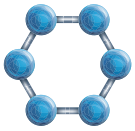


3D나노융합소자연구센터 기술리포트

2024년 11월호
(발행일 2024년 11월 1일)

3DC²

Contents

HBM 기술동향

서 론

HBM(High Bandwidth Memory)은 고 대역폭의 메모리 성능을 구현하기 위해 제안된 적층형 메모리로 2013년 SK하이닉스가 세계 최초로 개발하였고, 그해에 세계 반도체 표준 협회인 JEDEC에 의해 표준 규격으로 채택되었다. 초기에는 단순하게 그래픽 성능 강화를 목적으로 GDDR 계열의 DRAM을 대체하기 위해서 개발이 시작되었다. 최근에는 빠른 디지털 시대로의 전환과 인공지능 산업의 발전에 따른 초고성능 컴퓨팅 시장의 급격한 성장에 의해 HBM 수요와 시장 규모가 지속적으로 확대되고 있다.

인공지능(Artificial Intelligence)이란 인간의 지능이 가지는 학습, 추리, 적응, 논증 등의 기능을 인공적으로 구현하는 것으로 일반적으로는 인간의 지능을 모방한 기능을 갖춘 컴퓨터 시스템을 의미한다. 1955년 미국 다트머스 대학교에서 개최한 학회에서 존 매카시가 이 용어를 사용한 이 후 수십년 동안 많은 사람들의 연구에 의해 다양한 이론이 개발되고 발전 되어왔다. 특히 2016년에는 구글 브레인(현 구글 딥 마인드)에서 발표한 알파고에 의해 특정 분야에서는 인간의 수준을 뛰어 넘는 결과를 보여줄 수 있다는 사실을 입증하여 대중화에 성공하게 된다. 2017년에는 대규모의 언어 데이터를 학습해서 이해하고 상호 작용이 가능한 생성 형 인공지능이 출현하였으며, (OpenAI의 Chat GPT 등.) 이 후 클라우드 컴퓨팅을 기반으로 한 빅테크 기업들의 생성 형 인공지능 개발 경쟁이 시작되면서 본격적으로 인공지능이 대중적으로 상용화되기 시작했다.

인공지능 시스템을 구현하기 위해서는 대용량의 데이터 학습과 추론을 더 빠르고 효율적으로 수행하기 위해 설계된 맞춤형 하드웨어 장치가 필요한데 이를 인공지능 가속기(AI accelerator)라고 한다. 인공지능 가속기는 인공지능 및 딥 러닝에 최적화된 연산 구조를 제공하는 연산 알고리즘, 다수의 연산을 동시에 수행하여 연산 속도를 크게 향상시키는 병렬처리 아키텍처, 높은 연산 성능을 유지하면서 낮은 전력 소모를 실현하는 전력의 효율성 등의 기능을 가지고있어야 하는데, 이런 측면에서 현재까지 가장 최적화된 Solution 제품이 GPU(프로세서)와 HBM(메모리)을 조합한 인공지능 반도체라고 할 수 있다.

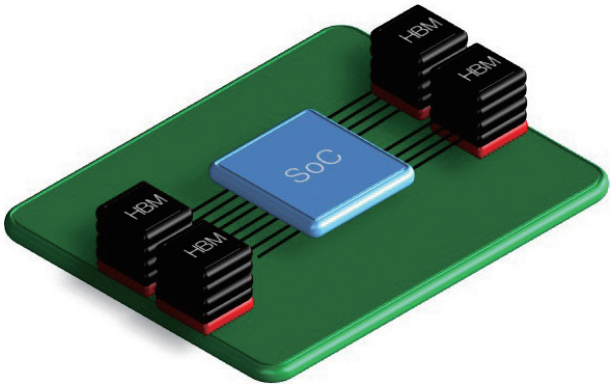
본 기술리포트에서는 최근 메모리 반도체 시장의 새로운 성장동력으로 부상한 HBM의 개발 과정과 제작에 필요한 요소기술 및 향 후 개발 동향에 대해 순서대로 기술해 나갈 예정이다.

HBM 개요

1. HBM 역사

HBM의 개발은 다국적 비디오게임 기업인 닌텐도의 제안에서 시작되었다. 닌텐도는 그래픽 성능 강화를 위해 메모리의 대역폭을 늘리는 방안을 제안했으며, 이에 따라 GPU를 제작하는 AMD와 메모리를 제작하는 SK하이닉스가 공동으로 참여해 프로젝트가 시작된 것이다.

초기 HBM은 커스터마이징 된 제품으로, 게임 업체나 GPU 업체의 요구에 맞추어 제작해야 했기 때문에 큰 수익을 기대하기는 어려웠다. 특수 목적의 제품이다 보니 수요가 한정적 이었으며, 불과 10년 전만 해도 HBM은 전체 메모리 시장에서 1%도 채 안되는 비중을 차지하고 있었다. 하지만 SK하이닉스는 향후 초고속 대용량 데이터 처리 수요가 늘어날 것이라는 판단하에 기술 개발을 지속해서 이어갔고, 그 결과 2013년 세계 최초로 HBM을 개발하는 데 성공했다. ([그림 1]



[그림 1] HBM을 프로세서와 결합한 SiP 제품

HBM 개발에서는 SK하이닉스가 선구자 역할을 했지만, 이후에는 삼성전자가 HBM 개발에 주도권을 잡는 양상을 보였다. 삼성전자는 2016년 기존 HBM1(1세대) 보다 2배 빠른 속도로 정보를 처리하는 HBM2(2세대)를 세계 최초로 개발했고 양산에 성공했다. 그러나 삼성전자는 2017년 HBM2E(3세대) 제품을 출시한 이후 2019년부터 HBM 투자를 중단했다. 이 회사 경영진들은 당시 HBM 시장 규모가 매우 작고 시장 성장에 확신이 안 서자 투자 중단을 결정한 것인데, 특히 연구개발 전담 팀을 해체하면서 HBM 선두 경쟁에서 완전히 이탈했다.

SK하이닉스는 SK 그룹의 지원으로 2017년부터 HBM 관련 R&D에 대규모 투자를 단행하기 시작했으며, 2년 뒤에는 HBM 공정의 핵심 기술인 실리콘관통전극(TSV) 공정 안정화를 목표로 대규모 예산을 집행하기도 했다. 그 결과 4세대 제품인 HBM3부터 시장의 주도권을 완전히 가져왔다. (출처: 'SK하이닉스 투자할 때, 삼성전자는 중단', 뉴스웍스)

일각에서는 HBM 투자는 SK하이닉스였기에 가능했을 수도 있다고 말하기도 한다. SK하이닉스에게 메모리는 반도체 사업의 거의 전부였지만, 삼성전자는 반도체 사업에서 메모리에만 집중하지 않았으며 파운드리 사업과 AP, 이미지센서 등 시스템 반도체 사업도 함께 진행하고 있었고 이러한 상황에서 삼성전자와 같은 대규모 조직이 HBM 시장에 적극적으로 대응하기에는 시장 규모가 너무 작았기 때문이다.

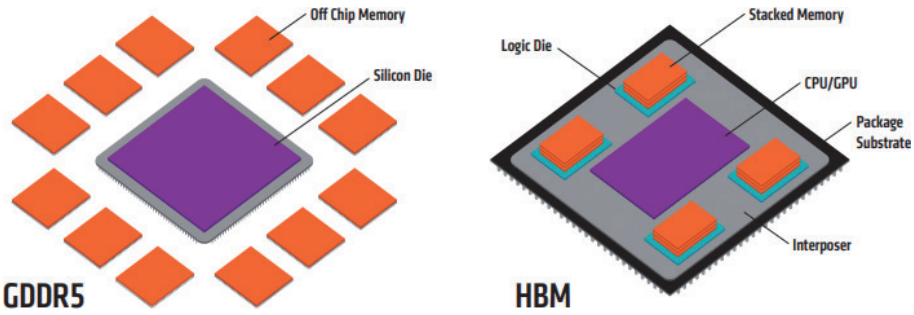
2. HBM 기본 구조와 원리

HBM은 많은 수의 신호 입출력 단자(I/O, Input/Output)를 형성해서 대량의 데이터를 한 번에 전송할 수 있는 고성능 메모리 반도체이며 여러 개의 DRAM 칩을 수직으로 적층하여 상호 연결하여 제작된 제품이다. 일반 DRAM보다 고속으로 데이터를 전송할 수 있는 성능으로 인해 빠른 속도의 대량 연산 작업이 필요한 초고성능 컴퓨팅이나 인공지능 시스템 구동에 최적화된 메모리 제품으로 평가되고 있다.

HBM은 DRAM 다이를 수직으로 적층하여 실리콘을 관통하는 금속 배선(TSV, Through Silicon Via)를 통해 주프로세서와 통신을 하는 원리이다. 이를 위해서 칩을 직접 인쇄 회로 기판(Substrate) 위에 올려놓는 GDDR 계열의 DRAM과는 달리, DRAM 칩(Core die)을 DRAM control 목적의 Logic 칩(Base die) 위에 적층하고, 다시 기판과 Logic 칩 사이에 Interposer라는 중간 단계를 삽입 한다. ([그림 2] ~ [그림 4])

GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V

[그림 2] GDDR과 HBM1의 단면 구조 비교 (출처: AMD)

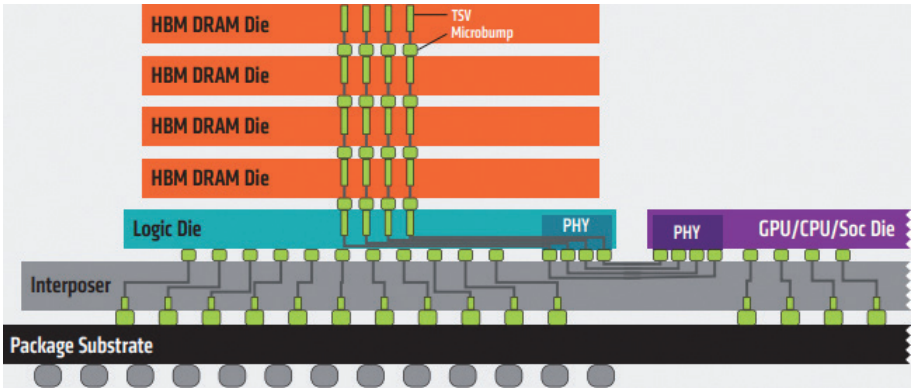


[그림 3] GDDR과 HBM1의 평면 구조 비교 (출처: AMD)

[GDDR의 경우 신호 입출력 단자를 연결하기 위해서 메모리 다이 표면 32개의 데이터 전송 핀을 단순한 후공정 구리 배선 공정으로 기판과 연결하면 되므로 따로 미세 공정이 필요 없었다. 그러나 HBM에서는 1024개나 되는 미세한 핀을 연결해야 하기 때문에 그대로 기판에 붙일 수 없다. 설령 그대로 붙인다고 하더라도 1024개나 되는 배선을 기판에 구현하여 CPU/GPU에 연결하는 것도 쉽지 않은 일이라, 중간에 Interposer를 추가하여 여기에

CPU/GPU와 HBM을 가깝게 배치해서 미세 공정을 통해 연결하자는 아이디어가 나왔다.

2012년에 실리콘 인터포저 위에 이종 반도체 소자 다이를 수평 배치해서 연결하는 2.5D 패키지 기술인 TSMC의 CoWoS(Chip-on-Wafer-on-Substrate)가 개발되고, 2014년에 AMD와 SK하이닉스가 협력하여 TSV(Through Silicon Via) HBM 제품 개발에 성공하면서, 이후 본격적으로 프로세서와 HBM 메모리를 활용한 2.5D SiP(System in Package) 제품이 나오게 되었다. ([그림 4])



[그림 4] TSMC의 2.5D 패키지 기술을 이용한 이종집적소자 (출처: AMD)

HBM은 다음과 같은 여러가지 장점을 가지고 있어서 초고성능 컴퓨팅이나 인공지능 시스템 구현에 최적화된 제품으로 인정받고 있다.

1) 짧은 레이턴시와 높은 메모리 대역폭 : 메모리 칩을 관통하는 구멍을 뚫어 1,024개나 되는 채널을 구성하므로 프로세서와 HBM 간 거리가 매우 짧아졌고, 신호 노이즈와 같은 간섭 현상이 거의 없어서 HBM의 높은 대역폭을 손실 없이 그대로 실제 성능으로 구현 시키는 것이 가능하다.

2) 작은 칩 면적과 작은 컨트롤러 면적 : 수직 적층으로 PCB 기판에서 차지하는 메모리 칩의 총 면적을 줄일 수 있고, 프로세서 내부에 탑재되는 내장 메모리 컨트롤러 크기도 감소시킬 수 있다.

3) 낮은 전력 소모 : Interconnection 거리의 최소화로 인해 메모리 용량 1GB 당 전력 소비 W를 비교하면 GDDR 보다 1/4 수준의 전력 소모량을 보인다.

4) 메모리 용량의 확장성 : 적층 과정에서 메모리 다이를 얇게 만드는 그라인딩 장비, 본딩 장비, 접합 물질, 본딩 기술 등이 발전하면서 수직으로 용량을 확장시키는 것이 가능 해졌다.

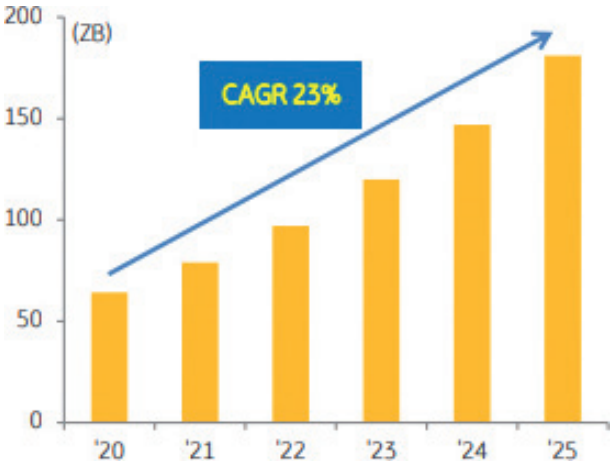
그러나 향 후 더 큰 메모리 용량과 더 우수한 성능을 가지는 차세대 HBM 개발을 위해서는 다음과 같은 기술적인 난제를 극복해야 한다.

1) 높은 제작 난이도와 낮은 수율 : 다이를 한층 한층 접합(본딩)하는 과정에서 불량이 발생할 수 있어 최종 수율이 감소된다. (예를 들어 한 층을 쌓을 때 90% 확률로 성공한다고 하더라도, 8층의 최종 수율은 0.9^8~43% 밖에 되지 않는다. 2024년 8단 기준 마이크론과 삼성전자의 수율이 30~50%로 추정되며, SK하이닉스의 경우 70~80% 정도로 추정되고 있다.

2) 높은 발열 현상으로 인한 성능 저하 : GDDR에 비해 구조가 복잡하며 열원이 모든 방향으로 분산되는 GDDR에 비해 하나의 칩에 수직으로 적층되어 있어 열원이 더욱 집중되어 나타난다. 그러므로 방열에 매우 불리한 구조를 가지고 있다.

3. HBM 시장 현황과 전망

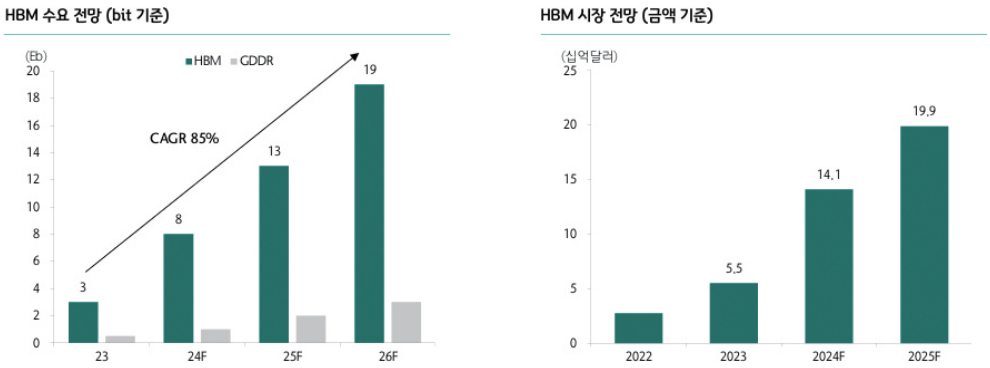
[그림 5]와 같이 최근 몇 년간 디지털 시대로의 전환이 가속화되면서 정보통신기술(ICT) 분야에서 생성되고 사용되는 데이터의 양이 폭발적으로 증가했으며(2023년: 10^{21}바이트), 향후 5년 동안 인공지능·자율주행·빅데이터 등의 ICT 기술이 일상화 되면 지금보다 3배가 넘는 데이터가 생성될 전망이다. 이러한 방대한 양의 데이터를 빠르게 처리하기 위해서는 초고성능 컴퓨팅이 요구되며, 특히 최근 공개된 초거대 인공지능(Super-Giant AI)인 Chat GPT의 상업적인 성공을 계기로 대용량 데이터의 생성과 빠른 연산에 대한 수요는 더욱 확대될 전망이다.



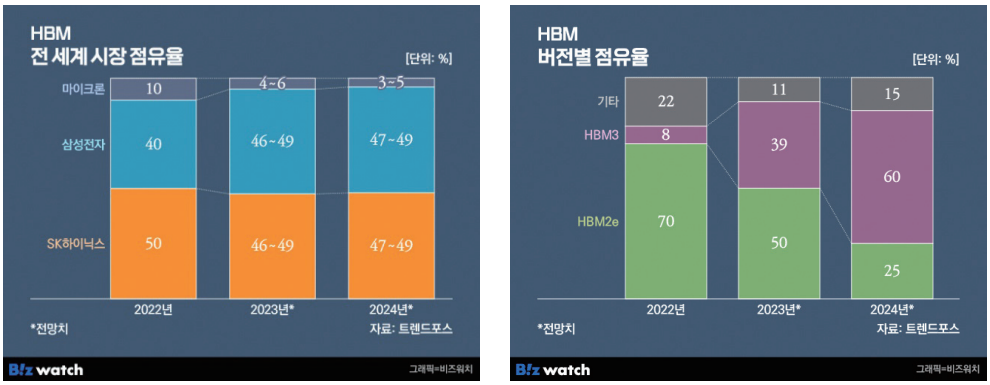
[그림 5] 디지털 데이터 생성량 변화

[그림 6]과 같이 초고성능 컴퓨팅에서는 대량의 데이터를 지연 없이 동시에 처리할 수 있는 메모리 반도체가 필요한데, 현재까지는 HBM이 거의 유일한 해결책이라 할 수 있으며, 이러한 HBM 수요 증가에 따라 최근 몇 년 간 HBM 시장은 폭발적인 성장을 보여왔다. 2023년 기준 전체 메모리 반도체 시장에서 HBM이 차지하는 비중은 bit 기준 1% 미만이지만 매출은 10%, 영업이익률은 50%에 달할 정도로 고부가가치(일반 DRAM 보다 5배 이상의 높은 수익성)의 하이 엔드 반도체 제품으로 자리잡았다.

[그림 7]과 같이 2023년 세계 HBM 시장에서는 하이닉스와 삼성전자가 각각 53%, 38%로 도합 90% 수준의 압도적인 점유율을 기록하고 있고 나머지 10%는 마이크론이 차지하고 있다. 2024년은 수요가 공급을 초과한 상황으로(수요가 전년 대비 3배 이상 증가) 공급량이 증가하는 만큼 시장 규모가 커질 전망이며, 특히 최신 HBM3의 경우 SK하이닉스가 시장의 90%를 차지할 정도로 거의 독주하고 있다.

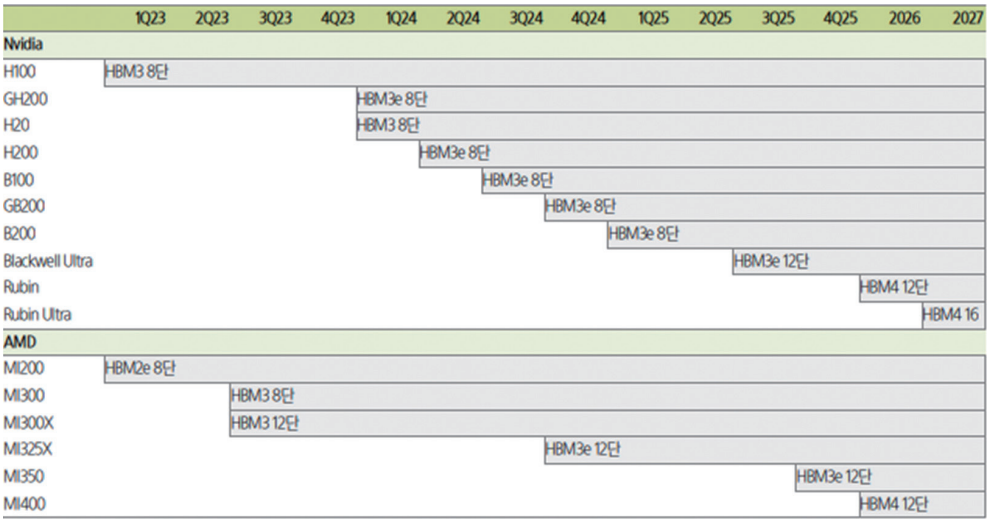


[그림 6] HBM 수요 전망(출처: TechInsight) 및 시장 전망 (출처: Yole group)



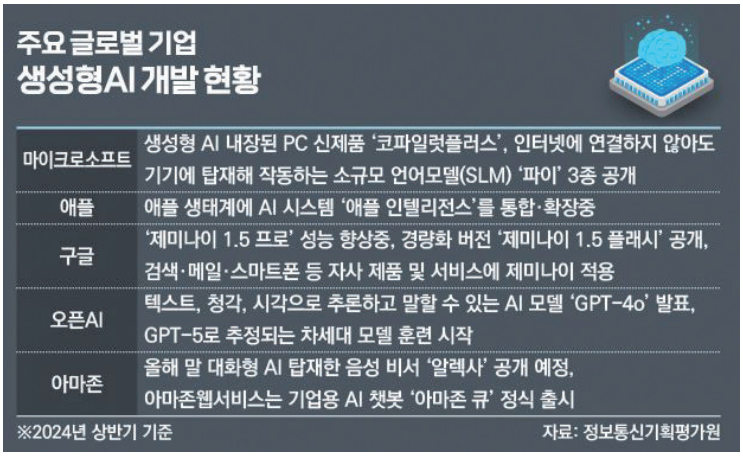
[그림 7] 메모리 반도체 3사의 HBM 시장 점유율 (출처: 트렌드포스)

인공지능 산업의 또 다른 특징은 수요의 확장성이 매우 크다는 점이다. 서버, IT 디바이스와 같은 하드웨어 외에도 자율주행, 공장자동화, 의료서비스 등 다양한 산업 영역으로 발전해 나갈 수 있으며, 이는 미래 경쟁력 확보 차원에서 인공지능에 대한 투자는 지속적으로 증가될 수밖에 없다는 것을 의미한다. 올해 대만에서 개최된 Computex 2024에서 Nvidia와 AMD는 AI용 GPU 신제품 출시 로드맵을 구체화했고, 인공지능 생태계 확대를 위한 업계의 노력은 한층 더 강화되는 추세이며, 이에 따라 HBM 수요는 향 후 몇 년간 급증하게 될 것으로 예상된다.



[표 1] Nvdia와 AMD의 신제품 출시 및 HBM 사용 계획 (출처: 삼성증권)

그동안 인공지능 투자에 있어서 가장 큰 장벽이 높은 인프라 투자 비용이었다. 미래 산업 경쟁력 확보에 있어 인공지능이 반드시 필요한 것은 알고 있지만, 높은 인프라 투자 부담으로 인해 많은 빅테크 기업들이 투자를 과감하게 진행하지 못했는데, 이러한 고민은 Nvidia의 Blackwell과 같은 차세대 가성비 제품의 출시로 완화되는 분위기이다. GPU 업계의 전략 변화(신제품 가격 경쟁력 확대 및 물량 확대)에 따라 보다 저렴한 비용으로 인공지능 기술 개발에 나설 수 있는 기회가 늘어나고 있고, 이는 보다 많은 고객의 인공지능 투자 참여로 이어질 전망이다.



주요 글로벌 기업 생성형AI 개발 현황

마이크로소프트	생성형 AI 내장된 PC 신제품 '코파일럿플러스', 인터넷에 연결하지 않아도 기기에 탑재해 작동하는 소규모 언어모델(SLM) '파이' 3종 공개
애플	애플 생태계에 AI 시스템 '애플 인텔리전스'를 통합·확장중
구글	'제미나이 1.5 프로' 성능 향상중, 경량화 버전 '제미나이 1.5 플래시' 공개, 검색·메일·스마트폰 등 자사 제품 및 서비스에 제미나이 적용
오픈AI	텍스트, 청각, 시각으로 추론하고 말할 수 있는 AI 모델 'GPT-4o' 발표, GPT-5로 추정되는 차세대 모델 훈련 시작
아마존	올해 말 대화형 AI 탑재한 음성 비서 '알렉사' 공개 예정, 아마존웹서비스는 기업용 AI 챗봇 '아마존 큐' 정식 출시

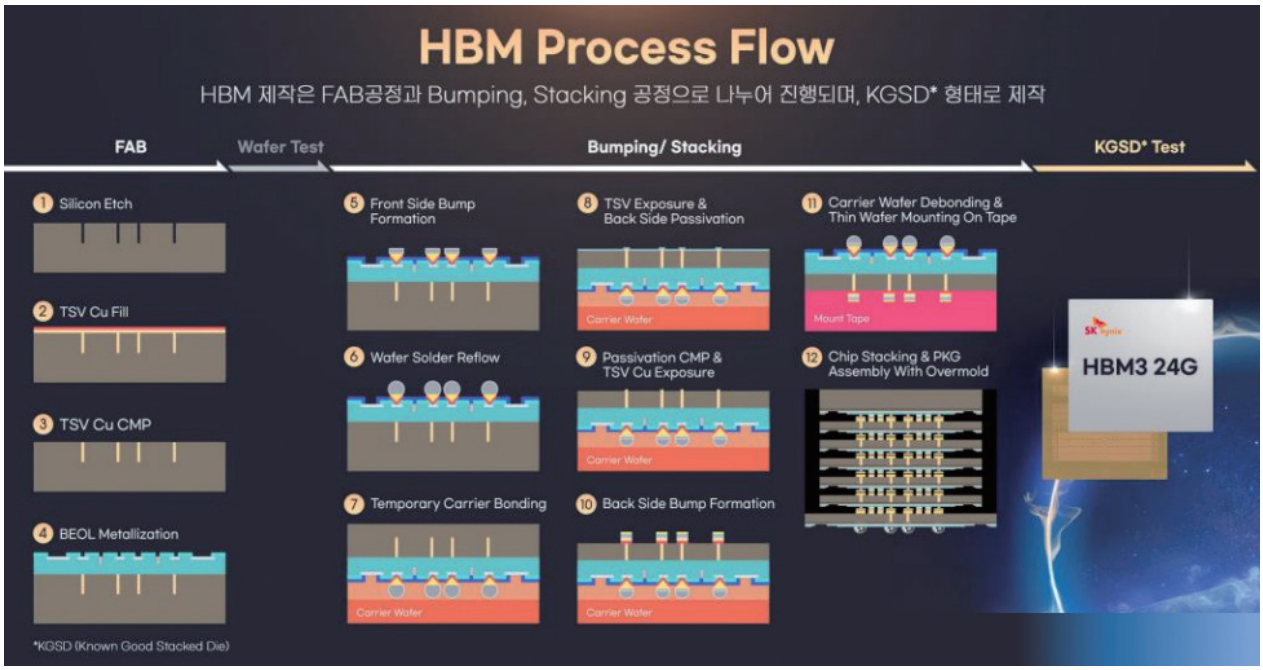
※2024년 상반기 기준 자료: 정보통신기획평가원

[표 2] 주요 빅테크 기업의 인공지능 투자 계획 (2024년)

HBM 요소 기술

1. HBM 제조공정

SK하이닉스에서 대외적으로 발표한 HBM 제조공정 순서도는 [그림 8]과 같다.



[그림 8] SK하이닉스의 HBM 제조공정

1) FAB : 전공정에서 DRAM 소자 제작 완료 후 TSV 형성 공정을 진행한다.
   ① TSV etch : 칩을 관통하는 Via 홀 형성을 위한 Silicon etch 진행
   ② TSV Cu fill : Barrier metal과 Seed Cu 증착 후 Electroplating Cu 증착
   ③ TSV Cu CMP : TSV Cu의 전기적인 절연과 평탄화를 위한 CMP 진행
   ④ BEOL metallization : 외부와 전기적인 연결을 위한 금속 배선과 패드 형성(재배치)

2) Bumping : 후공정에서 Wafer level packaging 기술을 사용해 TSV가 형성된 칩의 Front-side/Back-side bumping 공정을 진행한다.
   ⑤ Front-side bump formation : Cu 패드 위에 절연막을 증착하고 패턴을 형성한 후 Electroplating 공정으로 Cu-pillar 및 Solder(Sn-Ag) micro-bump를 형성
   ⑥ Solder reflow : 고온 열처리 공정을 통해 구형의 Solder bump를 형성
   ⑦ Temporary carrier bonding : Temporary carrier wafer를 TSV가 형성된 Core wafer front-side에 부착

⑧ TSV exposure & Backside passivation : Core wafer back-side thinning(grinding)으로 TSV를 노출시키고 보호막 코팅을 위한 Passivation 공정 진행

⑨ Passivation CMP & TSV Cu exposure : CMP 공정으로 평탄화를 진행하고 Back-side TSV를 노출시킴

⑩ Back-side bump formation : Front-side와 유사한 방법으로 Back-side에 노출된 TSV 위에 Cu-pillar 및 Solder(Sn-Ag) micro-bump를 형성

3) Stacking : Front-side/Back-side에 형성한 bump들을 본딩하여 칩을 수직으로 적층한다.

⑪ Carrier wafer debonding & Thin wafer mounting on tape : Wafer front-side에 부착했던 Carrier를 제거하고 그라인딩 된 Wafer back-side에 마운팅 테이프 부착

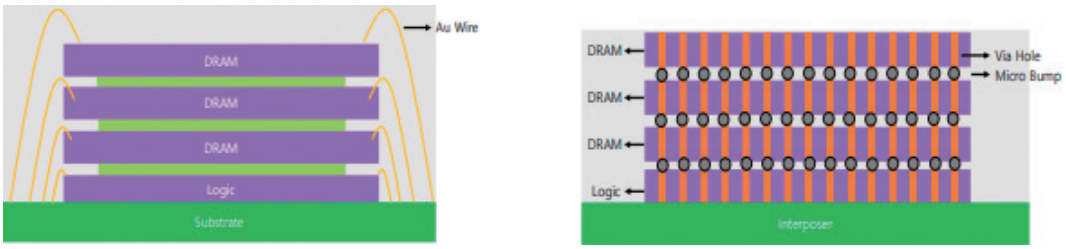
- Core(DRAM) wafer로 사용될 메모리 칩을 절단(Sawing/Dicing)

⑫ Chip stacking & PKG assembly with Overmold : Base(Logic buffer) wafer 위에 TSV가 형성된 Core wafer 칩을 Bump bonding 공정을 이용해 적층

- Chip to wafer bonding 방식으로 적층하여 KGSD(Known Good Stack Die) 형태로 제작
- 적층이 완료되면 Base wafer 위에서 Molding 하고 Base carrier wafer를 제거시킨 후 칩 단위로 절단

2. 실리콘 관통 전극(TSV) 기술

[그림 9]와 같이 TSV는 실리콘에 구멍을 뚫어서(Via hole) 전도성 재료로 채운 전극을 의미하며 칩을 수직으로 적층하기 위한 요소 기술이다. 칩을 적층할 때 기존에는 칩과 칩, 칩과 기판을 와이어로 연결하던 것을 TSV 전극을 이용해 수직으로 연결한다.



[그림 9] 와이어 본딩과 TSV 적층 기술의 비교 (출처: Skhynix newsroom)

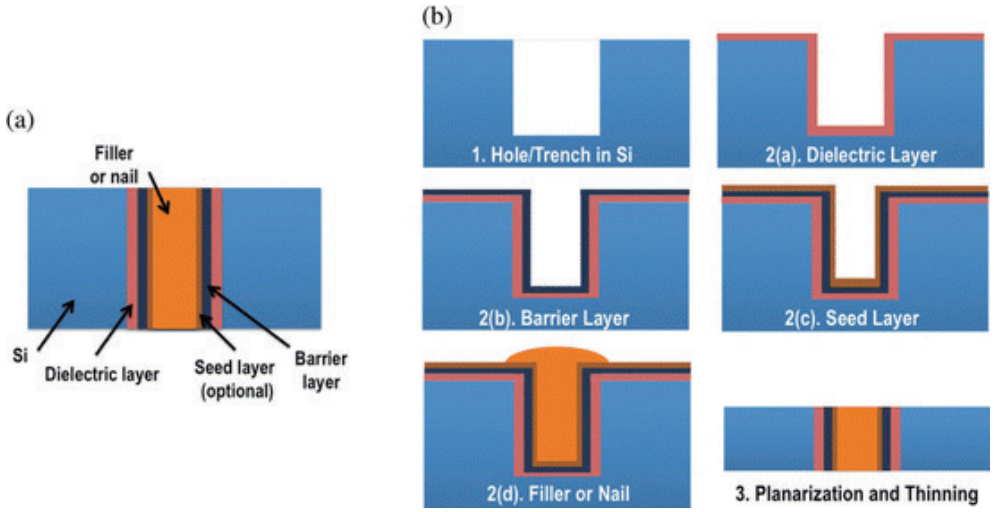
TSV 공정은 해당 공정을 CMOS 공정 이전에 진행하는가, CMOS 공정과 금속 배선 공정(BEOL metallization) 사이에 진행하는가, 금속 배선 공정 이후에 진행하는가에 따라 Via First, Via Middle, Via Last로 불리는데 현재 가장 널리 사용되고 있는 방법은 Via Middle이다. [그림 10]에 일반적인 CMOS 소자 적층에 사용되는 Via middle TSV 제조공정을 도시 하였으며 다음과 같은 순서대로 공정을 진행한다.

1) Silicon etch : CMOS 공정이 완료된 실리콘 웨이퍼에 Trench 홀을 형성하는 단계로, 수직의 Via 홀 모양을

유지하기 위해서 일반적으로 이방성 식각 특성을 나타내는 Bosch process와 Cryogenic process를 적용하는 DRIE(deep reactive ion etching) 식각법을 사용한다.

2) TSV Cu fill : ①절연 및 Cu 확산 방지 목적으로 CVD SiO2, 혹은 SiN 등의 Dielectric layer를 증착. ②접합 강도 향상과 Cu 확산 방지 목적의 PVD Ti, 혹은 Ta 등의 Barrier layer를 증착. ③후속 Cu의 전해 도금을 균일하게 하기 위한 목적의 PVD Cu Seed layer를 증착. ④전해 도금을 통해 Via 홀을 Cu로 채운다. 전해 도금을 이용한 Cu 증착 방법은 공정 비용이 비교적 저렴한 장점이 있으며, 도금 전류 파형의 조절이나 첨가제(Accelerator/Suppressor)를 혼합한 도금액을 사용하여 결함 없는 Via 홀의 충전(Super-filling)이 가능하다.

3) TSV Cu CMP : CMP 공정으로 평탄화 시키며 동시에 웨이퍼 윗면에 있는 Cu를 모두 제거해 Via 홀에만 Cu가 채워질 수 있게 한다.

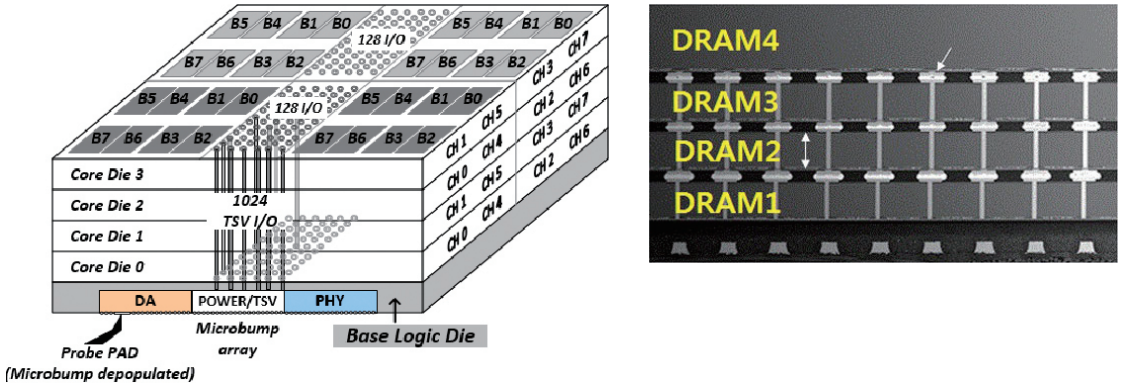


[그림 10] TSV 제조공정 (출처: 'Materials and Processing of TSV', Springer Nature)

TSV를 이용한 3D 패키지의 가장 큰 장점은 패키지 크기와 성능이다. [그림 9]에서 비교한 것처럼 와이어를 이용한 칩 적층에서는 적층된 칩의 옆면에 거미줄처럼 와이어들이 연결된 것을 볼 수 있다. 적층되는 칩의 개수가 많고, 연결할 핀(I/O) 수가 많을수록 와이어는 더욱 복잡해지며, 와이어를 연결할 공간도 많이 필요하다. 그러나 TSV를 이용한 칩 적층 사진을 보면 복잡한 와이어가 없으며, 그만큼 패키지의 크기를 줄일 수 있다.

TSV를 이용하여 칩을 적층한 패키지의 전기적 특성이 좋은 이유는 칩 간 전기 신호를 전달하고자 할 때 TSV를 이용해서 최단 거리로 신호가 전달되기 때문이다. 다시 말해 칩과 칩을 연결하는 핀을 원하는 위치에 형성하기 쉽고, 개수도 늘릴 수 있으며, 전기 신호 전달 경로가 짧아지기 때문이다. 보통 DRAM의 스펙에서 X8 이라고 표현된 것은 정보(Data)를 전달할 수 있는 핀의 개수가 8개라는 것을 의미한다. 즉, DRAM에서 동시에 내보낼 수 있는 정보가 8-bit이라는 뜻이다. (X16이면 16-bit, X32이면 32-bit.) 이 핀의 개수를 더 늘리면 더 많은 정보를 동시에 내보낼 수 있으므로 더 늘리고 싶지만, 와이어를 이용한 적층에서는 공정 미세화의 한계 때문에 X32가 최대

였다. 하지만 TSV를 이용한 적층에서는 공정 미세화로 이런 한계를 극복하였으며, HBM의 경우 X1024를 구현하였다. ([그림 11])

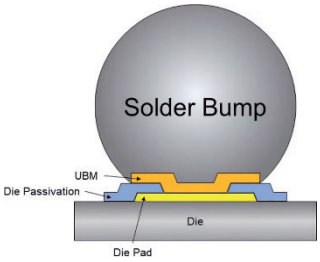


[그림 11] 8Gb HBM stacked DRAM architecture and micrograph with TSV. (출처: SK하이닉스)

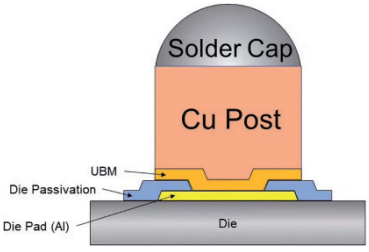
3. Solder bump 형성 및 Reflow 기술

HBM의 TSV가 칩 사이에서 전기적으로 연결되는 부분이 Solder bump인데, 칩 간의 본딩 접착제 역할을 하며 동시에 전기 신호의 입출력 단자 역할을 한다. Bump가 중요한 이유는 단위 면적당 입출력 단자 개수를 증가시키려면 Bump의 크기가 작아져야 하며, 단위 면적당 Bump 집적도를 높이는 것이 대역폭을 증가시키는데 가장 핵심이 되는 기술이기 때문이다. 현재 HBM에 사용되는 Bump는 20um 크기의 Micro bump이며 HBM 1개당 Bump의 수는 보통 5천개 수준이다. HBM과 함께 사용되는 Logic 칩도 2만개 정도의 Bump를 사용하므로, HBM이 4개 사용된다면 SiP 패키지 내 Micro-bump 수는 4만 개 정도가 된다.

Solder bump를 형성하는 전통적인 방법은 Metal pad 위에 구형의 Solder를 직접 올려 놓는 C4(Controlled Collapse Chip Connection) 기술이지만, Bump 크기 감소에 한계(75μm 이상)가 있다. 최근 HBM과 같은 미세 피치의 I/O에 적용하는 Micro-bump(100μm 이하) 형성에 사용되는 방법은 Cu 기둥을 만들어서 칩과 칩 사이의 간격(Bonding gap)을 높이고, 그 대신 Bump의 크기를 줄여 Bump 사이의 간격을 줄이는 CPB(Copper Post Bump 혹은 Copper Pillar Bump) 기술이다.



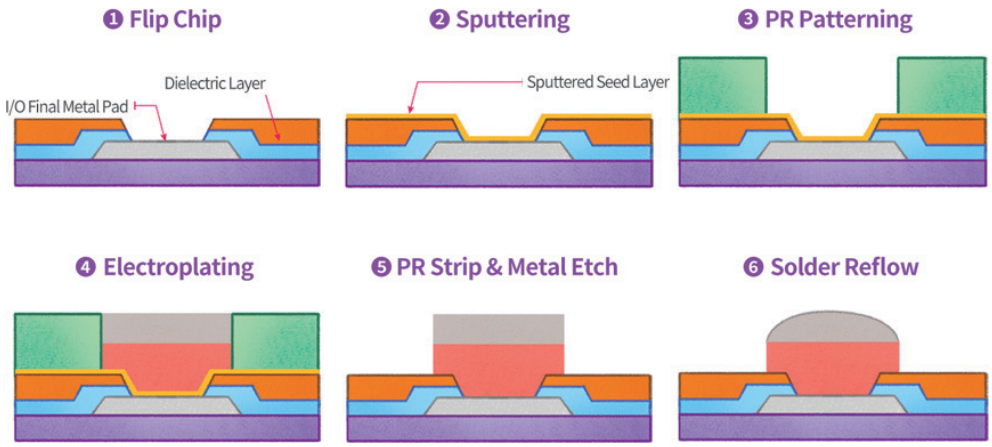
C4 bump(75~200μm)



Cu pillar bump(10~100μm)

[그림 12]는 SK하이닉스의 Wafer level packaging 기술을 이용한 CPB Micro-bump 형성 과정이다. 먼저 I/O pin이 형성될 Metal pad 위에 UBM(Under Bump Metallurgy) 박막을 증착하고(예를 들어 Adhesion layer

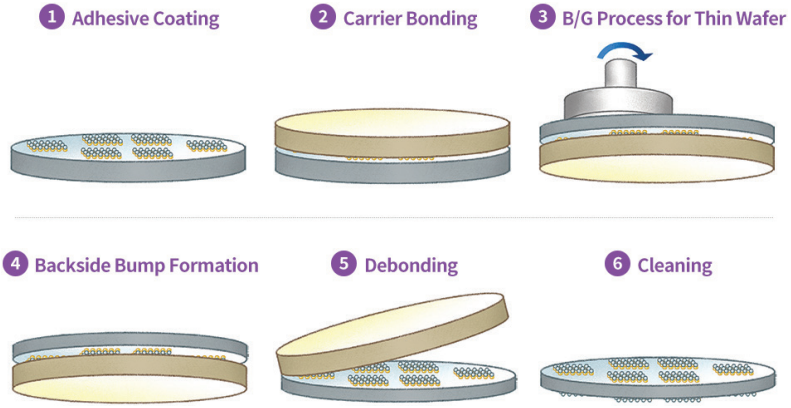
Ti / Seed layer Cu / Diffusion barrier Ni), 그 위에 다시 포토 레지스트를 도포하여 패턴을 형성한다. 그리고 전해도금으로 Solder bump를 만드는데, Cu를 도금한 뒤 다시 Solder로 사용될 물질을 도금한다. 도금이 완료되면 포토 레지스트를 제거하고, 잔류하는 UBM 박막을 Etching 공정으로 제거한다. 이후 열을 가해주는 Reflow 공정을 통해 Solder가 녹으며 Cu 기둥 위에 구형의 Bump가 형성된다. Reflow 작업이 필요한 이유는 범프 간 높이 차이를 최소화하고, Solder bump의 거칠기를 줄이며, Solder의 산화물을 제거하여 본딩 공정 시 접합성을 높이기 위해서다.



[그림 12] CPB Micro-bump 형성 공정

4. WSS(Wafer Supporting System) 기술 (출처: SKhynix newsroom)

[그림 13]과 같이 Core wafer의 Back-side thinning(grinding) 전에 Carrier wafer를 붙인 후 thinning 공정을 진행하여 얇아진 Core wafer의 Back grinding 된 면에 추가 공정이 가능할 수 있게 핸들링 하는 시스템을 의미한다. Core wafer에 Carrier를 붙이는 Carrier bonding 공정과 Core wafer에서 Carrier를 떼어내는 Carrier debonding 공정으로 구성되어있다. Carrier bonding은 가접착용 접착제를 웨이퍼 전면에 도포한 뒤 Carrier에 붙이는 공정이다. Carrier debonding은 웨이퍼 후면의 Bump 형성 공정이 완료된 후 Carrier를 떼어내고, 웨이퍼에 접착제 성분이 남아있지 않도록 세정하는 공정으로 이루어진다.

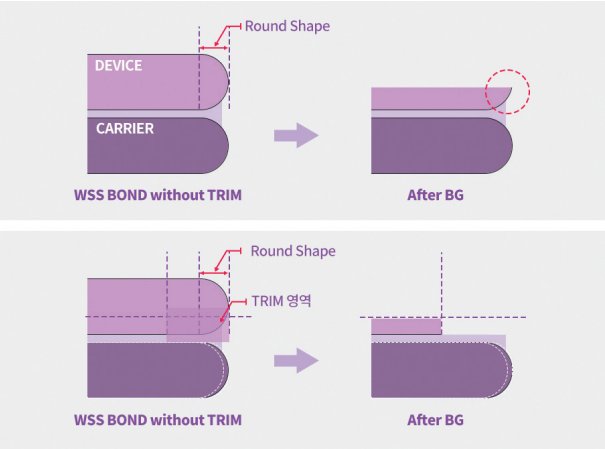


[그림 13] WSS 공정 순서

5. Wafer Edge Trimming 기술 (출처: SKhynix newsroom)

[그림 14]와 같이 Core wafer를 Carrier wafer와 본딩 후에 그대로 Back grinding 공정을 진행하면 TSV 패키지를 만들 웨이퍼는 그림의 우측 빨간 원으로 표시한 것처럼 가장자리가 날카로워진다. 이 상태에서는 웨이퍼 후면에 Bump를 형성하기 위한 포토 공정, 금속 박막 형성 공정, 전해 도금 공정 등 많은 후속 공정 진행 과정에서 가장자리가 깨질 위험이 커지며 수율에서 매우 큰 손실이 발생할 가능성이 높아진다.

따라서 이러한 문제를 해결하기 위해서 Carrier wafer와 본딩하기 전에 미리 TSV 패키지를 만들 웨이퍼의 전면 가장자리를 Trimming해서 제거한다. 이렇게 가장자리 쪽이 제거된 상태에서 Carrier wafer와 본딩하고 Back grinding을 진행하면 아래 그림처럼 웨이퍼 가장자리의 날카로운 영역이 사라지고, 후속으로 여러 공정을 진행해도 가장자리가 깨질 위험도 사라진다.



[그림 14] Wafer Edge Trimming 공정 순서

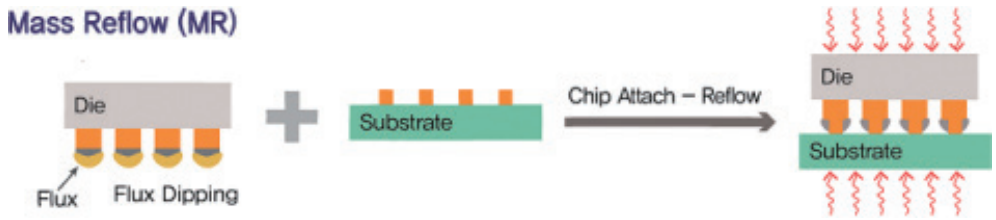
6. 적층(Stacking) 기술

HBM의 적층은 Chip to Wafer 혹은 Chip to Chip 방식이 가장 흔하게 사용되고 있으며 Wafer to Wafer 방식을 사용하지 않는 이유는 수율 때문이다. 예를 들어 TSV 공정을 마친 Wafer의 수율이 50% 정도라고 가정하고 이러한 Wafer 끼리 8단(수율: 0.5^{8}~0.39%), 12단(수율: 0.5^{12}~0.02%)으로 적층해버리면 양품의 HBM이 거의 나오지 않게 된다.

HBM의 적층은 이전 공정에서 형성해 두었던 하부 칩의 Front side Micro-bump solder를 녹인 후 상부 칩 뒷면에 형성되어 있는 Backside Micro-bump와 접합 시켜서 진행하는데, 이 때 가장 중요한 핵심 기술이 Bump 사이의 본딩 기술과 Bump 사이의 미세 공간을 채우는 Underfill 기술이다. HBM에 적용되는 대표적인 Bump bonding 및 Bump underfill 기술은 다음과 같다.

1) MR(Mass Reflow): [그림 15]와 같이 Bump가 형성된 칩을 기판 위에 수직으로 여러 개 정렬하여 안착시킨 후 한꺼번에 Reflow 장비를 통과 시키면서 칩 전체에 열을 가해 Solder가 녹아서 접합이 되게 하는 공정이다. 접합이 진행되는 동안 Bump 사이의 공간은 비어있는 상태이며 본딩이 모두 완료된 후 Underfill 공정을 진행한다. 속도가 빠른 대량 본딩 방식이지만 Bump의 크기와 간격이 작아지면서 Bridge 문제가 발생하고 칩의 두께가 얇

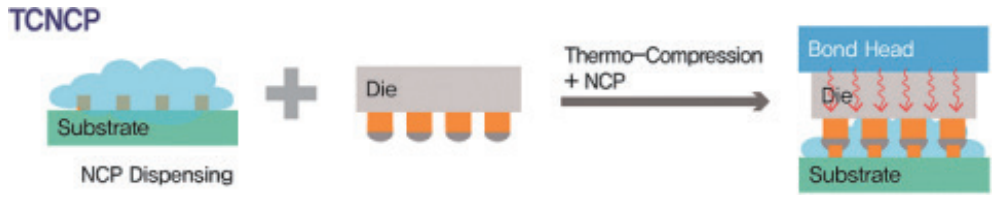
아지면서 열에 의한 휨 현상(Warpage) 등의 문제점이 발생하여 Micro-bump의 접합에는 일반적으로 Thermo compression 방식을 많이 사용한다.



[그림 15] Mass reflow 방식의 적층 기술 (출처: ECTC)

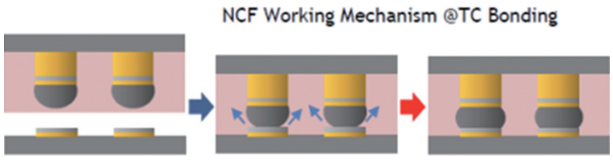
2) TC(Thermo Compression): MR 본딩의 기술적인 한계를 극복하기 위해서 사용하는 기술로 열과 압력을 가하여 Solder를 접합하는 공정이다. 적층하려는 칩과 칩 혹은 칩과 기판 사이에 절연 물질(Nonconductive paste 혹은 film)을 삽입하고 열과 압력을 가하여 최소한의 어긋남으로 Bump 간 연결을 진행하며 동시에 Underfill이 형성된다. Underfill 공정에 사용되는 소재에 따라 TC-NCP와 TC-NCF로 구분된다.

* TC-NCP([그림 16]) : 절연체와 접착제 역할을 수행하는 반죽 형태의 Paste를 칩과 칩 사이에 도포하고 열과 압력을 가하는 방식으로 Paste의 두께 조절이 어려워 이후 TC-NCF 방식으로 대체된다.



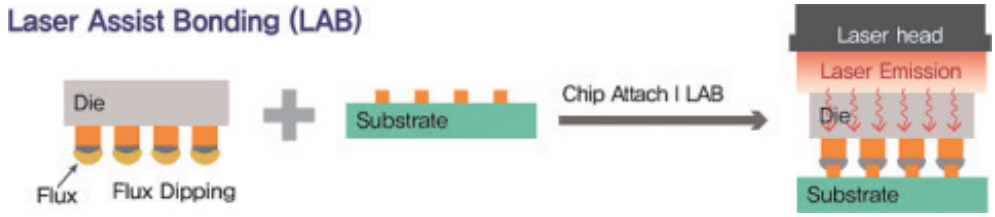
[그림 16] Thermo compression Nonconductive paste 방식의 적층 기술 (출처: ECTC)

* TC-NCF([그림 17]) : 절연체와 접착제 역할을 수행하는 Film을 칩과 칩 사이에 샌드위치 처럼 넣고 열과 압력을 가하는 방식으로 열과 압력에 의해 Solder가 녹아 접합되며 동시에 Film이 녹아 빈 공간을 채우면서 Underfill이 진행된다. 개별 칩 하나씩 본딩을 진행하여 적층을 쌓아 나가는 방식으로 여러 회사에서 HBM 제작의 표준처럼 사용되어 왔으며, 삼성전자는 최근에 출시된 HBM3E 까지도 적용하고 있다.



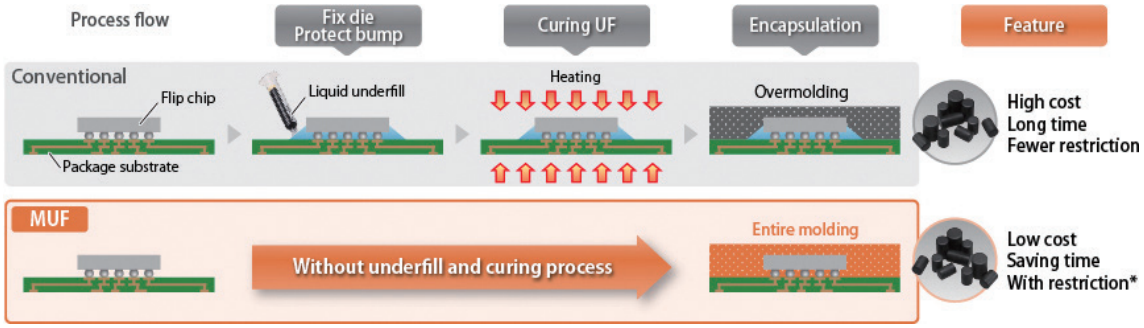
[그림 17] Thermo compression Nonconductive Film 방식의 적층 기술

3) LAB(Laser Assist Bonding) : [그림 18], 레이저로 1~2초간 칩의 접합 부분에만 열을 가해 본딩하는 방식으로 칩 전체에 고열을 가하는 MR이나 TC 본딩과 다르게 열이 가해지는 부분이 국부적이라 Thermal stress와 손상이 작다는 장점을 가지고 있다. 그러나 다이 하나씩 개별적으로 본딩을 진행해야 하는 관계로 양산성 확보에 어려움이 있다.



[그림 18] Laser Assist Bonding 방식의 적층 기술 (출처: ECTC)

4) MR-MUF(Mass Reflow-Molded UnderFill) : [그림 19]와 같이 먼저 여러 개의 칩을 고온에서 가 접합하여 적층한 후 Reflow를 통해 모든 칩을 한번에 접합하는 MR bonding을 진행하고, 이후 MUF 공정을 통해 칩 사이의 범프 간 공백을 채워주는 Underfill을 진행함과 동시에 Molding 공정까지 한번에 진행하는 공정이다. Underfill과 Molding을 동시에 진행하므로 대량 생산에 적합하고 SK하이닉스가 세계 최초로 개발하여 HBM2E 부터 적용하기 시작했으며 HBM3E 까지도 적용하고 있다.



[그림 19] Mass Reflow-Molded UnderFill 방식의 적층 기술 (출처: 신한투자증권)

HBM 개발 동향

1. 차세대 HBM 개발 로드맵

HBM의 최대 수요처인 Nvidia와 AMD가 차세대 인공지능 반도체 로드맵을 발표하면서 인공지능 반도체를 지원하는 HBM 시장에서 SK하이닉스, 삼성전자, 마이크론 등 메모리 3사 간 개발 경쟁이 더욱 치열해질 전망이다. 인공지능 반도체의 성능이 지속적으로 향상됨에 따라 HBM의 성능과 용량 요구도 높아지고 있기 때문이다.

Nvidia는 2025년 출시되는 '블랙웰 울트라'까지는 HBM3E 12단 8개를 탑재하고, 2026년 출시되는 '루빈' 플랫폼에 HBM4 8개를 처음으로 탑재하며, 2027년 출시되는 '루빈 울트라'에는 HBM4 12개를 탑재한다고 밝혔다. AMD는 올해 4분기 출시되는 'MI325X'에 HBM3E 12단 제품을 세계 최초로 탑재할 계획이며, 2025년 출시되는 'MI350'에도 HBM3E 12단 제품을 탑재하고, 2026년 출시되는 'MI400'에 HBM4를 탑재할 계획이다.

메모리 제조사들의 HBM 개발 로드맵을 살펴보면 3개 회사가 HBM 개발을 시작한 시기는 모두 다르지만, 최근 기술 격차가 많이 줄어들어 공통적으로 2025년 차세대 HBM4(6세대) 제품 개발 및 2026년 양산 계획을 가지고있는 것으로 추정된다. ([표 3])

	2020	2021	2022	2023	2024	2025	2026
삼성전자	HBM2e	HBM-PIM		HBM3	HBM3e (24Gb)	HBM4 Sample	HBM4
SK하이닉스	HBM2 (16Gb)		HBM3 (16Gb)		HBM3E (24Gb)"	HBM4 Sample	HBM4
Micron Technology		HBM2E			HBM3E (24Gb)	HBM4 Sample	HBM4

[표 3] 메모리 제조사의 HBM 개발 로드맵 (출처: 삼성증권)

JEDEC에서 올해 발표한 HBM4의 표준 스펙을 살펴보면 HBM3와 비교해서 층당 채널 수가 1024bit 에서 2048bit 으로 두배 증가하며, 적층은 최대 12단에서 16단으로 증가하여 DRAM의 용량이 최대 36Gb에서 48Gb으로 확장된다. 채널 당 전송 속도는 6.4Gbps로 초기 합의가 되었지만 이 속도는 시장의 요구로 8Gbps 이상으로 더 높아질 가능성이 크다. 적층 높이에 대해서는 회원사 간 이견이 있어서 두께 스펙을 기존 720um에서 775um로 완화 시키려는 논의가 활발하게 진행되어 왔으며, 최근 775um로 최종 확정한 것으로 알려져 있다.

1) Bandwidth의 변화

HBM3E까지 1,024개의 입출력 단자(I/O)를 통해 데이터의 전송이 이루어졌다면, HBM4에서는 I/O 수가 2,048개로 2배로 확대된다. 이를 위해 HBM4에서는 I/O 간 거리(Pitch)가 HBM3 대비 절반 수준으로 줄어들거나,

TSV 지름을 더 좁게 만드는 방식이 사용될 것으로 예상된다.

2) Core die(DRAM)의 변화

HBM3E에서는 DRAM 1a ~ 1b 공정을 사용했지만 HBM4에서는 성능 향상과 미세화를 위해 DRAM 1b ~ 1c 공정을 사용할 것으로 전망된다. ([표 4])

	HBM2e	HBM3	HBM3e	HBM4	HBM4e
삼성전자	DRAM 1y	DRAM 1z	DRAM 1a	DRAM 1c	DRAM 1c
SK하이닉스	DRAM 1y	DRAM 1z	DRAM 1b	DRAM 1b	DRAM 1c
Micron Technology	DRAM 1z	-	DRAM 1b	DRAM 1b	DRAM 1c

[표 4] HBM 세대별 Core die 생산 방식의 변화 (출처: 삼성증권)

3) Base die(Logic)의 변화

Base die는 Core die의 DRAM을 control 하는 기능을 담당하는 Logic die이며, HBM과 GPU를 연결하여 고속으로 데이터를 처리하는 데 기여하고 있다. 그동안 HBM 제조사는 DRAM에 사용하는 공정으로 Logic die를 양산해 왔으나, HBM4부터는 파운드리 공정이 신규 적용될 것으로 예상된다. 파운드리 공정을 적용하면, 선단공정에서 사용하는 다양한 Logic 기능을 추가할 수 있어서 성능이 향상되며 고객 맞춤형(Customizing) 반도체 제품의 공급이 가능해지기 때문이다. ([표 5])

	HBM2e	HBM3	HBM3e	HBM4	HBM4e
삼성전자	DRAM 1y	DRAM 1z	DRAM 1a	삼성파운드리 5nm	삼성파운드리 2nm
SK하이닉스	DRAM 1y	DRAM 1z	DRAM 1b	TSMC 12nm	TSMC 3/5nm
Micron Technology	DRAM 1z	-	DRAM 1b	DRAM 1b	TSMC 3nm

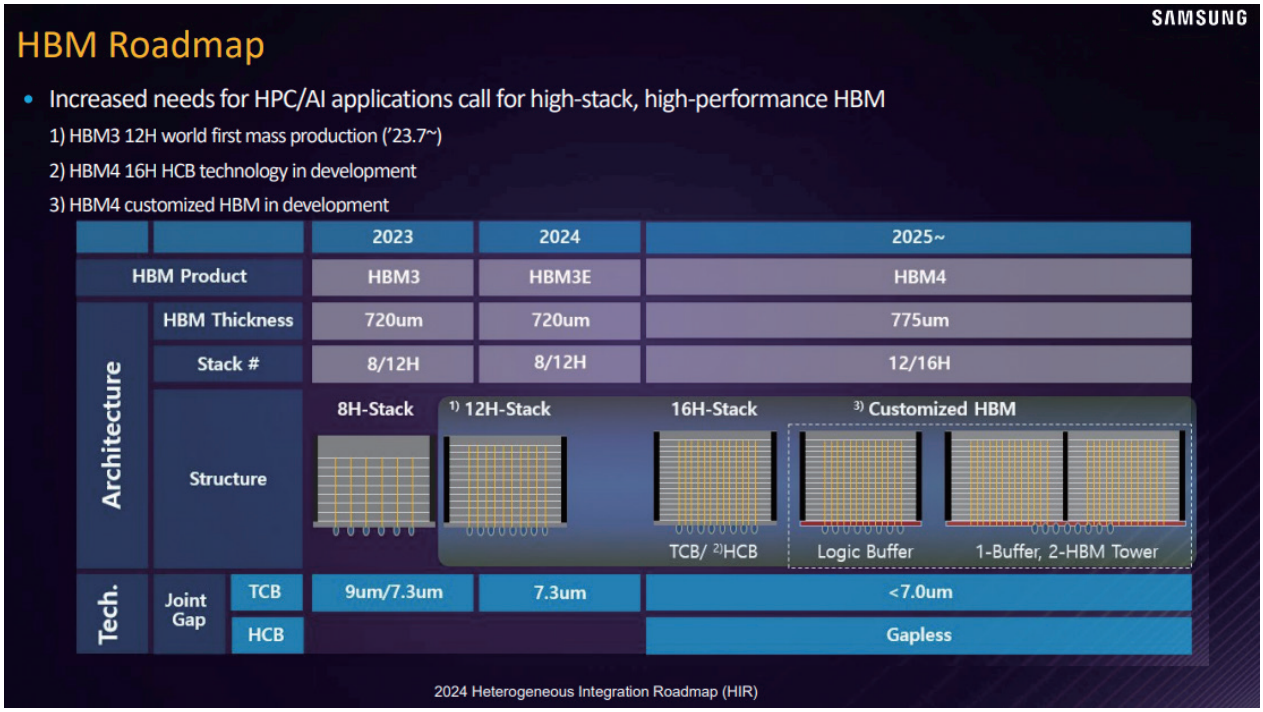
[표 5] HBM 세대별 Base die 생산 방식의 변화 (출처: 삼성증권)

Base(Logic) die의 변화로 인해 향후 파운드리와 메모리 업체 간의 협업이 더욱 중요해질 전망이며, 이에 Nvidia는 TSMC-SK하이닉스 연합과 제휴하고, AMD는 삼성전자와 협업을 강화하는 추세이다. HBM3E까지는 SK하이닉스가 D램과 Logic 다이를 자체 제작하고, TSMC가 이를 받아 기판 위에 GPU와 HBM을 나란히 조립(패키지)했지만, HBM4부터는 TSMC가 Logic 다이를 직접 제작하는 방식으로 바뀌고 HBM3E와 동일하게 2.5D 패키징이 유지된다. 삼성전자는 첨단 공정 파운드리와 메모리 사업을 동시에 공급하는 맞춤형 토탈 솔루션을 최대 강점으로 내세우고 있다. 이는 메모리에서 HBM을 만든 다음 자체 파운드리 팹에서 패키지까지 모두 가능하기 때문이다.

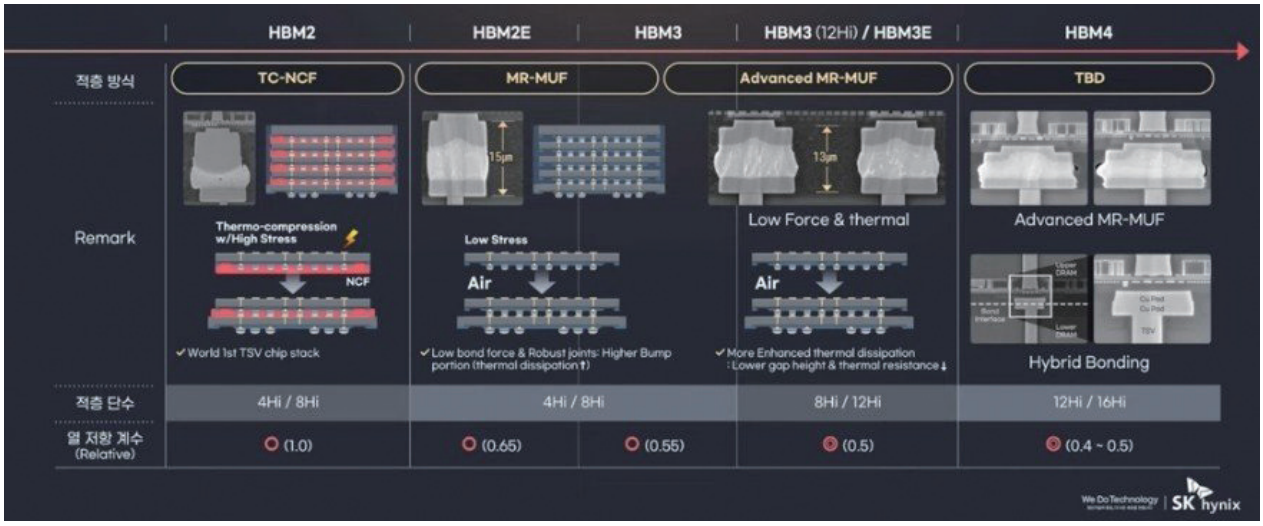
4) 적층 기술의 변화

HBM4에서 적층 단수가 16단으로 확대되기 때문에 기존의 두께 스펙(720um) 내에선 Micro-bump를 통한 본딩으로는 대응이 어렵다. 만약 12단으로 쌓는다면 한 칩 당 60μm의 공간이 확보되지만, 16단으로 쌓게 되면 기존보다 25% 감소한 45μm 공간이 주어지기 때문에 본딩에 필요한 두께를 최소화시키기 위해서 업계에서는 최근까지 Bump를 사용하지 않는 미세 본딩 기법인 하이브리드 본딩 적용에 대한 연구를 지속해왔다. DRAM을 더 많이 쌓으면서 높이(두께) 기준을 충족시키고 I/O 수를 두배로 증가시키기 위해서는 DRAM 층 사이에 위치한 수십 um 크기의 Micro-bump를 제거하는 것이 가장 효과적인 해결책이기 때문이다.

그러나 하이브리드 본딩은 높은 기술 개발 난이도와 낮은 경제성(신규 투자)으로 인해 현재 CIS, NAND 등 일부 제품에만 적용되고 있으며 단기간에 HBM에 요구되는 고밀도의 2.5D Packaging 혹은 3D Packaging에 적용하기에는 무리이다. 따라서 당분간은 HBM4에 대한 대비로 기존의 적층 기술(TC-NCF, MR-MUF)과 차세대 적층 기술(하이브리드 본딩) 적용에 대한 연구개발이 투 트랙으로 진행될 것으로 전망된다. ([그림 20], [그림 21])



[그림 20] 삼성전자의 HBM 적층기술 로드맵(2024년 발표)
(TCB: Thermo-Compression Bonding, HCB: Hybrid Cu Bonding)



[그림 21] SK하이닉스의 HBM 적층기술 로드맵(2023년 발표)

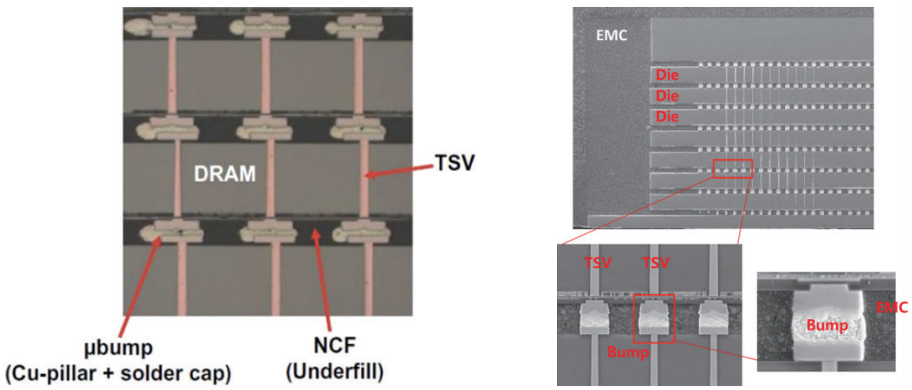
2. 차세대 HBM 적층 기술

HBM 적층 공정의 핵심 기술은 다음과 같다.

- 각 칩을 정확하게 정렬해서 1024개 이상의 TSV를 정밀하게 연결시키는 것
- 열과 압력에 의한 웨이퍼의 휨 현상(Warpage)과 깨지는 현상(Crack)을 방지하는 것
- 미세한 범프 사이 공간에 Underfill 소재가 균일하게 채워지도록 하는 것(Void, 방열 특성)
- 공정 시간의 단축으로 생산성을 확보하는 것

1) TC-NCF vs MR-MUF

[그림 22]와 같이 현재 HBM의 본딩 공정은 삼성전자, 마이크론의 TC-NCF 방식과 SK하이닉스의 MR-MUF 방식으로 양분되어 경쟁하고 있다.



삼성전자 TC-NCF　　SK하이닉스 MR-MUF

[그림 22] 적층 기술에 따른 HBM 단면 구조 비교

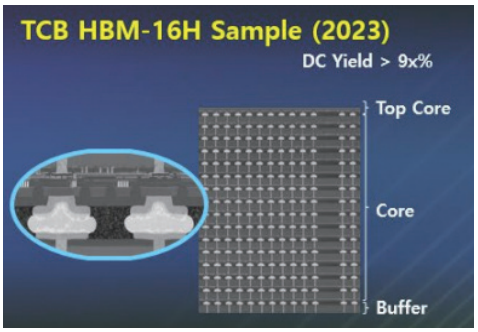
TC-NCF 방식의 장점은 ①개별 칩을 하나씩 쌓으면서 본딩과 Underfill을 동시에 진행하므로 칩의 Warpage 현상이 적고 칩 간 Align이 틀어지는 정도가 감소한다는 것과, ②NCF 필름 두께 감소에 의해 칩 간 Pitch를 줄이기에 용이하다는 것이다. 이는 HBM의 적층 단수 확대가 더 용이하다는 것을 의미한다.

TC-NCF 방식의 단점은 ①필름 소재의 균일한 물성 변화 조절이 어렵고(본딩 과정에서 열과 압력으로 필름이 녹아 칩 가장자리로 흐르는 Fillet 현상, 칩이 미끄러지는 Slip 현상 및 Bump crack 현상, 필름이 Bump 사이를 균일하게 채우지 못해서 발생하는 Void 및 이로 인한 방열 특성의 저하), ②개별 칩 별로 본딩 공정을 진행하는 관계로 시간이 오래 소요되어 생산성이 낮다는 것이다.

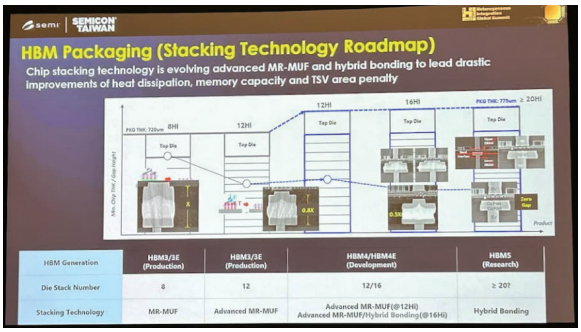
MR-MUF 방식의 장점은 ①모든 칩을 가 접합으로 적층하여 Reflow 장비에서 한번에 접합을 진행하고 이후 EMC 재료를 이용해 Underfill과 Molding 공정을 동시에 진행하므로 생산성이 매우 높다는 것과, ②액체 형태의 EMC 재료로 미세한 범프 사이 공간을 균일하게 채울 수 있어 Void 발생 가능성이 낮으며 방열 특성이 우수하다는 것이다.

MR-MUF 방식의 단점은 ①Underfill을 진행하지 않은 상태로 고온에서 장시간 Reflow 과정에서 Warpage 발생 및 Align이 틀어질 가능성이 높고, ②EMC 소재의 점도가 높아 적층 단수가 높아지거나 범프 간 Pitch가 감소하는 경우 모든 공간을 균일하게 채우기 어렵다는 것이다. 이는 HBM의 적층 단수 확대가 용이하지 않다는 것을 의미한다.

차세대 HBM4에서 적층 두께 스펙이 720um에서 775um로 완화되는 것이 확정되는 경우 TC-NCF, MR-MUF와 같은 기존 기술을 지속적으로 사용할 가능성이 높아지며, 실제로 최근 16단 적층에 기존 기술을 적용해서 유효성 있는 전기적 특성 결과들이 확보되는 것으로 발표되고 있다. ([그림 23])



삼성전자 TC-NCF



SK하이닉스 MR-MUF

- 삼성전자의 경우, NCF 소재 두께를 지속적으로 축소하여 최근 7um까지 축소한 것으로 알려져 있으며, 2024년 3월에 TC-NCF 방식으로 적층해서 동작이 확인된 16단 HBM sample을 발표했다.

- 도레이에서는 새로운 collective 결합 방식을 도입한 TCB-NCF 방법을 제안했는데, 기존 방식으로 4개의 칩을 쌓는데 40초가 소요된 반면, TCB-NCF 방법을 사용하면 14초 이내로 가능해지므로 생산성이 개선되는 것으로 발표했다.

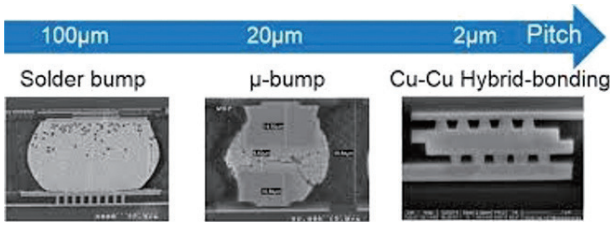
- SK하이닉스의 경우 Advanced MR-MUF 방식을 개발해 HBM3 부터 적용하고 있는데, ①가 접합 기술을 통해 칩이 휘는 문제를 해결하여 40% 얇은 칩 생산이 가능하고, ②신규 에폭시 밀봉재(EMC) 사용과 TC-NCF 대비 더 많은 수의 Thermal dummy bump를 적용해 열 방출을 극대화 시켰으며, ③Underfill과 Molding을 동시에 진행하여 생산성을 3배 정도 개선시킨 바 있다. 2024년 9월 대만에서 개최된 '세미콘 타이완'에서는 이 기술을 적용한 HBM3E 16단 제품 개발 결과를 발표하기도 했다.

2) 하이브리드 본딩(Hybrid bonding)

HBM의 차세대 적층 기술로 가장 큰 관심을 받고있는 핵심 기술이 하이브리드 본딩이다. '하이브리드(Hybrid)'라는 용어는 두 가지 유형의 계면 본딩(유전체 사이의 본딩, 금속 사이의 본딩)이 동시에 진행되는 것을 의미하며 HBM에서는 각 각 산화막과 구리가 사용된다.

하이브리드 본딩 기술은 Research Triangle Institute(이후 Ziptronic으로 분사)에서 처음 고안되어 산화막을 저온에서 플라즈마를 이용해 접합이 잘되는 상태로 바꾼 후 정렬해서 붙이고, 그 후 온도를 높여 Cu 패드를 붙이는 방식인 DBI(Direct Bond Interconnect) 기술로 발전하였다. 현재 소니를 중심으로 스마트폰과 이미지센서 등의 디바이스에 사용되는 CMOS 제작 과정에서 이 기술을 적용하고 있으며, 최근 YMTC가 Ziptronic의 DBI 기술을 사용하여 232단 3D 낸드(X3-9070)를 출시하기도 했다.

[그림 24]와 같이 HBM에 하이브리드 본딩 기술을 적용하면 Bump 접합을 이용한 기존 본딩 기술에 비해 다음과 같은 장점을 가질 수 있으며, 이로 인해 HBM4 이후에는 하이브리드 본딩 기술이 채택될 가능성이 매우 높다.



[그림 24] Solder bump vs Hybrid bonding

첫째, Bump를 사용하지 않고 본딩이 가능하여 본딩 층의 두께를 줄이고 전기 경로를 짧게 하여 저항을 낮출 수 있다. 이로 인해 마치 단일 칩처럼 성능 저하 없이 고속으로 작동할 수 있다.

둘째, Cu와 Cu를 직접 연결함으로써 Bump 간격을 획기적으로 줄일 수 있다. 보통 Solder를 사용할 때 Bump 간격을 10μm 이하로 구현하기 어렵지만, Cu-Cu 하이브리드 본딩의 경우에는 Bump 간격을 1μm 이하 수준으로 줄일 수 있어 I/O 밀도를 지속적으로 증가시키는 것이 가능하다.

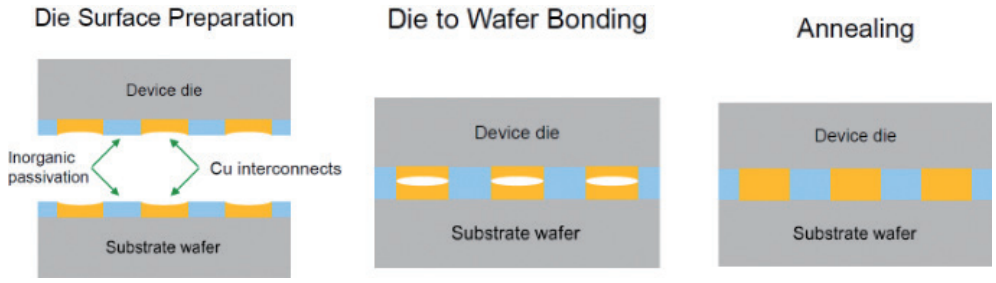
하이브리드 본딩 장비 시장 점유율 1위 업체인 BESI에서 제시하는 DBI 방식의 표준 Die to Wafer 하이브리드 본딩 공정의 순서는 [그림 25]와 같다.

A) Die surface preparation 단계

CMP를 통해 Oxide와 Cu를 평탄화 시키는데 Cu는 dishing을 nm 수준으로 정밀하게 조절해서 Oxide에 비해 약간 낮은 단차를 가지도록 한다. 이후 세정 공정으로 표면 오염과 Particle을 최대한 제거하고 플라즈마 전처리로 표면을 활성화 시킨다.

B) Die to wafer bonding(Pre-bonding) 단계

상온에서 Die와 Wafer를 15N 이하의 낮은 압력으로 눌러서 붙여준다. 이때 Van der Waals 인력에 의해 SiO2 분자 간에 수소 결합(Hydrogen bonding)이 발생한다.



[그림 25] 하이브리드 본딩 공정 순서도

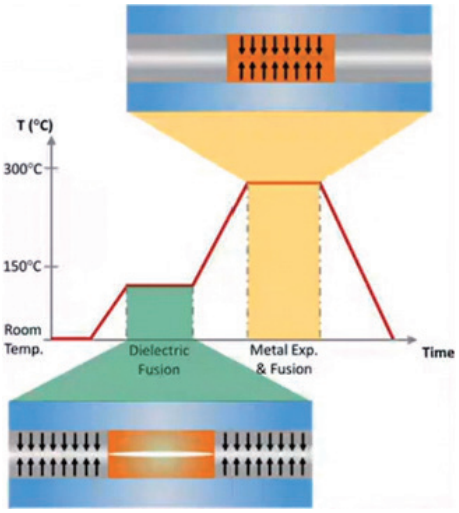
C) Annealing 단계([그림 26])

C-1) Low temperature annealing

이전 단계에서 형성된 수소 결합은 약한 분자간의 결합이기 때문에 이 결합을 더 강하게 만들기 위해서 저온(~150°C) 열처리를 진행한다. 이 과정에서 SiO2 간의 Interdiffusion이 발생하며 수소 결합 보다 더 강한 원자간 공유 결합이 형성된다.

C-2) High temperature annealing

이후 고온(~300°C) 열처리를 진행하면 Cu가 팽창하면서 CMP dishing으로 인한 간극을 채우고 상단의 Cu와 하단의 Cu가 맞닿게 된다. 그리고 Cu 간의 Interdiffusion이 발생하여 금속 접합이 진행되고 최종적으로 금속(Cu)과 유전체(SiO2)의 접착이 모두 완료된다.



[그림 26] Annealing 단계의 접합 과정

이러한 하이브리드 본딩 기술은 여러가지 장점에도 불구하고 HBM에 적용하기 위해서는 다음과 같은 문제점을 해결해야 하며, 2026년이 HBM4 양산 목표 시점이라면 기간 내 안정적인 본딩 기술 역량 확보까지는 시간적으로 부족하기 때문에 메모리 3사는 기존의 TC-

NCF 혹은 MR-MUF 방식과 병행해서 로드맵을 제시하고 있다.

- 기술 성숙도: 아직 업계의 기술 표준이 형성되지 않았고, 칩과 칩을 적층하면서 본딩시켜야 하기에 Multi Stacking에 대한 정교한 기술 개발이 추가로 필요하다. (10μm 이하의 미세 피치에서의 정렬도, 계면 접합의 신뢰성, 계면 Defects와 Void 발생에 의한 수율 저하, Uniformity 확보 등)

- 방열 기술: 하이브리드 본딩은 칩 간의 매우 밀접한 적층을 가능하게 하며, 이는 칩 당 발생하는 열의 양을 증가시킨다. 또한 전통적인 패키지 방식에서 사용되는 Underfill 재료가 제거됨에 따라 열 확산 경로가 제한되어 고밀도 패키지에서 열을 효과적으로 분산시키기 매우 어렵게 만든다. 이러한 발열 문제를 해결하기 위해서는 열 인터페이스 재료(TIM)의 사용이나 액체 냉각 시스템을 사용하는 등 능동적인 냉각 기술의 개발이 필요하다.

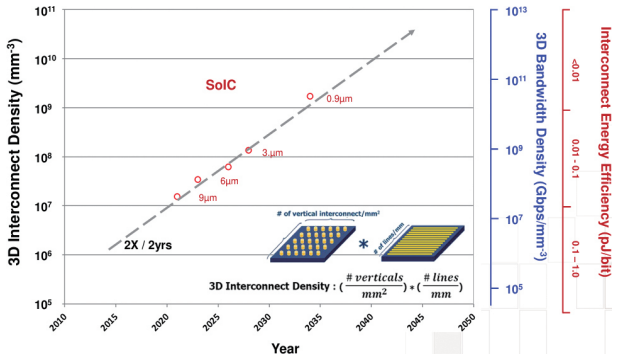
- 제조 비용: 전공정 기술이 융합된 기술이므로 신규 장비 투자 비용이 증가 한다.

3) 하이브리드 본딩 기술개발 동향

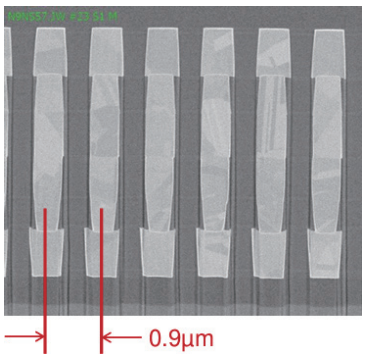
향 후 16단 혹은 20단 이상의 HBM4 개발을 위해서는 하이브리드 본딩 기술의 적용이 반드시 필요하므로 업계의 하이브리드 본딩 기술 개발이 매우 활발하게 진행되고 있으며, 주요 연구 동향은 다음과 같다.

A) TSMC

[그림 27]과 같이 하이브리드 본딩 기술 분야에서 가장 앞서 나가는 리더십을 보여주고 있으며, 2019년 3D Chiplet packaging 기술인 SoIC(System on Integrated Chips)에 하이브리드 본딩 기술을 도입해서(SoIC-X) AMD 등의 고객사에 양산 제품을 제공하고 있다. 2021년에 이미 9μm I/O pitch 수준을 구현하는데 성공하였고 2030년 이후 10분의 1 수준인 0.9μm I/O pitch 구현을 목표로 하고 있다.



TSMC의 Inter-chip Interconnect Scaling Roadmap



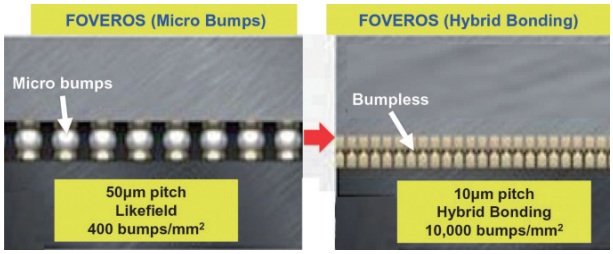
TSMC의 9μm I/O pitch Hybrid bonding

[그림 27] TSMC의 Hybrid bonding 기술

B) Intel

[그림 28] 2020년 3D Chiplet packaging 기술인 Foveros에 하이브리드 본딩을 적용한 Foveros Direct를

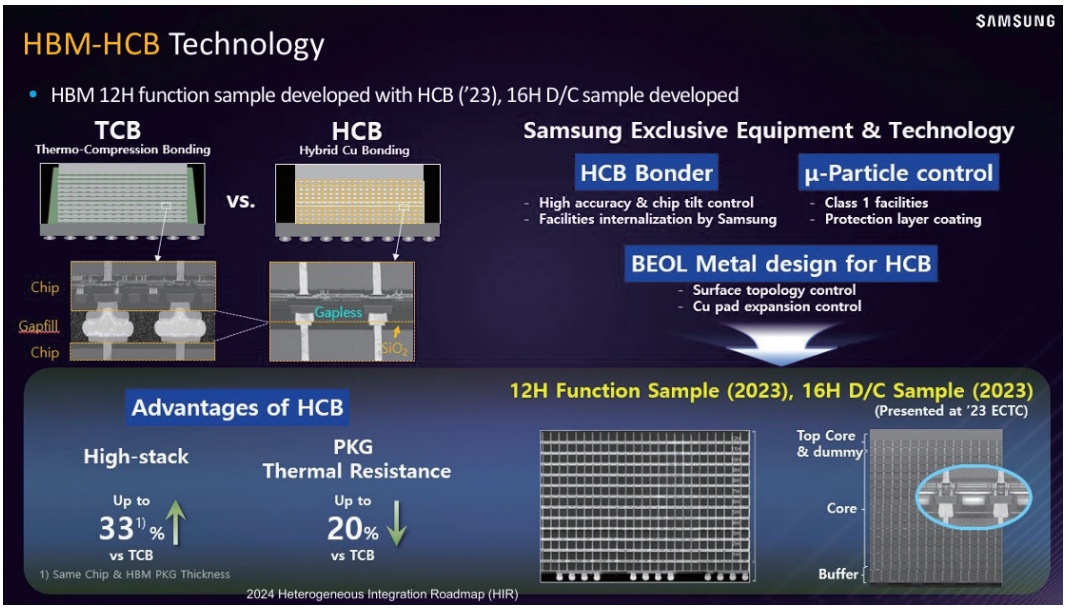
개발했으며 2023년까지 10μm I/O pitch 수준을 구현하고, 2025년 출시 예정인 Xeon Clearwater Forest CPU의 Core 칩 적층에 활용한다는 계획을 발표했다.



[그림 28] Intel의 Hybrid bonding 기술

C) 삼성전자

16단 이상의 HBM에서는 제한된 폼팩터(775μm) 안에 17개(베이스 다이 1개 + 코어 다이 16개) 칩을 적층하기 위해서 다이 두께 감소에 한계가 있어 하이브리드 본딩이 반드시 필요하다는 입장이며 매우 적극적으로 하이브리드 본딩 기술 개발에 투자하고 있다. 2023년 하이브리드 본딩 기술을 활용한 12단 HBM을 개발하여 정상 동작을 확인하였고, 이후 16단 HBM3 샘플 제작에 성공하였다. 세메스 본딩 장비를 사용해서 기술개발을 진행하고 있으며, 최근 HBM에서는 10μm I/O pitch를 구현했으며, 3D Logic 소자에서는 3μm 수준의 미세 pitch 구현에도 성공한 것으로 알려져 있다.

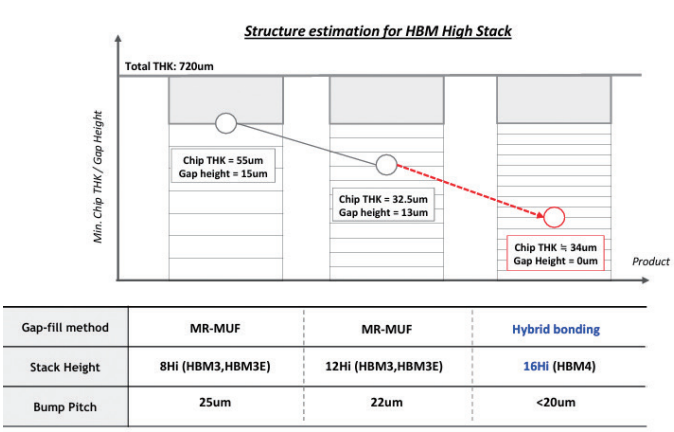


[그림 29] 삼성전자의 Hybrid bonding 기술(2024 Heterogeneous Integration Roadmap)

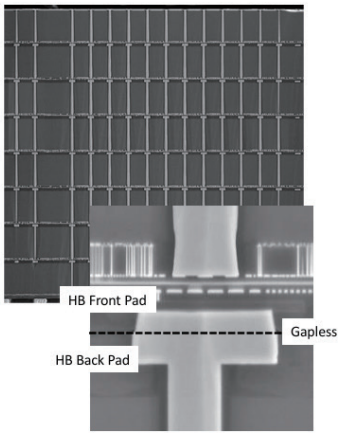
D) SK하이닉스

2022년 HBM2E에 하이브리드 본딩을 적용하여 8단 적층을 구현하고 전기적인 테스트까지 완료하여 기본적인 신뢰성을 확보한 바 있다. HBM4에서는 Advanced MR-MUF와 하이브리드 본딩 투 트랙 전략으로 대응하고 있지만, 올해 11월 부산에서 진행된 'ISMP-IRSP 2024'에서는 2026년 양산을 목표로 하는 7세대 HBM4E부터는 하이브리드 본딩을 필수로 적용한다는 계획을 발표했다.

16Hi 이상 HBM Solution으로 Gap-less Hybrid Bonding 기술 개발 중으로 HBM4 이후 채용 예상



Gap-fill method	MR-MUF	MR-MUF	Hybrid bonding
Stack Height	8Hi (HBM3,HBM3E)	12Hi (HBM3,HBM3E)	16Hi (HBM4)
Bump Pitch	25um	22um	<20um



[그림 30] SK하이닉스의 Hybrid bonding 기술(2023 반도체공학회)

시사점

- 현재 인공지능 반도체 시장을 주도하는 Nvidia, AMD 등 팹리스 업체에서 차세대 가성비 제품을 연달아 출시하여 인공지능 산업에서의 투자 비용이 상대적으로 저렴해지고 있으며 향후 Goggle, Microsoft 등 빅테크 기업들의 투자도 지속적으로 증가될 전망이다. (젠슨 황: '블랙웰에 대한 수요는 엄청나다', '모두가 최대한(물량을) 원하며 가장 먼저 받고 싶어 한다', 'HBM4 6개월 빨리 달라') HBM의 판매 단가는 일반 DRAM DDR5에 비해 5배 정도 비싸며, HBM이 전체 DRAM 매출에서 차지하는 비중은 2023년 8%에서 2024년에는 21%, 2025년에는 30% 이상으로 증가할 것으로 예상된다.

- HBM의 경쟁력을 좌우하는 가장 중요한 핵심 기술은 Chip stacking(bonding) 기술이며 삼성전자와 마이크론은 기존 방식인 TC-NCF 기술을, SK하이닉스는 독자적으로 개발한 Advanced MR-MUF 기술을 사용하고 있다. TC-NCF 기술은 다이 사이에 Film을 사용하므로 열에 의한 Warpage 현상이 적고, 층 간 Align 정확도가 높으며, 적층 단수를 높이는데 유리하다는 장점을 가지고 있다. MR-MUF 기술은 공정 단순화로 생산성이 좋고, 액상의 Underfill 물질이 Bump 사이의 빈 공간을 균일하게 채워 결함 발생이 적으며, 열 방출에 유리하다는 장점을 가지고 있다.

- 메모리 3사는 2025년에 출시되는 차세대 HBM4(6세대)에 적용될 적층 기술에 기존 방식인 TC-NCF/MR-MUF 방식과 하이브리드 본딩의 투 트랙 전략을 로드맵에 제시하고 있으며, 삼성전자는 하이브리드 본딩 기술에, SK하이닉스는 Advanced MR-MUF에 조금 더 큰 비중을 두고있는 양상이다. 그러나 2026년에 출시되는 HBM4E(7세대)에는 필수적으로 하이브리드 본딩을 적용해야 한다는 것이 업계의 공통 의견이다.

- HBM4 부터는 시장에서 고객 맞춤형(Customizing) 제품을 요구하고 있어서, 단순한 메모리 기능 외에도 AI 연산(In memory computing), 특정 데이터 처리 기능(IP) 등 고객사 요구에 맞는 다양한 Logic 기능을 제공해야 한다. 이는 결국 Base(Logic) 다이 제작을 위한 설계 역량과 파운드리 선단 공정기술을 확보해야 경쟁력을 가질수 있다는 것을 의미한다. 삼성전자의 경우 자체적으로 보유한 로직 설계 및 파운드리 공정 기술을 활용할 가능성이 있으며,(더 안정적인 공정 기술 확보를 위해 TSMC와의 제휴 가능성도 보도되고 있다.) SK하이닉스의 경우 Nvidia-TSMC와 연합할 가능성이 매우 높다.