미국은 지금

Google I/O 2026:
Agentic AI 시대의 개막

키움증권 리서치센터 글로벌리서치팀
US Equity 박기현 kihyun.park@kiwoom.com

Issue Brief

Google I/O 를 통해 확인한 AI 투자 사이클의 지속성

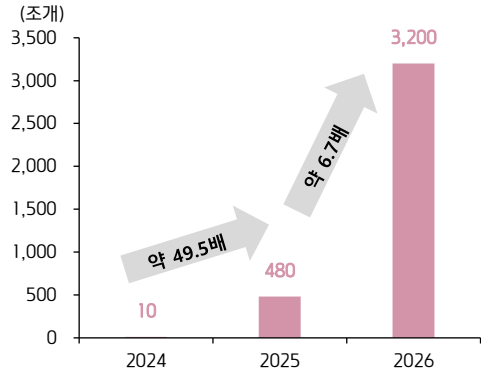
Google I/O 는 구글이 매년 개최하는 연례 개발자 컨퍼런스로, 차세대 소프트웨어 아키텍처, 핵심 하드웨어 로드맵, 오픈소스 프레임워크 및 개발자 도구의 표준을 선제적으로 제시하는 상징적인 이벤트이다. 과거 단발성 챗봇 형식의 대화형 생성 AI 가 출현했던 초기 국면에서 구글은 OpenAI, Anthropic 등 프론티어 모델 전문기업 진영에 시장 주도권을 일부 상실했다는 금융시장의 회의적 평가를 받기도 하였다.

그러나 이번 Google I/O 2026(5/19~20)에서 동사는 독보적인 가격 경쟁력과 압도적인 처리 속도를 갖춘 차세대 자체 반도체(TPU v8) 아키텍처, 프론티어급 성능의 초고속 경량화 모델(Gemini 3.5 Flash), 그리고 자율형 에이전트 소프트웨어 오케스트레이션 생태계를 전격 공개하였으며, 이를 통해 구글은 개별 모델 단위의 일대일 성능 경쟁을 넘어 '실리콘-인프라-모델-플랫폼' 전반을 관통하는 독점적 밸류체인의 보유자로 완전히 거듭났음을 입증하였다.

금융시장 관점에서 Google I/O 는 단순한 신제품 쇼케이스를 넘어 하이퍼스케일러의 CapEx 집행 강도와 그에 따른 전방 공급망의 매출 지속성을 정량적으로 추정할 수 있는 지표이다. 실제로 이번 I/O 에서 구글이 공개한 월간 토큰 처리량은 YoY 약 7 배('25.5 월 480T → '26.5 월 3,200T+), 연간 CapEx 는 2025 년 910 억 달러에서 2026 년 약 1,900 억 달러(약 2 배)로 동반 확장되었음이 확인되었다.

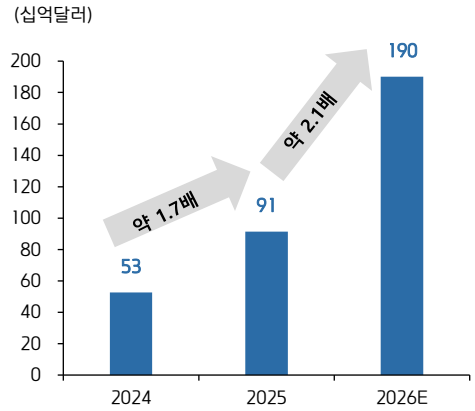
특히 이번 I/O 2026 은 단순한 프롬프트 명령어 시대의 종언을 고하고, 백그라운드 단에서 스스로 연산하고 판단하는 자율형 멀티 에이전트(Agentic AI) 시대의 개막을 선언하였다. Agentic AI 도입이 중요한 이유는 글로벌 데이터센터 워크로드 프로필을 '간헐적이고 가변적인 트래픽(Burst 형)'에서 '지속적이고 고정적인 기저 부하(Persistent Base-load 형)' 구조로 근본적으로 체질 개선시킬 것이기 때문이다. 결과적으로 차세대 연산 가속기 및 네트워킹 하드웨어 수요의 수명 주기가 시장의 우려보다 훨씬 더 길고 강하게 유지될 것임을 예고하는 중대한 분수령이다.

Google 월간 토큰 처리량



자료: 구글, 키움증권 리서치

Google 연간 CapEx 추이



자료: Bloomberg, 구글, 키움증권 리서치

구글 I/O 2026 핵심 발표 사안: Agentic AI 시대의 도래

금번 행사에서 구글이 공개한 기술 혁신의 핵심은 하드웨어, 기초 모델, 그리고 이를 결합한 에이전트 실행 플랫폼의 '풀스택(Full-stack) 수직 계열화'로 요약된다. 주요 발표 사안은 다음과 같다.

1) 8 세대 반도체(TPU 8t 및 TPU 8i) 도입

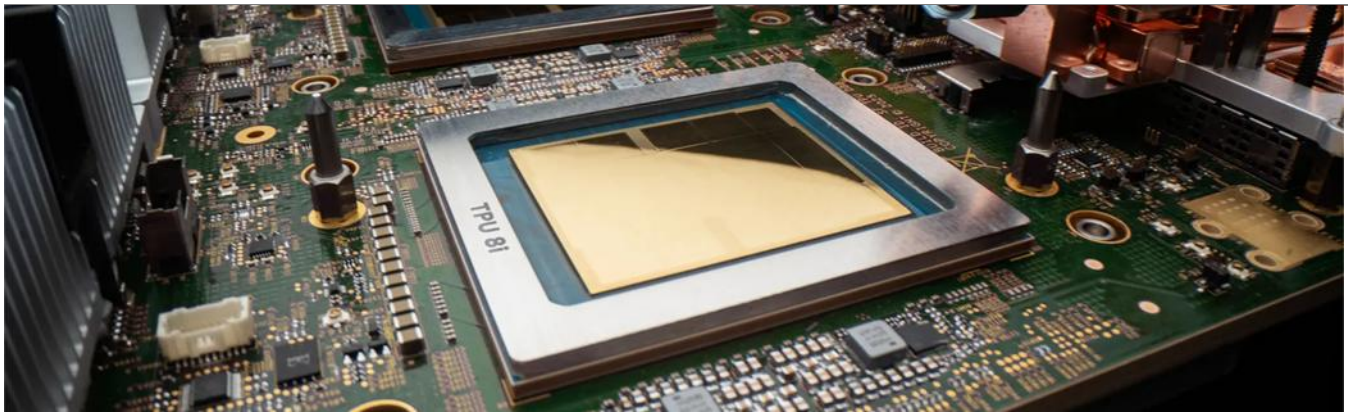
구글은 독자 설계 맞춤형 ASIC 라인업인 8 세대 텐서 처리 장치(TPU)를 최초로 학습 전용인 TPU 8t 와 추론 전용인 TPU 8i 의 듀얼 칩 구조로 이원화하여 발표했다.

학습 특화형 TPU 8t 는 전 세대 대비 원시 컴퓨팅 성능이 3 배 향상되었으며, 9,600 칩 슈퍼포드 및 2PB 공유 HBM 구성을 지원한다. 또한 JAX 및 Pathways 분산 컴파일러 프로토콜을 통해 단일 데이터센터의 한계를 극복하고 글로벌 100 만 개 이상의 TPU 규모로 연산 풀을 유기적으로 확장할 수 있는 스케일아웃 역량을 갖추었다.

반면, 추론 전용 TPU 8i 는 1,152 칩 클러스터 구성에 온칩 SRAM 을 3 배 증대시켜, 다중 자율 에이전트 구동 시 발생하는 지연 시간(Latency)을 단축하는 데 모든 리소스를 집중하도록 전면 재설계되었다.

두 칩셋 모두 와트당 성능(Performance-per-watt) 효율을 기존 대비 최대 2 배 개선하여, 랙당 전력 부하 압력을 완화하도록 설계된 점이 특징이다.

구글 8 세대 TPU 제품군(TPU 8t / 8i)



주요 평가 항목	TPU 8t (학습)	TPU 8i (추론)
주요 목적	대규모 모델 학습 최적화	저지연 추론 및 멀티 에이전트 구동 최적화
연산 성능 및 가치	포드(Pod)당 연산 성능 약 3 배 향상	이전 세대 대비 달러당 성능 80% 개선
전력 효율성	와트당 성능(Performance-per-watt) 최대 2 배 향상	와트당 성능(Performance-per-watt) 최대 2 배 향상
메모리 아키텍처	단일 슈퍼포드당 9,600 개 칩 및 2PB 공유 HBM 확장	288GB HBM + 384MB 온칩 SRAM (이전 대비 3 배)
인터커넥트 대역폭 (ICI)	인터칩 대역폭(ICI) 2 배 확장	ICI 대역폭 2 배 확장 (19.2 Tb/s)
CPU 호스트 환경	자체 Axion ARM 기반 CPU 최초 탑재	자체 Axion ARM 기반 CPU 탑재 (서버당 호스트 2 배 확충)
시스템 최적화 혁신	10x 빠른 스토리지 액세스 및 TPUDirect Virgo 네트워크 기반 100 만 칩 스케일링	Boardfly 아키텍처 (네트워크 직경 50% 축소) 온칩 가속 엔진(CAE) 도입으로 지연 5 배 단축

자료: 구글, 키움증권 리서치

2) Gemini 3.5 Flash 및 Gemini Omni

구글은 최첨단 인텔리전스 지표와 압도적인 실행 속도를 결합한 경량화 모델인 Gemini 3.5 Flash 를 출시했다. 해당 모델은 이전 세대 상위 모델인 Gemini 3.1 Pro 의 주요 벤치마크 및 실제 경제적 가치를 평가하는 GDPVal 성능을 상회하는 한편, 속도-비용 효율 측면에서 동급 프론티어 모델 대비 결정적 우위를 확보한 것으로 평가된다.

또한 입력과 출력 전 과정에서 동영상, 이미지, 텍스트를 동시에 처리하는 네이티브 멀티모달 라인업인 Gemini Omni 제품군을 유료 구독자 및 유튜브 쇼츠 생태계 전반에 제공하기 시작했다.

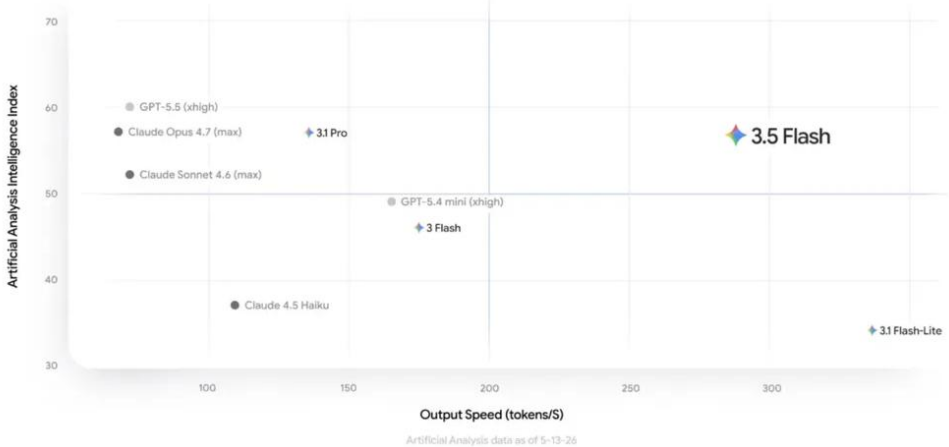
Gemini 3.5 Flash 대비 Anthropic(Opus 4.7) 및 OpenAI(GPT-5.5) 프론티어 모델 비교

	Benchmark		Gemini 3.5 Flash	Gemini 3 Flash	Gemini 3.1 Pro	Claude Sonnet 4.6	Claude Opus 4.7	GPT-5.5
Coding	Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	76.2%	58.0%	70.3%	-	66.1%	**78.2%**
	SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	55.1%	49.6%	54.2%	-	**64.3%**	58.6%
Agentic	MCP Atlas Multi-step workflows using MCP		**83.6%**	62.0%	78.2%	69.5%	79.1%	75.3%
	Toolathlon Real-world general tool use		**56.5%**	49.4%	-	-	-	55.6%
UI control	OSWorld-Verified Agentic computer use		78.4%	65.1%	76.2%	72.5%	78.0%	**78.7%**
Expert tasks	Finance Agent v2 Financial analysis and decision-making		**57.9%**	42.6%	43.0%	51.0%	51.5%	51.8%
	GDPval-AA Economically valuable knowledge work	Elo	1656	1204	1314	1676	1753	**1769**
Multimodal	CharXiv Reasoning Information synthesis from complex charts	No tools	**84.2%**	80.3%	83.3%	72.4%	82.1%	84.1%
	MMMU-Pro Multimodal understanding and reasoning	No tools	**83.6%**	81.2%	80.5%	74.5%	75.2%	81.2%
	Blueprint-Bench 2 Agentic spatial reasoning	Normalized score	33.6%	0.0%	26.5%	6.7%	24.5%	**36.2%**
Long context	MRCR v2 (8-needle) Long context performance	128k (average)	77.3%	67.2%	84.9%	84.9%	59.3%	**94.8%**
		1M (pointwise)	**26.6%**	22.1%	26.3%	-	-	-
Reasoning	Humanity's Last Exam Academic reasoning (full set, text + MM)		40.2%	33.7%	44.4%	33.2%	**46.9%**	41.4%
	ARC-AGI-2 Abstract reasoning puzzles		72.1%	33.6%	77.1%	58.3%	75.8%	**84.6%**

https://deepmind.google/models/evals-methodology/gemini-3-5-flash/

자료: 구글, 키움증권 리서치

Gemini 3.5 Flash 는 Artificial Intelligence Index 및 Output Speed 에서 모두 최상위권에 해당



자료: 구글, 키움증권 리서치

3) 자율형 에이전트 플랫폼(Antigravity 2.0 및 Gemini Spark) 상용화

단순 코딩 보조 도구를 넘어 복잡한 연속형 에이전트 그룹을 직접 개발 및 조율하는 독립형 데스크톱 플랫폼인 Antigravity 2.0 과 구글 클라우드 전용 가상머신(VM) 가동환경에서 24 시간 내내 백그라운드로 독립 실행되는 개인용 AI 비서 Gemini Spark 가 공개되었다.

Antigravity 2.0 의 시연에서는 93 개의 AI 서브 에이전트가 12 시간 동안 약 26 억 토큰을 생성하며 단일 OS 프레임워크를 구축하는 사례가 공개되었다. 이는 사용자의 능동적 세션 유지를 전제하지 않고, 사용자가 수면 중이거나 디바이스를 꺼놓은 상태에서도 장시간 소요되는 데이터 분석, 리서치 프로젝트, 자율 쇼핑 등의 연산 업무를 스스로 수행하는 환경을 구축해냈음을 시사한다.

Gemini Spark 는 MCP(Model Context Protocol) 연동을 기반으로 Canva, OpenTable, Instacart 등 글로벌 서드파티 파트너사들과의 연결 앱 생태계를 확장하고 있다. 이러한 외부 툴과의 유기적인 통합은 에이전트가 사용자를 대행해 복잡한 업무를 직접 처리할 수 있도록 자율 실행력을 강화하는 핵심 기반이 된다. 향후 맞춤형 하위 에이전트 생성 및 로컬 브라우저 제어 등의 신규 기능이 순차적으로 추가됨에 따라, 외부 생태계와 결합된 에이전트의 유효 워크로드 범위는 더욱 넓어질 전망이다.

Gemini Spark 는 글로벌 서드파티 파트너십 확보 및 MCP 연동을 통한 자율 실행력 강화를 추진 중

자료: 구글, 키움증권 리서치

금융시장 함의: AI 인프라 사이클 장기화

최근 시장에서는 'CapEx 피로(CapEx fatigue)' 우려, 즉 2H26 이후 AI 인프라 투자 사이클의 둔화에 대한 내러티브가 확산되고 있다. Gemini 3.5 Flash 와 같은 고효율 모델의 등장으로 토큰당 비용이 하락하면, 하이퍼스케일러의 AI 인프라 투자가 둔화될 것이라는 논리이다. 그러나 I/O 2026 에서 공개된 데이터는 이 논리와 정면으로 충돌한다. 프론티어 모델의 경량화 및 효율성 개선은 연산 수요를 감소시키는 것이 아니라, 한계비용 하락을 통해 총 토큰 소비량의 기하급수적 폭발을 촉발하는 '제번스의 역설(Jevons' Paradox)'을 유도한다. 따라서 AI 하드웨어 인프라 수요의 수명 주기는 시장의 예상보다 훨씬 길고 강력하게 지속될 것이다.

1) Agentic AI 가 유도하는 토큰 소비량의 지수적 증가

기존의 생성형 AI 가 인간의 입력에 의해서만 연산이 개시되는 1:1 단발성 채팅 구조였다면, Gemini Spark, Antigravity 2.0 과 같은 연속형 에이전트 시스템은 1:N 지속형 오케스트레이션(Always-on) 구조로의 전환을 의미한다.

실제로 구글 인프라 전반의 월간 토큰 처리량은 2024 년 5 월 9.7T, 2025 년 5 월 480T 에서 2026 년 5 월 현재 3,200T 이상으로 YoY 약 7 배 급증하는 성장 궤도를 보여주고 있다. Antigravity 시연에서는 93 개의 서브 에이전트가 단일 프로젝트를 수행하기 위해 12 시간 동안 생성한 토큰량만 26 억 개에 달했다.

토큰 소비량은 더 이상 인간의 물리적 타이핑 속도나 세션 길이에 종속되지 않으며, 인프라의 동시 처리량(Concurrent throughput)에 의해서만 제한되므로 총 연산 볼륨은 기하급수적으로 팽창할 수밖에 없다.

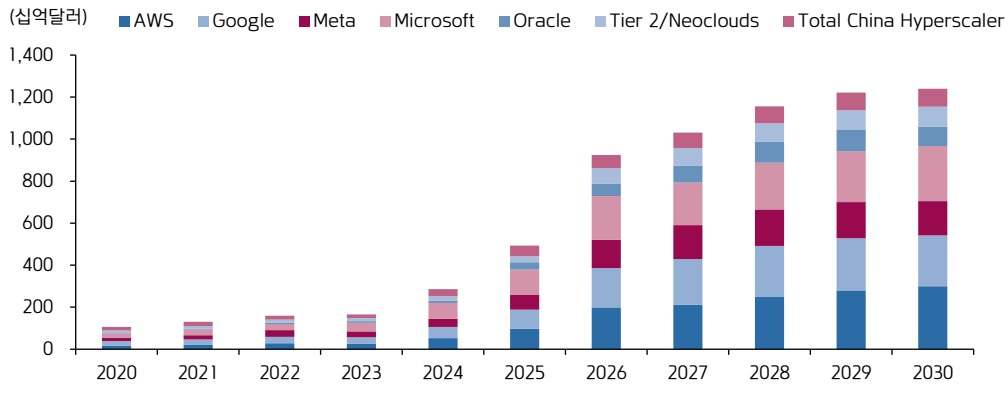
2) 하이퍼스케일러 CapEx 의 구조적 증가

폭발적으로 증가하는 토큰 소비량을 저지연으로 뒷받침하기 위해 구글의 연간 CapEx 규모는 2025 년 910 억 달러 수준에서 2026 년 예상 기준 약 1,900 억 달러로 약 2 배 상향 조정되었다. 여기서 핵심은 '총 토큰 처리량 증가율(7 배)'이 'CapEx 증가율(2 배)'을 명확히 상회한다는 점이다.

이는 토큰당 가격 하락 시 수요의 가격 탄력성이 3.5 배 수준(7 배 ÷ 2 배)으로, 1 을 대폭 상회하고 있음을 의미하며, 비용 하락 → 신규 워크로드 경제성 확보 → 수요 폭발 → 인프라 증설로 이어지는 양의 피드백 루프(positive feedback loop) 가 작동하는 구조이다.

빅테크 기업들은 이러한 구조적 경쟁에 노출되어 전방 인프라 확충 속도를 늦출 수 없으며, 글로벌 하이퍼스케일러의 2026 년 합산 CapEx 는 전년 대비 70% 증가한 약 8,300 억 달러(Bloomberg Intelligence 기준)에 달할 것으로 전망된다. 이는 AI 인프라 수요 지속성에 대한 명확한 시그널로 해석된다.

하이퍼스케일러 CapEx 추이 및 전망치



자료: Bloomberg Intelligence, 키움증권 리서치

결론: AI 인프라 사이클은 2H26~2027 더 길어진다

우리는 시장의 CapEx 피로 우려에 반하여, AI 인프라 투자 사이클이 단기에 종료되지 않고 오히려 2H26~2027 구간에 재가속될 것으로 판단한다. 토큰 수요의 지수적 확장과 Agentic 워크로드의 상시화는 하이퍼스케일러로 하여금 신규 가속기, 메모리, 인터커넥트, 냉각 인프라에 대한 투자를 지속하도록 강제하는 구조적 환경이 형성되었다.

비용 감소가 연산의 보편화를 부르고, 보편화된 자율형 에이전트가 데이터센터 가동률을 영구적으로 끌어올리는 인프라 확장 주기가 도래했다. 하이퍼스케일러의 대규모 하드웨어 발주는 일시적 유행이 아닌 필수적 Base-load 확충 단계로 판단하며, 글로벌 테크 인프라 밸류체인 전반에 대한 비중확대 의견을 유지한다. 특히 (i) 광 인터커넥트·CPO, (ii) 액냉 기반 AI 서버 ODM/OEM 영역의 구조적 수혜를 최선호 익스포저로 제시한다.