR 추이



자료: EpochAI, SK 증권

주요 AI Labs 기업가치 추이



자료: EpochAI, SK 증권

최적화는 AI Labs 경쟁 구도의 핵심 변수?

OpenAI 내부자: 신규 발견 최적화로 추론 비용 절반 이상 절감

AI Infra 가 비싸짐에 따라 같은 AI 워크로드 수행에 필요한 추론 비용 감소 노력 지속

6 월초 OpenAI 는 기존 모델을 실행하는 추론 비용을 절반 이상 줄이는 방법 발견

구체적 기법 비공개, 가능성으로 양자화(quantization)·KV 캐싱·쿼리 배치 처리·저전력 모델/부분 라우팅 등 거론

토큰 제공이 제한적인 무료 버전 사용자(또는 로그 아웃 사용자)에 한해서 서비스 시 필요 GPU 수를 수백대 수준으로 축소

현재 OpenAI 의 온라인 GPU 규모는 150~200 만대 수준으로 추정, 무료 티어라 하더라도 큰 최적화 발전

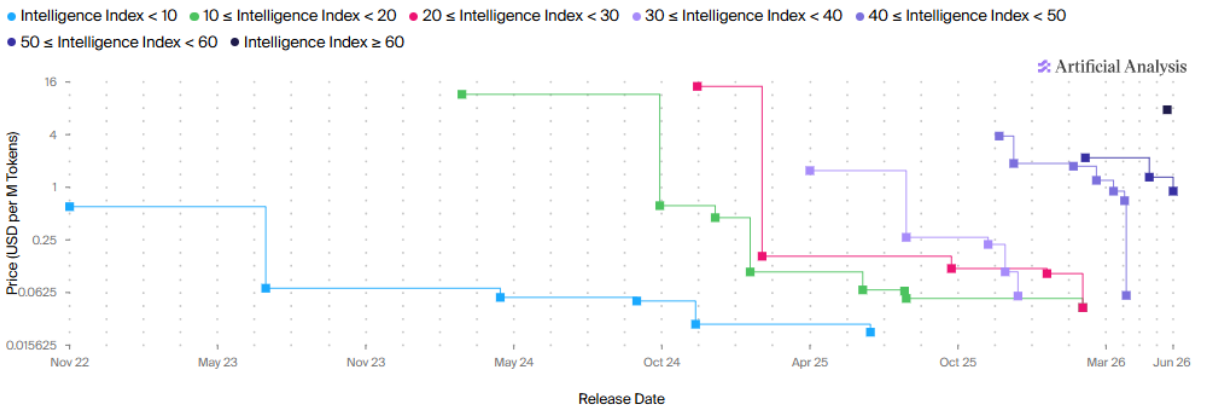
최적화 기법 지속, 확산될 경우 GPM 개선 기대. OpenAI 1Q26 GPM 39%, 2026 연말 목표 52%

Anthropic 의 동 개념 지칭 용어는 'Compute multipliers', 학습, 추론을 더 효율적으로 만드는 아키텍처 혁신들을 가르킴

적은 인프라를 지녀도 실제적은 인프라 양(Effective Compute)이 늘어난다고하여 칭해진 이름

Anthropic 은 복제 시 경쟁사 이점 제공 우려로 인지 인원 제한

지능 수준 별 LLM 추론 비용 추이: 최선단 모델의 추론 비용 유지, 차선단부터 급격한 하락



자료: Artificial Analysis, SK 증권

Compute Multiplier 는 어떤게 있을까?

기법이 비공개인 미국 AI Labs 과 달리 DeepSeek 등 중국 기업들은 논문을 통해 SW 최적화 기법을 많은 부분 공개

가장 대표적인 Multiplier 는 MoE(Mixture of Experts)로 모델 전체를 작동 시키지 않고 모델을 구획 후 필요 전문가만 호출

MLA(Multi-head Latent Attention)는 KV Cache 값을 저차원으로 저장하여 메모리 비용 축소

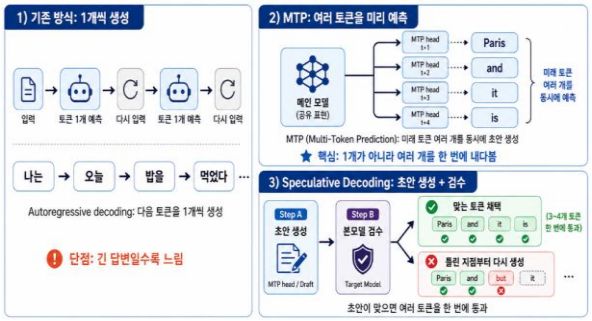
MTP(Multi-token Prediction)+speculative decoding 는 확률적 접근으로 Latency 축소

2026 년 6 월 DeepSeek 는 신규 V4 모델에 spec decoding 방식인 'Dspark' 적용하여 모델 속도 85% 향상 (Latency 축소)

중국 모델사들이 수출 통제에 의한 HW 제약을 Compute Multiplier 로 극복하려는 시도가 많은 것

SW 강점을 통해 미국 모델 대비 낮은 수준의 추론 비용을 제공, 낮은 성능 모델의 추론 비용을 더 낮추는 역할

기존 방식과 MTP + Speculative decoding 비교



자료: SK 증권

2026 년 6 월 Dspark 적용하여 동일 모델 속도 향상

자료: 언론 종합, SK 증권

OpenAI vs Anthropic, 누가 더 잘 최적화하나?

두 기업간 효율 비교를 Revenue per GW 로 적용할 경우 적은 비용으로 높은 ARR 달성 중인 Anthropic 우위

마진율 비교 또한 1Q26 의 추론 마진율(GPM)이 Anthropic 은 70% 상회, OpenAI 는 39% 로 보도되는 중

Compute multiplier 외 GPM 에 영향을 주었을 요인으로는 다음 사항들이 존재

1) 제품 믹스: Anthropic 은 마진율이 높은 Coding Agent 에 집중, 유료 티어 고객 집중. 반면 OpenAI 는 아직 챗봇 사업 이용자가 많은 상황이며 이 중 무료 이용자 비중이 90% 이상

2) 가격: Anthropic 의 '최고 성능 모델' 지위는 높은 토큰 가격을 지불 용의 형성. 낮은 C 보다 높은 P 로 인한 마진율

3) 칩 전력비: Anthropic 은 현재 ASIC 비중이 90% 이상 (TPU, Trainium), 엔비디아는 GPU 비중 90%. 엔비디아의 주장에 따르면 자사 GPU 시스템 활용 시 동일 전력에서 더 많은 토큰 생산 가능. 두 기업 모두 자사칩 준비 단계. 양산 타임라인은 OpenAI 의 할라피뇨(Jalapeño) 추론 칩이 2026 년 말 배포 시작 예정

Meta 서버 사업의 출시 파동, 맥락과 결론은?

Meta 는 GPU 를 어디에 활용하고 있을까?

Meta 의 GPU 는 1) AI Labs 로서 Muse Spark 등 서비스의 훈련 및 게시 2) 본업 광고 강화 에 나누어 활용

DAU 30 억명 이상인 플랫폼들에서 GPU 를 활용하여 컨텐츠 전환율, 광고 전환율을 높이는 방식으로 매출액 성장 중

최근 실적발표에서도 신규 추천 모델인 GEM 의 훈련 GPU 2 배 늘려 광고 전환율을 Instagram +5%·Facebook +3% 기록

현재 업계의 평균 광고 전환 성공률은 1000 번 중 10 번 내외 수준. AI 경쟁으로 지속적인 GPU 수요 전망

Muse Spark 은 4 월 공개된 Foundation Model 로, 이를 기반으로한 소비자용 Agent Service 기대 중

Agent Service 가 나올 경우 30 억명 이상의 사용자들에게 이를 즉시 배포할 예정. 사용 시간 증가로 이어진다면 매출 증폭 가능

따라서 Meta 는 현재 Compute 사용량은 적지만 제품이 나온다면 배포될 유통망은 확실한, 극단적인 Compute 수요 구조

더 이상 명분 없는 Capex 가 힘들어진 때

Meta 의 12MF FWD PER 는 17 배로 S&P 22 배, 빅테크 평균 25 배 수준 대비 강한 할인. 주주들의 강한 불만

Capex 통해 지은 AI DC 를 네오클라우드화한다고 가정할 경우 본업+AI DC 를 SOTP 하여 주가 상향 가능하다는 보고서 존재

주가가 AI DC 를 '외부 판매'할 경우, 또는 AI DC 를 외부 판매하는 '청산 가치' 대비해서도 할인돼있는 상황

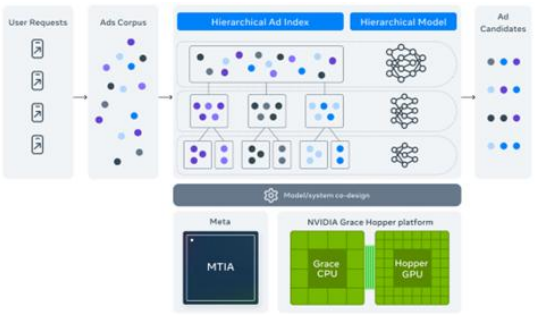
투자자들은 2020 년 이후 $80B 이상 투자했으나 $2B 수준의 제한적인 매출을 만들고 있는 Reality Labs 사업부 공포 기억

그러던 중 SpaceX 는 Anthropic, Google 에 비싼 가격에 Compute 를 임대하면서 3 개월 내에 통보 해지 조항 포함

이는 공격적 증설을 하면서 남는 Compute 를 판매, 이후 필요해질 경우 회수까지 가능한 계약 체결

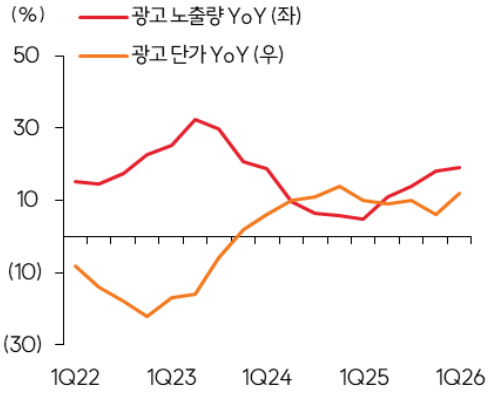
xAI 도 Meta 와 마찬가지로 미래의 꿈은 크나 현재 수요는 제한적인 기업, Meta 가 유사한 구조를 충분히 모방할 수 있는 상황

GPU 활용하는 Andromeda 광고 엔진 구조



자료: Meta, SK 증권

Meta 광고 단가 및 노출량 추이



자료: Meta, SK 증권

CSP 를 통해 전체 AI Capex 중 3% 내외 AI 인프라 추가 출회 가능

7/1 보도 이후 시장의 반응은 Meta 7% 상승, SOX 6% 하락, 네오클라우드 주식 10% 이상 하락으로 극단적으로 갈리는 모습

이는 Meta 가 그럴듯한 소비자용 Agent 를 발견하지 못한다고 보고 꾸준히 시장에 Compute 를 내놓는다는 가정

Meta 는 2026F 주요 AI Capex 의 17%, 2027F 주요 AI Capex 의 14% 차지

확보 Compute 의 20% 외부 판매 가정 시 주요 AI Capex 에서 3% 내외 AI 인프라 추가 출회 가능

그러나 외부 판매를 계기로 오히려 공격적인 AI Capex 집행 가능성도 염두에 두어야

기존에도 Meta 는 사용처를 찾지 못한 AI DC 구매로 많은 할인을 받는 상태, 언젠가는 해소했어야 하는 GPU 수요처

'남으면 판매하면 된다'는 명분이 생길 경우 저커버그는 텐트형 AI DC 를 더 공격적으로 증설할 가능성이 높음

안정적 B200, H100 대여 가격이 보여주는 결론: 주식 시장의 우려 선반영

다만 공격적으로 Capex 집행할 경우 네오클라우드 업체들의 우려가 불거질 수 있다고 판단

현재 Coreweave Backlog 의 1/3 이 Meta, 주요 고객이 경쟁사로 둔갑하는 셈

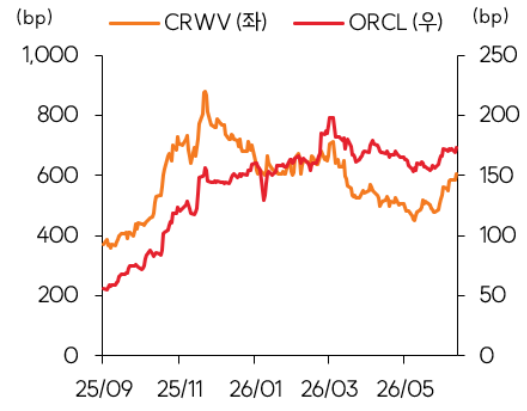
한동안 안정적이었던 Oracle, Coreweave 의 CDS spread 가 다시 올라올 수 있는 계기

아직까지는 B200, H100 대여 가격 견고, 산업계에는 과잉 공급 우려 제한적으로 보는 것으로 판단

결국 해당 우려가 주식 시장에 선반영된 상황, 실제로 GPU 대여 가격에 우려가 전가되는지 주목 필요

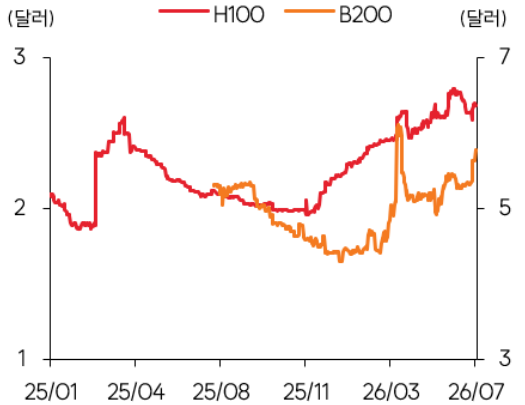
1) Meta 의 소비자형 Agent 출시 성공으로 자체 소화 2) AI Labs 업체 수요가 추가 공급 상쇄

Oracle, Coreweave CDS Spread



자료: Bloomberg, SK 증권

B200, H100 렌탈 가격 추이



자료: Bloomberg, SK 증권

Nvidia Update: Rubin Ultra 연기, CSP 사업 진출

Rubin Ultra 지연설: 반복되는 차세대 플랫폼 노이즈

(6/30) SemiAnalysis, Rubin Ultra GPU 원설계 quad-die(4-reticle) 취소·dual-die(2-reticle) 축소 확인 보도

(7/5) SemiAnalysis, Rubin Ultra 탑재용 Kyber NVL144 랙 2028 년으로 1 년 이상 지연됐다는 보도

(7/6) 엔비디아 공식 반박 "Our roadmap is intact"(AI 칩 로드맵 유지) 입장. 기존 로드맵은 2027 년 내 Ultra 출시

엔비디아는 2025 년 Blackwell 출시 당시 design flaw 논란을 인정한 이후에도 지연 없이 Blackwell 공급에 성공

GB200 NVL 72 Rack 의 경우 액체냉각 누수 이슈로 1 개 분기 랙 출시가 실질적으로 지연

Vera Rubin 도 Redesign 이슈가 있었으나 엔비디아 부인 후 타임라인에 맞추어 출시. 즉 엔비디아가 부인한 지연은 여태 없었음

CSP 사업 진출: 신용 지원을 통한 자금 조달 지원

엔비디아, AI 클라우드 사업자 대상 신규 비즈니스 모델 "revenue-sharing + credit-support" 도입 발표

네오클라우드에 최소매출을 보장(6 년 take-or-pay)하고, 백스톱 초과분 매출을 일부 나눠 갖는 구조

엔비디아는 1) 칩 판매 2) 지원 용량(supported capacity) 클라우드 매출의 일부(revenue share) 이중 수취

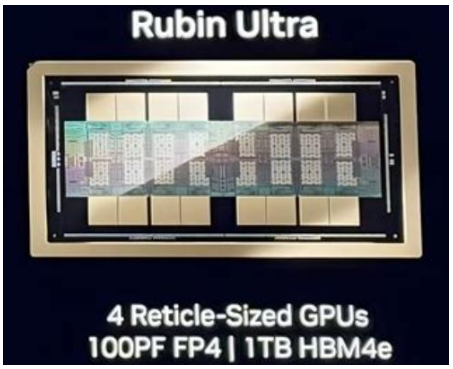
클라우드 파트너 핵심 수혜는 신용 확보를 통한 자금 조달 용이. 파트너가 GPU 임차 수요처 확보 실패 시 엔비디아가 유휴 GPU 용량을 사전 합의 가격에 재매입

엔비디아 측 효과: 하드웨어 일회성 매출 위에 usage-linked·recurring 매출 스트림 추가

초기 계약자는 Sharon AI, Firmus 로 아직 500MW 내외 규모. 소버린 및 엔터프라이즈향 위주 서버가 타겟일 것으로 판단

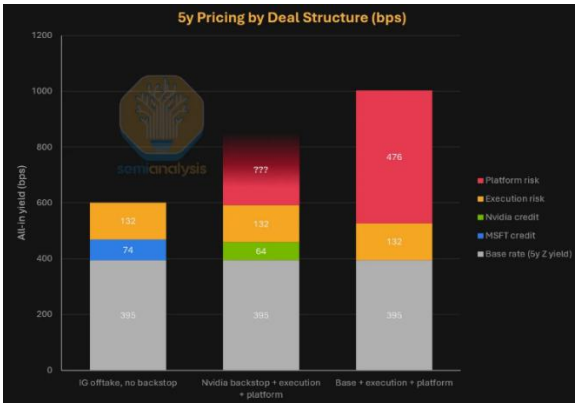
AI Infra 사업자들의 신용 지원은 사업자를 막론하고 지속 중(Google, Broadcom). 전방 수요에 대한 의문 제기 가능성 증가

Rubin Ultra 의 quad die 구조



자료: Nvidia, SK 증권

Nvidia Backstop 구조로 네오클라우드 금리 이점 수취 가능



자료: Semianalysis, SK 증권

Compliance Notice

작성자(박제민)는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.

당사는 자료공표일 현재 해당기업과 관련하여 특별한 이해 관계가 없습니다.