SECTOR
UPDATE

2026. 3. 17

Tech팀

이종욱 팀장
jwstar.lee@samsung.com

문준호 Senior Analyst
joonho.moon@samsung.com

정승호 Research Associate
seungho1.jeong@samsung.com

리서치센터 리포트
바로가기

Tech (OVERWEIGHT)

GTC 2026: Disaggregated Inference

- GTC 2026의 주인공은 단연 Groq 3 LPU로, SRAM 기반 솔루션의 도입
- LPU-SRAM은 GPU-HBM을 대체하기 보다는 보완하는 것으로 이해할 필요
- AI가 유발한 메모리 계층 구조의 심화 속, 스펙과 함께 신규 솔루션의 중요성 확대

WHAT'S THE STORY?

Groq 3 LPU 마침내 출시: 시장 예상대로 Groq 3 LPU (Language Processing Unit) 기반 솔루션이 공개되었다. HBM이 아닌 SRAM (500MB)을 메모리로 활용하여, 보다 더 빠른 연산이 가능하다는 장점이 있다. Groq 3 LPU 256개(SRAM 128GB)를 탑재한 서버 랙 LPX도 함께 공개되었으며, 3분기부터 출하 개시될 예정이다.

Rubin + LPX 구조 도입의 배경: LPX의 등장은 기존 단일 GPU 아키텍처에서 처리하던 추론 과정을 Prefill과 Decode로 세분화하고, 각각에 최적화된 GPU-HBM과 LPU-SRAM을 조합하는 이종(disaggregated) 구조로의 전환을 의미한다. 추론 과정에서도 연산량이 크고 병렬 처리가 용이한 Prefill 구간에서는 GPU-HBM 조합이 활용되는 반면, 병렬성이 낮고 KV cache 접근 비중이 높은 Decode 구간 중 일부에서는 지연시간(latency)이 낮은 LPU-SRAM 조합이 적용된다. 이러한 계층(Hierarchy) 구조의 확장은 단일 디바이스 성능이 아니라 서버 전체 추론 경로(end-to-end inference path)를 최적화하는 관점에서 이해할 필요가 있다.

SRAM이 기존 메모리 수요를 잠식할까? SRAM 계층의 등장은 기존 메모리 수요를 대체하기보다 전체 메모리 수요를 확장하는 방향이라고 해석한다. Rubin NVL을 대체하는 것이 아닌 LPX 서버가 추가되어, 추론 성능 향상을 꾀한다. 즉, HBM 용량 수요가 줄지 않은 채 오히려 전체 메모리 footprint를 확대시키는 방향으로 작용한다.

Memory-centric AI의 증거: 결국 NVIDIA의 CMX(NAND)와 LPX(SRAM)는 메모리 계층 구조의 심화를 가속화하는 변화로 이해할 필요가 있다. 이는 HBM의 중심적 역할이 유지되는 가운데 DRAM, NAND, SRAM의 추가 수요를 자극하는 memory-centric AI 구조의 강화로 해석된다. 서버용 메모리는 이제 한두개의 단일 제품을 잘 만드는 것으로는 경쟁력을 유지하기 힘들고, AI 방향성을 메모리에 접목하여 새로운 스펙과 솔루션을 제안할 수 있는 기업의 시너지가 강화될 것이다.

삼성 파운드리의 등장: Groq의 LPU (LP30)은 삼성전자 파운드리 4nm 공정에서 생산하는 것으로 알려져 있다. 올해와 내년 각각 15k, 30k의 웨이퍼가 소요될 것이라 추정한다. 단일 GPU의 등장에서 복수의 프로세서가 복수의 고객사로부터 등장하는 방향성은 TSMC에 집중되어 있는 파운드리 수요를 분산시키는 효과가 있을 것이다. 기존 수주 받은 테슬라 이외에 추가적인 ASIC 수주가 기대된다.

(다음 페이지에 계속)

GTC 2026 Key announcements

2025~2027년 매출액 1조 달러

작년 10월 GTC에서 2025~2026년 예상 매출 5천억 달러를 제시한 데 이어, 이번 GTC에서는 2025~2027년 예상 매출 1조 달러를 제시했다. 2025~2027년 Data Center 부문 매출액 컨센서스(Bloomberg)가 9,436억 달러에 형성되어 있음을 감안하면, 큰 서프라이즈가 아닐 수 있다.

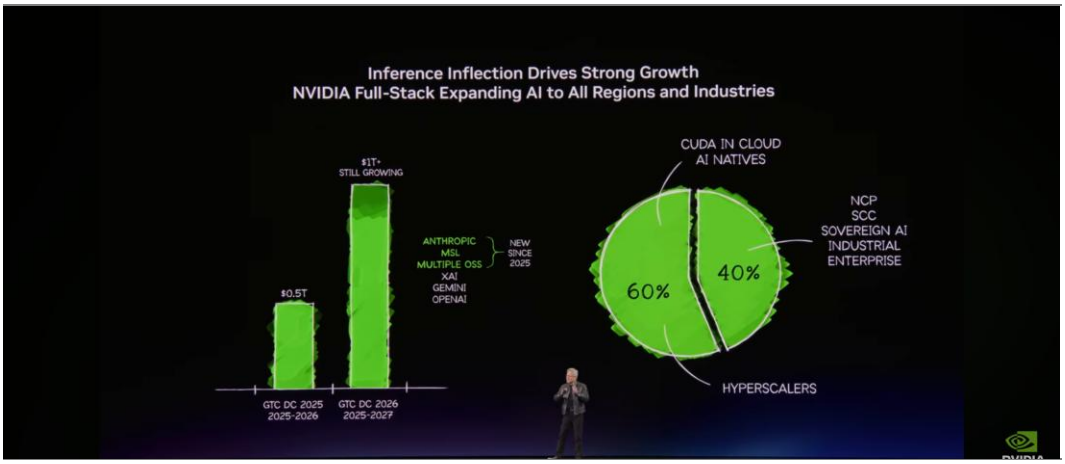
그러나 해당 수치는 Blackwell과 Rubin 계열 플랫폼에 한정된다. 즉, 여전히 판매되고 있는 Hopper 계열 플랫폼이나 CPX와 LPX 등 새로이 출시될 솔루션들 관련 매출 기회가 반영되지 않은 수치다. 이를 고려하면, 동사의 컨센서스는 여전히 상향 여력이 존재한다고 해석할 수 있다.

Groq의 등장과 추론의 분산 (feat. SRAM)

엔비디아는 작년 말 추론 특화 반도체 LPU (Language Processing Unit) 스타트업 Groq을 200억 달러에 우회인수한 바 있는데, 역시나 이번 GTC에서 Groq 3 LPU 기반의 솔루션들이 공개되었다. HBM이 아닌 SRAM (500MB)을 메모리로 활용하여, 보다 더 빠른 연산이 가능하다는 장점이 있다. 즉, 저지연 (low latency) 작업에 있어서는 오히려 GPU 보다 더 효율적일 수 있는 것이다. Groq 3 LPU 256개(SRAM 128GB)를 탑재한 서버 랙 LPX도 함께 공개되었으며, 3분기부터 출하 개시될 예정이다.

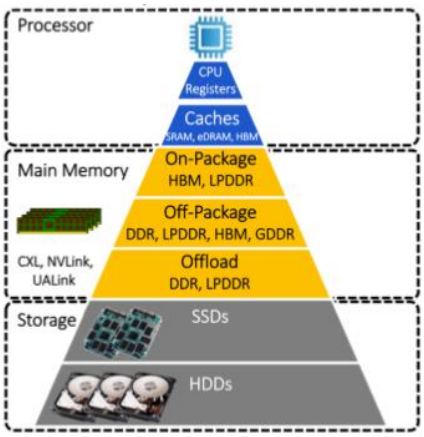
이를 도입하는 목적은 추론의 분산 (Disaggregated Inference)이다. 기존 단일 GPU 아키텍처에서 처리하던 추론 과정을 Prefill과 Decode로 세분화하고, 각각에 최적화된 GPU-HBM과 LPU-SRAM을 조합하는 구조다. 추론 과정에서도 연산량이 크고 병렬 처리가 용이한 Prefill 구간에서는 GPU-HBM 조합이 활용되는 반면, 병렬성이 낮고 KV cache 접근 비중이 높은 Decode 구간 중 일부에서는 지연시간(latency)이 낮은 LPU-SRAM 조합이 적용된다. 이러한 계층(Hierarchy) 구조의 확장은 단일 디바이스 성능이 아니라 서버 전체 추론 경로(end-to-end inference path)를 최적화하는 관점에서 이해할 필요가 있다. 젠슨 황에 따르면, Rubin 시스템의 성능 (throughput)은 이전작 Blackwell 대비 10배 증가하는데, Groq LPX 시스템을 함께 도입할 경우, 성능이 최대 35배 증가한다.

2025~2027년 매출 누계 1조달러 제시



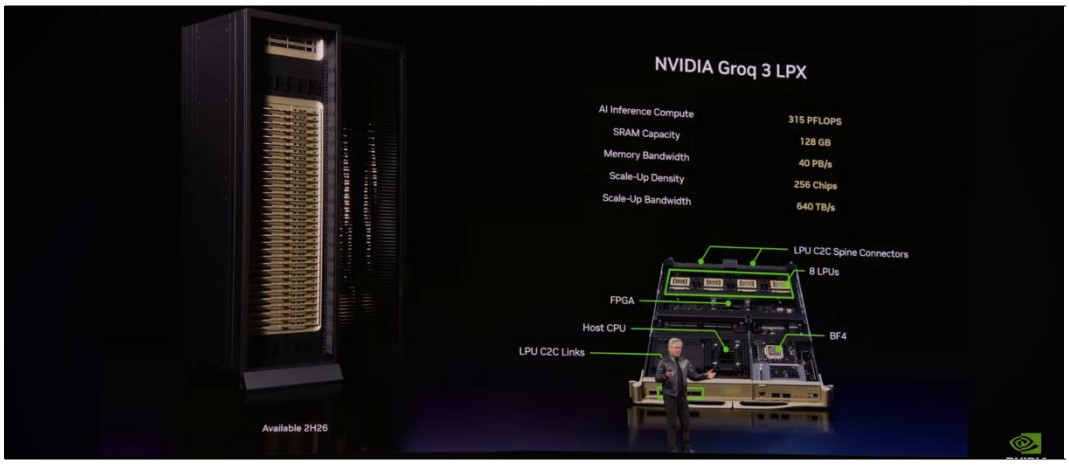
자료: NVIDIA, 삼성증권

메모리 계층 구조도



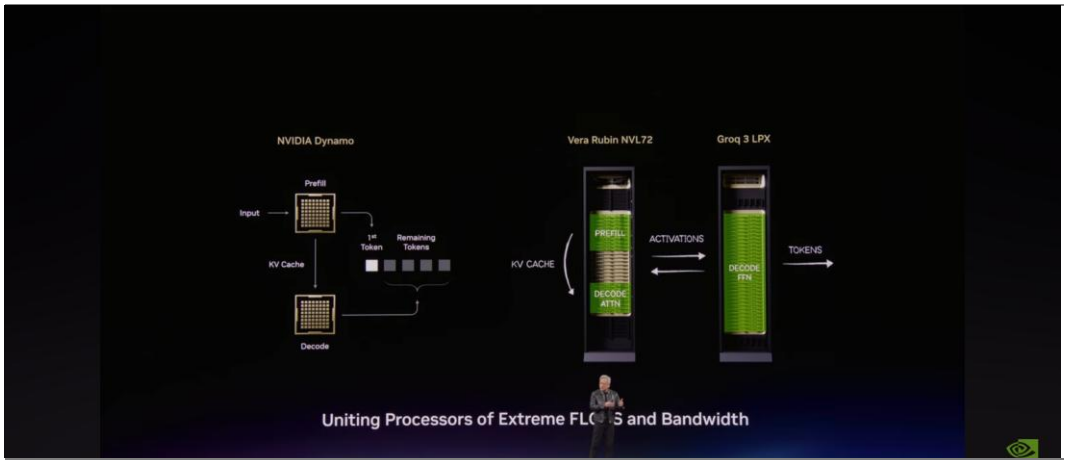
자료: Semiconductor Engineering, 삼성증권

Groq 3 LPX 트레이 공개



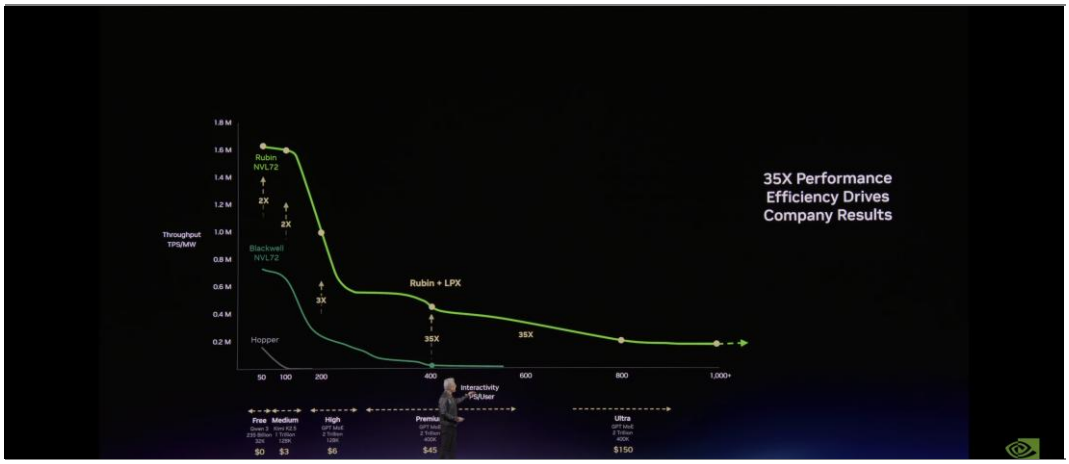
자료: NVIDIA, 삼성증권

Vera Rubin과 Groq LPX 간 역할 분담



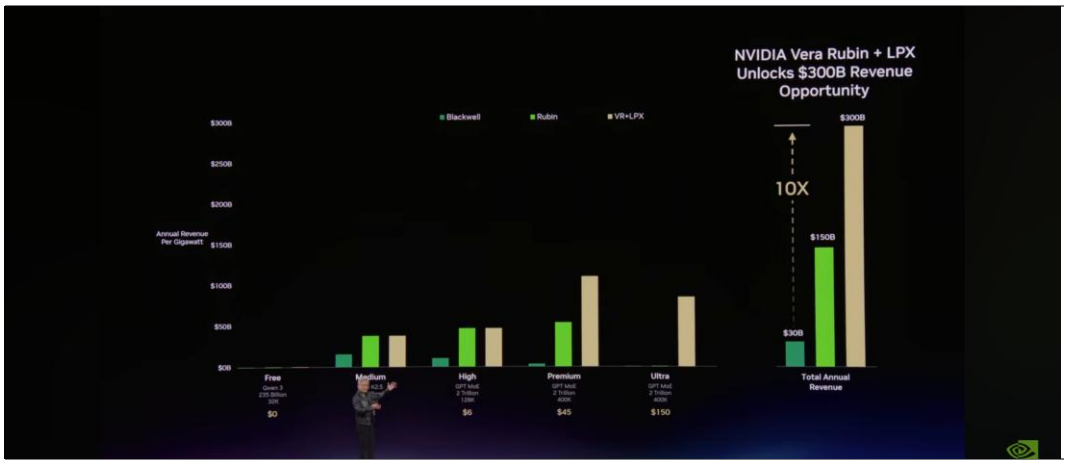
자료: NVIDIA, 삼성증권

Rubin 추론 성능 (throughput)은 Blackwell 대비 10배 증가하고, LPX 도입 시 최대 35배 증가



자료: NVIDIA, 삼성증권

LPX 도입 시 기가와트당 매출 기회는 10배 증가



자료: NVIDIA, 삼성증권

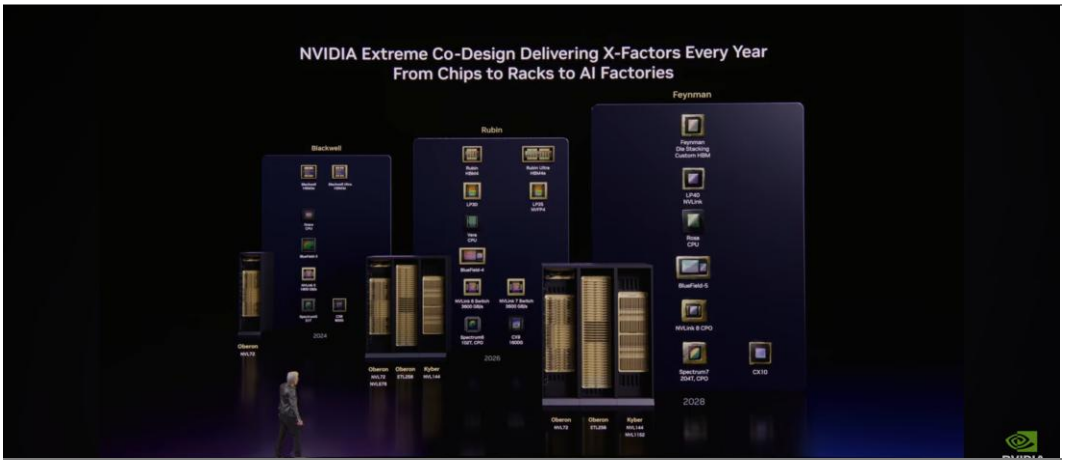
Rubin을 넘어, Feynman으로

차세대 아키텍처 Feynman (2028년)에 대한 업데이트도 있었다. TSMC 2nm가 아닌 1.6nm 공정을 채택하고, HBM 5 탑재 예정이다. Feynman부터는 Rosa라는 새로운 CPU가 함께 번들 된다. 서버 랙 솔루션 Kyber의 예상 소모 전력은 600kW에 이를 것으로 전망되었는데, 이는 현 GB300 NVL72 랙의 4배에 달하는 수준이다. 시장 관심을 모았던 CPO(Co-Packaged Optics) 스위치도 Rubin부터 도입 가능하나, 본격적으로 도입되기 시작하는 시점은 Feynman부터가 될 전망이다.

NemoClaw를 통해 AI 에이전트 도입 가속화

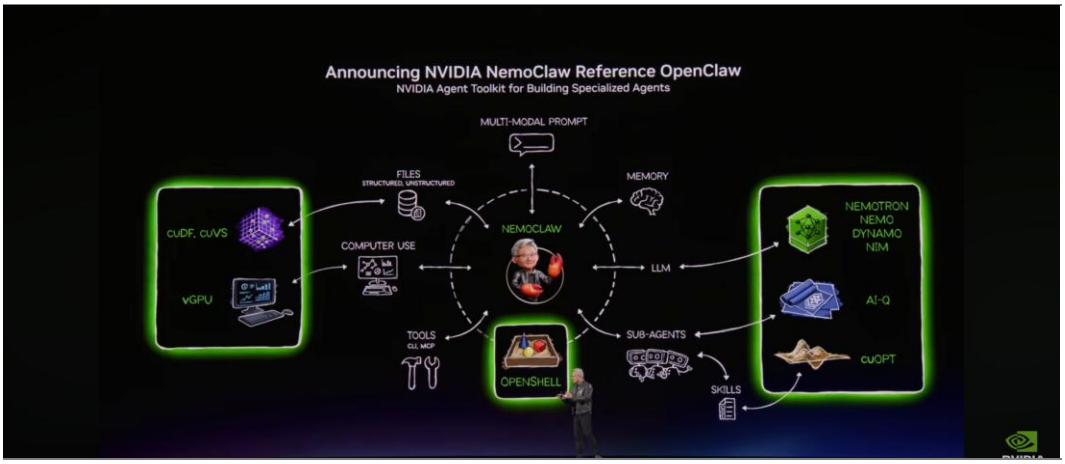
이번 GTC의 소프트웨어 하이라이트는 NemoClaw였다. 엔터프라이즈용 오픈소스 AI 에이전트 플랫폼이다. 기존의 오픈소스 에이전트 OpenClaw에 프라이버시 정책과 보안 가드레일 기능을 추가하였다. 기업들이 안전하고, 빠르게 AI 에이전트를 구축할 수 있도록 돕기 위함인데, 이를 통해 기업들의 컴퓨팅 파워 수요를 더욱 자극하는 일종의 선순환 구조가 기대된다.

하드웨어 로드맵 업데이트



자료: NVIDIA, 삼성증권

NemoClaw: 오픈소스 AI 에이전트 OpenClaw를 보완할 플랫폼

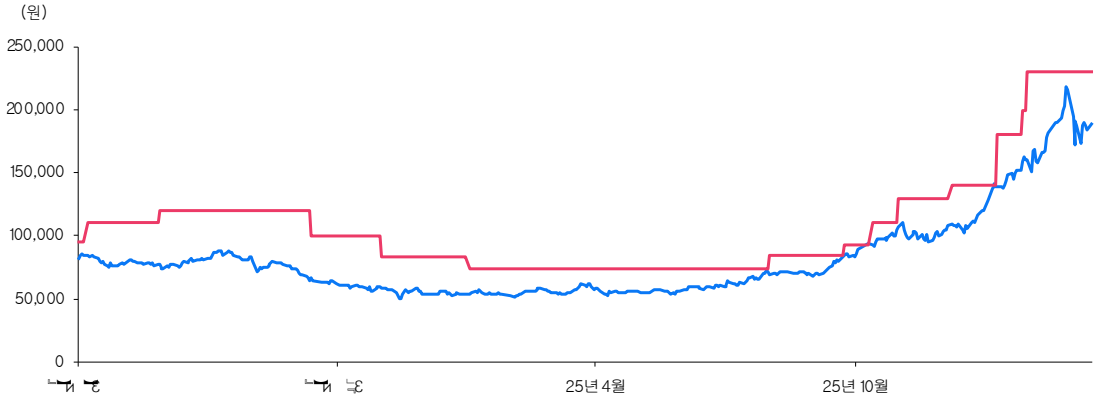


자료: NVIDIA, 삼성증권

Compliance notice

- 당사는 2026년 3월 16일 현재 삼성전자와(과) 계열사 관계에 있습니다.
- 본 조사분석자료의 애널리스트는 2026년 3월 16일 현재 위 조사분석자료에 언급된 종목의 지분을 보유하고 있지 않습니다.
- 당사는 2026년 3월 16일 현재 위 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 조사분석자료에는 외부의 부당한 압력이나 간섭 없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.
- 본 조사분석자료는 당사의 저작물로서 모든 저작권은 당사에게 있습니다.
- 본 조사분석자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다.
- 본 조사분석자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다.
- 본 조사분석자료는 기관투자가 등 제3자에게 사전 제공된 사실이 없습니다.

2년간 목표주가 변경 추이



최근 2년간 투자의견 및 목표주가 변경 (수정주가 기준)

삼성전자												
일 자	2024/1/2	4/8	5/28	9/12	11/1	2025/1/2	8/1	9/23	10/13	10/31	12/8	2026/1/9
투자의견	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY
TP (원)	95000	110000	120000	100000	83000	74000	85000	93000	110000	130000	140000	180000
괴리율 (평균)	-20.74	-28.36	-34.45	-39.49	-33.92	-21.93	-15.05	-6.95	-10.64	-21.96	-16.54	-18.69
괴리율 (최대or최소)	-10.21	-23.55	-26.83	-35.30	-29.28	-1.89	-1.76	1.51	-5.36	-14.54	0.71	-15.39
일 자	1/27	1/30										
투자의견	BUY	BUY										
TP (원)	200000	230000										
괴리율 (평균)	-19.23											
괴리율 (최대or최소)	-18.80											

투자기간 및 투자등급: 삼성증권은 기업 및 산업에 대한 투자등급을 아래와 같이 구분합니다.
*2023년 7월 27일부로 기업 투자 등급 기준 변경

기업

BUY (매수)	향후 12개월간 예상 절대수익률 15% 이상 그리고 업종 내 상대매력도가 평균 대비 높은 수준
HOLD (중립)	향후 12개월간 예상 절대수익률 -15%~ 15% 내외
SELL (매도)	향후 12개월간 예상 절대수익률 -15% 이하

산업

OVERWEIGHT (비중확대)	향후 12개월간 업종지수상승률이 시장수익률 대비 5% 이상 상승 예상
NEUTRAL (중립)	향후 12개월간 업종지수상승률이 시장수익률과 유사한 수준 (±5%) 예상
UNDERWEIGHT (비중축소)	향후 12개월간 업종지수상승률이 시장수익률 대비 5% 이상 하락 예상

최근 1년간 조사분석자료의 투자등급 비율 2025.12.31 기준
매수(85.8%)·중립(14.2%)·매도(0%)

삼성증권

삼성증권주식회사
서울특별시 서초구 서초대로74길 11(삼성전자빌딩)
Tel: 02 2020 8000 / www.samsungpop.com

삼성증권 Family Center: 1588 2323

고객 불편사항 접수: 080 911 0900

NCSI KS-SQI KCSI KCPI 소비자중심 Best Korea Brands 2024
Member of Dow Jones Sustainability Indices Powered by the S&P Global CSA