

HBM3E 12-Hi 적층 구조 및 열관리(Thermal Management) 기술 분석 보고서

문서번호 CRSM-AI-2026-AUTO

작성일 2026-06-13

작성 Cressem AI 시스템 (자동 생성)

보안등급 사내 비밀 (Confidential)

버전 v1.0

목 차

HBM3E 12-Hi 적층 구조 및 열관리(Thermal Management) 기술 분석 보고서	3
개요/배경	3
HBM 패키징 기술의 진화: 1D에서 3D로	4
차세대 열 제어 솔루션: Hybrid Bonding 및 iHBM	7
결론/시사점	9

HBM3E 12-Hi 적층 구조 및 열관리(Thermal Management) 기술 분석 보고서

본 보고서는 HBM3E 12단(12-Hi) 패키지의 고단 적층 구조를 분석하고, 적층 기술의 진화 과정과 열 저항 저감을 위한 핵심 방안을 고찰한다. 특히 Advanced MR-MUF 및 차세대 Hybrid Bonding 기술이 열 방출 경로에 미치는 영향을 중점적으로 다룬다.

개요/배경

1. AI 가속기 시장의 폭발적 성장과 HBM의 역할

최근 생성형 AI(Generative AI) 기술의 급격한 발전과 대규모 언어 모델(LLM, Large Language Model)의 확산으로 인해, 방대한 양의 데이터를 초고속으로 처리할 수 있는 컴퓨팅 인프라에 대한 수요가 전례 없는 수준으로 급증하고 있습니다. 이러한 AI 워크로드를 뒷받침하기 위한 핵심 하드웨어인 AI 가속기(AI Accelerator), 대표적으로 NVIDIA의 H100 및 Blackwell(B200) 시리즈 등은 연산 능력만큼이나 데이터 전송 대역폭(Bandwidth) 확보를 최우선 과제로 삼고 있습니다.

기존의 DDR(Double Data Rate) 메모리 구조로는 AI 가속기가 요구하는 초고속·대용량 데이터 병목 현상(Memory Wall)을 해결하는 데 한계가 있습니다. 이에 따라 메모리 다이(Die)를 수직으로 적층하여 데이터 통로를 획기적으로 확장한 고대역폭 메모리(HBM, High Bandwidth Memory)가 AI 컴퓨팅의 필수 솔루션으로 자리 잡았습니다. 특히, 5세대 HBM 규격인 HBM3E(HBM3 Extended)는 1,024-bit의 광폭 인터페이스와 9.2 Gbps 이상의 핀당 데이터 전송 속도를 통해 1.2 TB/s 이상의 대역폭을 구현하며 AI 및 HPC(High Performance Computing) 시장의 핵심 동력으로 작용하고 있습니다 [출처: blogs.sw.siemens.com].

2. HBM3E 12-Hi 고단 적층 기술로의 패러다임 전환

HBM의 기술적 진화는 단순히 데이터 전송 속도를 높이는 것을 넘어, 제한된 패키지 높이 내에서 얼마나 더 많은 메모리 용량을 확보할 수 있는냐는 '적층 단수(Stacking Height)'의 싸움으로 전개되고 있습니다. 과거 HBM3 세대에서는 8단(8-Hi) 적층이 주류를 이루었으나, AI 모델의 파라미터 수가 기하급수적으로 증가함에 따라 더 높은 용량의 메모리가 요구되고 있습니다.

이에 따라 시장의 중심은 12단 적층 구조인 HBM3E 12-Hi로 급격히 이동하고 있습니다. SK하이닉스는 세계 최초로 36GB 용량을 구현한 12단 HBM3E 양산에 돌입하며 시장 선점을 가속화하고 있으며 [출처: news.skhynix.co.kr], 삼성전자 또한 고성능 메모리 솔루션으로서 12단 적층 기술을 고도화하고 있습니다 [출처: semiconductor.samsung.com]. 12단 이상의 고단 적층은 단위 면적당 용량을 극대화할 수 있다는 강력한 이점이 있지만, 이는 동시에 공정 난이도와 물리적 한계를 시험하는 기술적 난제를 동반합니다.

3. 고단 적층에 따른 핵심 기술적 난제: 수율과 열관리(Thermal Management)

HBM3E 12-Hi와 같은 고단 적층 구조가 상용화됨에 따라, 제조 공정 및 신뢰성 측면에서 두 가지 치명적인 기술적 장벽이 부각되고 있습니다.

첫째, **공정 정밀도 및 수율(Yield) 확보의 어려움**입니다. 수십 개의 TSV(Through Silicon Via, 실리콘 관통 전극)와 마이크로 범프(Micro-bump)를 통해 다이를 수직으로 연결하는 과정에서, 단 한 개의 접합 결함(Void, Crack)만 발생하더라도 전체 스택(Stack)이 불량으로 처리되는 구조적 취약성을 가집니다. 특히 적층 단수가 높아질수록 다이 간의 정렬(Alignment) 오차와 휨(Warping) 현상이 심화되어, 이를 제어하기 위한 초정밀 검사 및 본딩(Bonding) 기술이 필수적으로 요구됩니다 [출처: 매뉴얼HBM검사및테스트공정가이드라인20260509001.pdf].

둘째, **열 축적(Heat Accumulation) 및 방열 문제**입니다. 칩을 수직으로 촘촘하게 쌓을수록 내부에서 발생하는 열이 외부로 방출되는 경로가 차단되는 물리적 한계가 발생합니다. 고성능 연산 시 발생하는 열은 메모리 셀의 데이터 유지 능력을 저하시키고, 심할 경우 패키지의 물리적 변형을 초래하여 시스템 전체의 안정성을 해칠 수 있습니다. 따라서 HBM3E의 성능을 온전히 구현하기 위해서는 Advanced MR-MUF나 TC-NCF와 같은 적층 소재의 열 저항(Thermal Resistance) 개선뿐만 아니라, 차세대 Hybrid Bonding 및 iHBM(발열 저감 기술)과 같은 혁신적인 열관리 솔루션 도입이 필수적인 상황입니다 [출처: news.skhynix.co.kr, <https://www.mdpi.com/2079-9292/14/13/2682>].

본 보고서에서는 이러한 HBM3E 12-Hi의 구조적 특징을 상세히 분석하고, 적층 공정 기술의 비교를 통해 최적의 열 방출 경로를 도출하며, 향후 차세대 패키징 시장을 주도할 기술 로드맵을 제시하고자 합니다.

HBM 패키징 기술의 진화: 1D에서 3D로

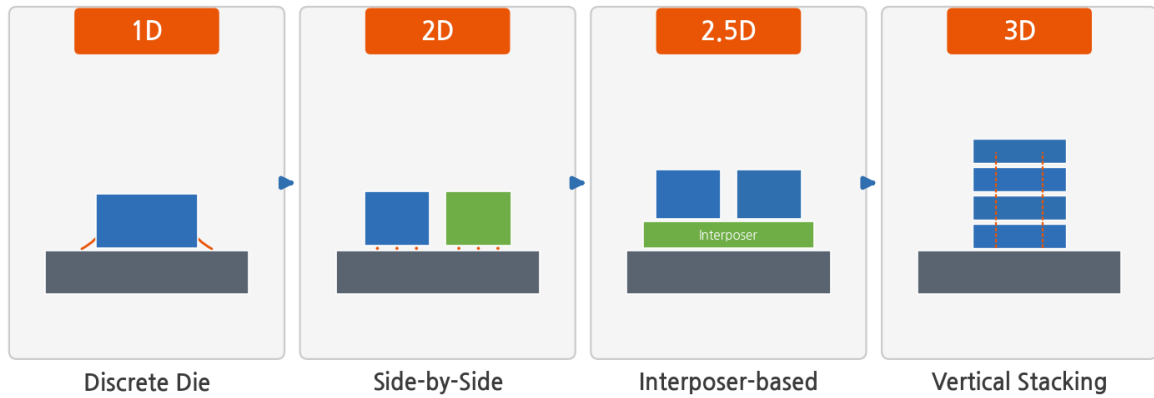
반도체 산업의 패러다임이 전공정(Front-end)의 미세화 한계에 직면함에 따라, 후공정(Back-end)에서의 패키징 기술은 단순한 보호 기능을 넘어 성능을 결정짓는 핵심 요소로 부상하였습니다. 특히 AI 가속기와 고성능 컴퓨팅(HPC) 시장의 폭발적 성장으로 인해, 데이터 전송 속도(Bandwidth)를 극대화하고 전력 효율을 높이기 위한 '패키징의 입체화'는 선택이 아닌 필수가 되었습니다. 과거 평면적인 구조에서 시작된 패키징 기술은 인터커넥트(Interconnect) 밀도를 높이는 방향으로 진화하며 1D, 2D를 거쳐 현재의 2.5D 및 3D 적층 구조로 발전해 왔습니다.

기존의 전통적인 패키징 방식은 칩을 기판(Substrate) 위에 수평적으로 배치하는 2D 구조에 머물러 있었습니다. 이 방식은 제조 공정이 단순하고 비용이 저렴하다는 장점이 있으나, 칩과 칩 사이, 혹은 칩과 메모리 사이의 물리적 거리가 멀어 데이터 전송 시 신호 지연(Latency)이 발생하고 전력 소모가 크다는 치명적인 한계가 있습니다. AI 연산과 같이 방대한 데이터를 실시간으로 처리해야 하는 워크로드에서는 이러한 물리적 거리에 따른 병목 현상이 전체 시스템 성능을 저하시키는 주요 원인이 됩니다.

이를 해결하기 위해 등장한 기술이 2.5D 패키징입니다. 2.5D 패키징은 칩들을 수평적으로 배치하되, 그 사이에 실리콘 인터포저(Silicon Interposer)라는 미세 회로 층을 삽입하는 방식입니다. 인터포저는 일반적인 기판보다 훨씬 미세한 회로 패턴을 구현할 수 있어, 로직 칩(GPU 등)과 고대역폭 메모리(HBM) 사이의 연결 밀도를 획기적으로 높여줍니다. 현재 NVIDIA의 Blackwell 아키텍처와 같이 CoWoS(Chip on Wafer on Substrate) 공정을 사용하는 고성능 AI 가속기들이 이 2.5D 구조를 기반으로 구현되고 있습니다.

현재 HBM의 핵심이자 기술적 정점에 있는 것은 3D 적층(3D Stacking) 기술입니다. 3D 기술은 칩을 수평이 아닌 수직으로 쌓아 올려 데이터가 이동해야 하는 물리적 경로를 최소화합니다. 이를 위해 실리콘 웨이퍼를 관통하는 미세 전극인 TSV(Through Silicon Via, 실리콘 관통 전극) 기술이 필수적으로 사용됩니다. TSV를 통해 수직으로 연결된 칩들은 마치 하나의 거대한 칩처럼 동작하며, 기존 방식과는 비교할 수 없는 압도적인 대역폭(Bandwidth)과 전력 효율을 제공합니다. HBM3E와 같은 최신 규격은 이러한 3D 적층 기술을 고도화하여 12단(12-Hi) 이상의 초고단 적층을 구현하고 있습니다.

반도체 패키징 구조의 기술적 진화 과정



개별 칩을 독립적으로 배치하여 연결하는 1D 구조에서 칩을 수평으로 배치하는 2D 구조로, 인터포저를 활용하여 칩 간 연결 밀도를 높인 2.5D 구조를 통해 칩을 수직으로 적층하여 대역폭을 극대화한 구조

그림 1. 반도체 패키징 구조의 기술적 진화 과정

HBM3E 12-Hi DRAM Die 단면 구조도

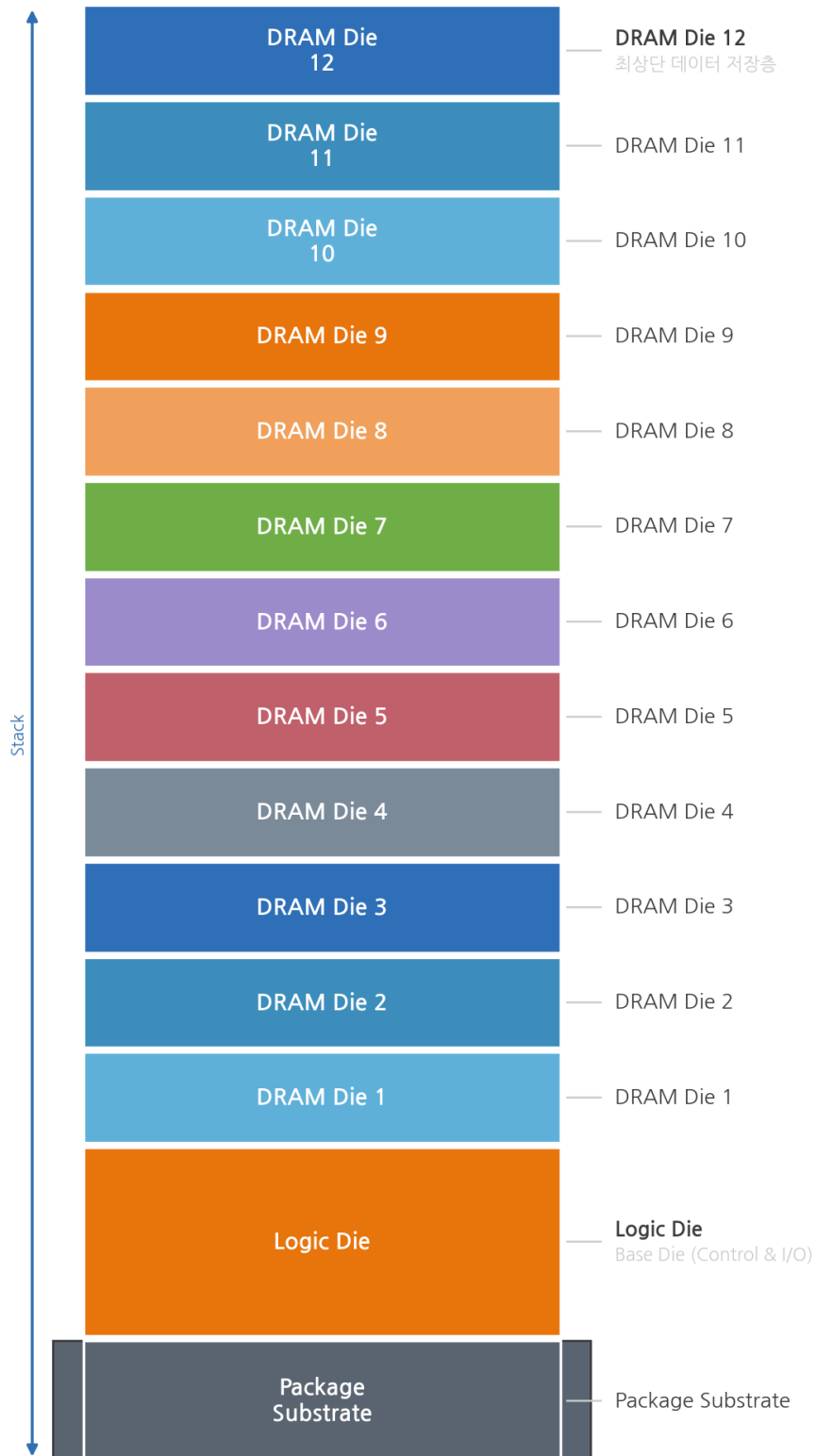


그림 2. HBM3E 12-Hi DRAM Die 단면 구조도

HBM3E 열 방출 및 관리 프로세스

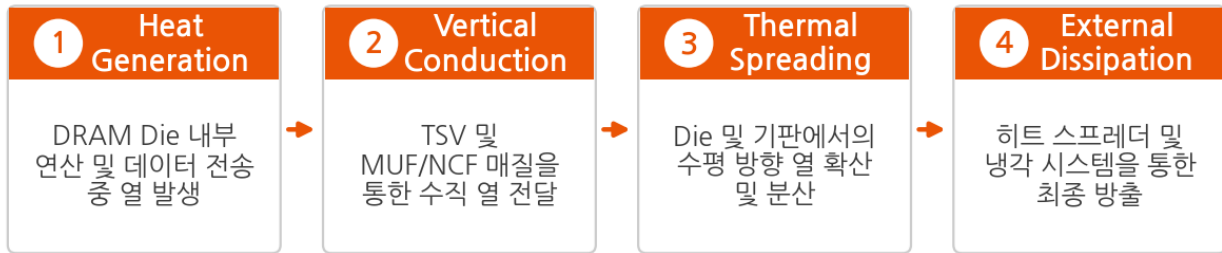


그림 3. HBM3E 열 방출 및 관리 프로세스

3. 열 관리 기술의 최신 동향 및 대응 전략

HBM3E의 고성능화를 뒷받침하기 위해 제조사들은 열 저항을 낮추기 위한 다양한 기술적 혁신을 도입하고 있습니다.

3.1 적층 소재 및 공정의 최적화

SK하이닉스는 Advanced MR-MUF 기술을 통해 칩 사이의 간극을 효과적으로 메우고, 열전도 특성이 강화된 신규 필러(Filler)를 첨가하여 방열 성능을 극대화하고 있습니다 [출처: HBM3e | SK hynix]. 또한, 적층 과정에서 발생하는 물리적 압력을 제어하여 휨(Warping) 현상을 방지함으로써, 접합부의 밀착력을 높여 열 저항을 낮추는 전략을 사용합니다 [출처: SK하이닉스 HBM3E 양산 동향 및 기술 경쟁력 분석20260509001.pdf].

3.2 적응형 전력 관리 및 냉각 기술 (Adaptive Power & Cooling)

최신 HBM3E 설계에는 실시간 열 조건과 워크로드 유형에 따라 스택의 주파수(Frequency)와 전압(Voltage)을 동적으로 조절하는 **적응형 전력 관리(Adaptive Power Management)** 기술이 적용됩니다. 이를 통해 성능 저하를 최소화하면서도 에너지 소비를 약 15% 절감하고 발열을 억제할 수 있습니다 [출처: HBM3e and HBM4: IC design guide for next-generation high bandwidth memory]. 또한, SK하이닉스가 공개한 'iHBM' 기술과 같이 냉각 방식 자체를 혁신하여 발열을 약 30%까지 낮추려는 시도가 이어지고 있습니다 [출처: SK hynix supplies 'early' HBM4 memory samples, qualification tests ...].

3.3 차세대 구조: Hybrid Bonding으로의 전환

HBM4 이후의 공정에서는 기존의 Bump 방식 대신 구리(Copper)를 직접 접합하는 **하이브리드 본딩(Hybrid Bonding)** 기술이 핵심이 될 전망입니다. 하이브리드 본딩은 범프(Bump)라는 중간 매질을 제거하고 Die와 Die를 직접 연결함으로써, 열 저항을 획기적으로 낮추고 매우 직접적인 열 방출 경로(Direct Thermal Pathway)를 제공할 수 있는 유망한 솔루션으로 평가받고 있습니다 [출처: Thermal Issues Related to Hybrid Bonding of 3D-Stacked High Bandwidth Memory: A ...].

결론적으로, HBM3E의 열관리 이슈는 단순한 부품의 문제를 넘어 패키징 구조, 소재 공학, 그리고 차세대 접합 기술이 통합되어야 해결 가능한 복합적인 과제입니다. 특히 고단 적층이 심화될수록 열 방출 경로 상의 미세한 결함(Void, Delamination 등)이 치명적인 성능 저하로 이어질 수 있으므로, 이를 정밀하게 검출할 수 있는 검사 기술의 중요성이 더욱 증대되고 있습니다.

차세대 열 제어 솔루션: Hybrid Bonding 및 iHBM

HBM의 적층 단수가 12단(12-Hi)을 넘어 HBM4(16단), 향후 HBM4E(20단)까지 증가함에 따라, 기존의 범프(Bump) 기반 접합 방식은 열 저항(Thermal Resistance)의 한계에 직면하고 있습니다. 적층 수가 늘어날수록 칩 내부에서 발생하는 열이 외부로 방출되는 경로가 길어지고, 칩 사이의 간격이 좁아지면서 열 축적(Heat Accumulation) 현상이 가속화되기 때문입니다. 이를 해결하기 위해 차세대 패키징 기술인 하이브리드 본딩(Hybrid Bonding)과 혁신적인 냉각 기술인 iHBM 기술이 핵심 솔루션으로 부상하고 있습니다.

1. Hybrid Bonding: 범프 없는 직접 접합을 통한 열 경로 최적화

Hybrid Bonding(하이브리드 본딩)은 기존의 마이크로 범프(Micro-bump)를 사용하지 않고, 구리(Cu) 패드와 절연막(Dielectric)을 직접 맞붙여 접합하는 기술입니다. 이는 HBM4 이후의 초고단 적층 구조에서 열 관리와 전기적 성능을 동시에 잡을 수 있는 게임 체인저로 평가받습니다.

- **구조적 혁신과 열적 이점:** 기존 방식은 Die와 Die 사이에 솔더(Solder) 범프와 언더필(Underfill) 소재가 존재하여, 이 층들이 열 저항체로 작용합니다. 반면, Hybrid Bonding은 Cu-to-Cu 직접 접합을 통해 칩 간의 물리적 간극을 극도로 줄입니다. 이는 열이 흐를 수 있는 직접적이고 저항이 낮은 경로(Low-resistance Thermal Pathway)를 제공하여, 칩 내부의 열을 상부 또는 하부로 전달하는 효율을 극대화합니다 [출처: MDPI - Thermal Issues Related to Hybrid Bonding of 3D-Stacked HBM].
- **전기적 성능 향상:** 범프의 크기가 사라짐에 따라 I/O(입출력) 밀도를 획기적으로 높일 수 있으며, 이는 데이터 전송 속도 향상과 함께 신호 무결성(Signal Integrity) 개선으로 이어집니다.
- **검사 난이도의 급증:** 하이브리드 본딩은 범프가 없는 초미세 접합을 요구하므로, 기존의 광학 검사(AOI)를 넘어선 극도로 높은 정밀도의 검사 솔루션이 필수적입니다. 크레셈은 이러한 기술 변화에 대응하여 초미세 간극을 측정할 수 있는 차세대 검사 모듈 개발에 집중하고 있습니다 [출처: 분석_크레셈CRESSEM20260509001.pdf].

2. iHBM: 냉각 방식의 패러다임 전환을 통한 발열 제어

하이브리드 본딩이 구조적 접합 방식을 개선한다면, iHBM(innovative HBM) 기술은 냉각 메커니즘 자체를 변화시켜 발열을 직접적으로 억제하는 데 목적을 둡니다.

- **발열 감소 성능:** iHBM 기술은 기존의 냉각 방식과는 차별화된 새로운 구조적/재료적 접근을 통해 HBM의 발열을 약 30% 수준까지 낮추는 것을 목표로 합니다 [출처: SK하이닉스 인스타그램/뉴스 자료]. 이는 AI 가속기와 같은 고부하 워크로드 환경에서 HBM의 동작 안정성을 확보하는 데 결정적인 역할을 합니다.
- **적응형 전력 관리(Adaptive Power Management)와의 결합:** 최신 HBM3E 및 차세대 제품군에는 실시간 열 조건과 워크로드 유형에 따라 스택의 주파수와 전압을 동적으로 조절하는 기술이 적용됩니다. 이러한 기술은 에너지 효율을 약 15% 향상시키며, iHBM과 같은 냉각 기술과 결합될 때 시너지를 발휘하여 데이터 센터의 전력 소모와 열 관리 부담을 동시에 경감시킵니다 [출처: Wevolver - HBM Memory Engineering Guide 2025].

3. 차세대 열 제어 기술 비교 분석

구분	기존 범프 방식 (TC-NCF / MR-MUF)	Hybrid Bonding (차세대)	iHBM (냉각 혁신)
접합 방식	Micro-bump + Underfill	Cu-to-Cu Direct Bonding	구조적/재료적 냉각 최적화
열 저항	높음 (범프/언더필 층 존재)	매우 낮음 (직접 경로 확보)	획기적 감소 (발열 30% 저감)
적층 밀도	물리적 범프 크기에 제한됨	초고밀도 구현 가능	기존 구조 내 효율 극대화
주요 타겟	HBM3, HBM3E (현재 주력)	HBM4 및 그 이후 세대	고성능 AI 가속기용 HBM

검사 핵심	범프 높이 및 정렬(Alignment)	초미세 접합면 및 표면 평탄도	열 방출 효율 및 신뢰성 검증
-------	-----------------------	------------------	------------------

표 1. 차세대 열 제어 기술 비교 분석**

결론적으로, HBM의 고단화 추세는 단순히 칩을 더 많이 쌓는 것을 넘어, '어떻게 쌓고 어떻게 식힐 것인가'의 문제로 귀결됩니다. Hybrid Bonding을 통한 구조적 저항 제거와 iHBM을 통한 능동적/수동적 냉각 기술의 결합은 향후 AI 반도체 시장의 주도권을 결정짓는 핵심 기술 경쟁력이 될 것입니다.

결론/시사점

HBM3E 12-Hi로 대표되는 고단 적층 메모리 시장은 AI 가속기의 성능을 결정짓는 핵심 요소로 자리 잡았으며, 이에 따라 패키징 기술의 고도화와 열관리(Thermal Management) 솔루션의 확보가 제조사의 핵심 경쟁력이 되었습니다. 본 보고서에서 분석한 기술적 흐름을 바탕으로 향후 시장의 성패를 가를 주요 시사점을 다음과 같이 정리합니다.

1. 수율(Yield) 확보를 위한 고정밀 검사 기술의 필수성

HBM3E 12-Hi와 같이 적층 단수가 증가할수록, 개별 DRAM Die의 결함뿐만 아니라 적층 과정에서 발생하는 미세한 물리적 변형(Warping) 및 TSV(Through Silicon Via) 연결 불량은 전체 패키지의 치명적인 손실로 이어집니다. 특히 12단 이상의 구조에서는 Die-to-Die 간의 정렬(Alignment) 오차와 마이크로 범프(Micro-bump)의 접합 신뢰성이 수율을 결정짓는 핵심 변수입니다. 따라서 단순한 전기적 특성 검사를 넘어, X-ray/CT 기반의 비파괴 검사 및 고해상도 광학 검사(AOI)를 통한 실시간 결함 검출 역량이 수율 극대화의 필수 조건입니다. [출처: 매뉴얼HBM검사및테스트공정가이드라인20260509001.pdf]

2. 열관리 기술이 주도하는 차세대 패키징 로드맵

기존의 TC-NCF 방식과 Advanced MR-MUF 방식의 경쟁은 결국 '어떻게 하면 더 효율적으로 열을 배출할 것인가'라는 열적 안정성(Thermal Stability) 문제로 귀결됩니다. 적층 단수가 16단(HBM4), 20단(HBM4E)으로 늘어남에 따라 기존의 Bump 기반 접합 방식은 열 저항(Thermal Resistance)의 한계에 직면할 것으로 예측됩니다. 이에 따라 구리를 직접 접합하여 열 방출 경로를 최단화하는 하이브리드 본딩(Hybrid Bonding) 기술이 차세대 표준으로 부상할 것이며, 이는 검사 장비 측면에서도 초미세 간극을 측정할 수 있는 극도로 높은 정밀도를 요구하게 될 것입니다. [출처: HBM 두께완화산업보고서 GrowthResearch(26.04.06).pdf]

3. 요약 및 향후 전망

향후 HBM 시장은 단순히 용량과 대역폭(Bandwidth)을 높이는 단계를 넘어, '고적층-저발열-고수율'이라는 세 가지 난제를 동시에 해결하는 기업이 주도할 것입니다.

구분	현재 (HBM3E 12-Hi)	미래 (HBM4 및 그 이후)
주요 적층 기술	Advanced MR-MUF / TC-NCF	Hybrid Bonding (Direct Copper Bonding)
핵심 과제	적층 안정성 및 방열 성능 개선	초미세 간극 제어 및 극단적 열 관리
검사 요구사항	TSV 및 Bump 연결성 정밀 검사	초고해상도 3D 구조 및 접합부 결함 분석

표 2. 결론/시사점

결론적으로, 반도체 제조사는 공정 기술의 혁신과 더불어, 공정 중 발생하는 불량률 실시간으로 추적하고 원인을 분석(Root Cause Analysis)할 수 있는 AI 기반 스마트 검사 솔루션을 통합함으로써 차세대 AI 메모리 시장의 주도권을 확보해야 합니다. [출처: 분석_크레셈CRESSEM20260509001.pdf]