
2023년 진공기술실무수련회교육
"메모리반도체 패러다임, HBM 기술"
한국알박㈜

반도체 메모리 패러다임, HBM

Introduction

- Memory Paradigm
- Market Trend

Hyper scaled Data Center

HBM (High bandwidth Memory)

- HBM 정의
- HBM History
- Fab기술 & Hardware

고성능, 고용량, 초고속 반도체

Advanced Packaging

- WLP 기술
- Future

AI

artificial intelligence

ULVAC

Memory Paradigm

Memory DRAM 수요

PC DRAM → Mobile DRAM → Server DRAM

PC	Mobile	Server & Network	Consumer (TV/Gaming/Automotive)
Desktop	Smartphone	Data Center	Set top box, TV
Laptop	Tablet, Wearable	Work Station	Gaming Device, Automotive

자료 SK-Hynix

Memory Paradigm

Memory 수요

PC DRAM → Mobile DRAM → **Server DRAM**

저장 ⇨ 유통

「데이터센터」 크게 증가

AI
IoT
Autonomous
(자율주행)

Data 실시간(real-time) 처리
(* = low latency(지연시간))

「서버용 D램」 폭발적 증가

구글, 메타(페이스북), 아마존 등
거대 IT업체

전세계에서 8천여개의 데이터센터를 운영

aws Microsoft Azure

AI solution
GPU + HBM (server)

Advanced Package

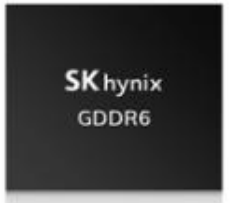
Chat GPT등 대화형 AI 진화
CPU → GPU 사용확대
→ 「GPU + **HBM**(메모리)」

ULVAC

Memory Paradigm

Server용 DRAM 종류 : DDR4 → DDR5, GDDR, HBM

DRAM 제품 라인업 및 응용처 (자료: Ryan, SK-Hynix)

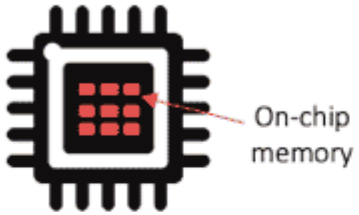
DDR (Double Date Rate)	LPDDR (Low Power DDR)	GDDR (Graphic DDR)	HBM (High Bandwidth Memory)
응용분야 1. 중-고급형/고용량 PC 제품 2. 서버 및 네트워크 3. 차량용 4. TV 및 가전제품	응용분야 1. 스마트폰/태블릿 등 모바일 2. 저사양/저용량 PC 제품 3. 차량용	응용분야 1. 고용량/고급형 PC 2. 서버 및 네트워크 3. TV 및 게이밍 기기	응용분야 1. 서버 및 네트워크
모듈 	단품 	단품 	특징 D램을 여러 개 쌓아 올려 한 번에 전송할 수 있는 데이터의 용량(메모리 대역폭)을 크게 늘림
단품 			단품 
		→ GPU	→ GPU

Memory Paradigm

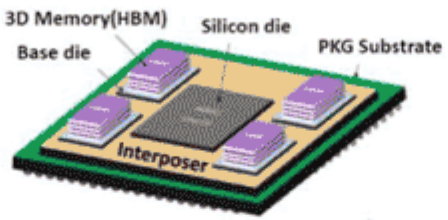
AI용 DRAM 종류 : GDDR(Graphic DDR), HBM (High Bandwidth Memory)

Common Memory Systems for AI Applications (자료: Rambus)

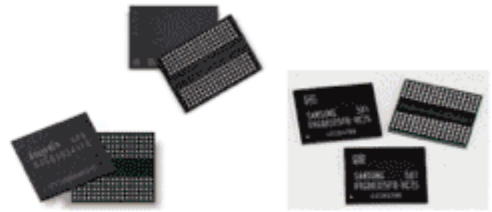
On-Chip Memory



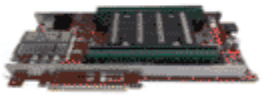
HBM



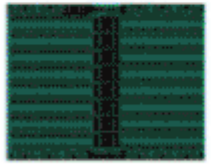
GDDR



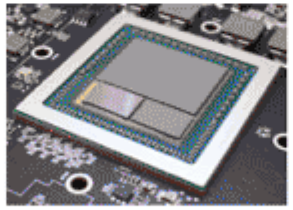
Microsoft BrainWave



Graphcore IPU



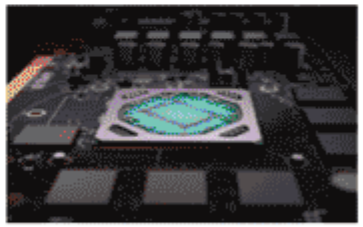
AMD Radeon RX Vega 56



nVidia Tesla V100

nVidia GeForce RTX 2080Ti

AMD Radeon RX580



Highest Bandwidth and Power Efficiency

Very High Bandwidth and Density

Good tradeoff between bandwidth, power efficiency, cost, and reliability

Multiple options suited to different needs

Memory Paradigm

Data Center = 인터넷데이터센터(IDC, Internet Data Center), 클라우드 데이터센터

ex) 페이스북, 구글, 애플, MS, 야후 등

'하이퍼 스케일(HyperScale)'급 데이터센터

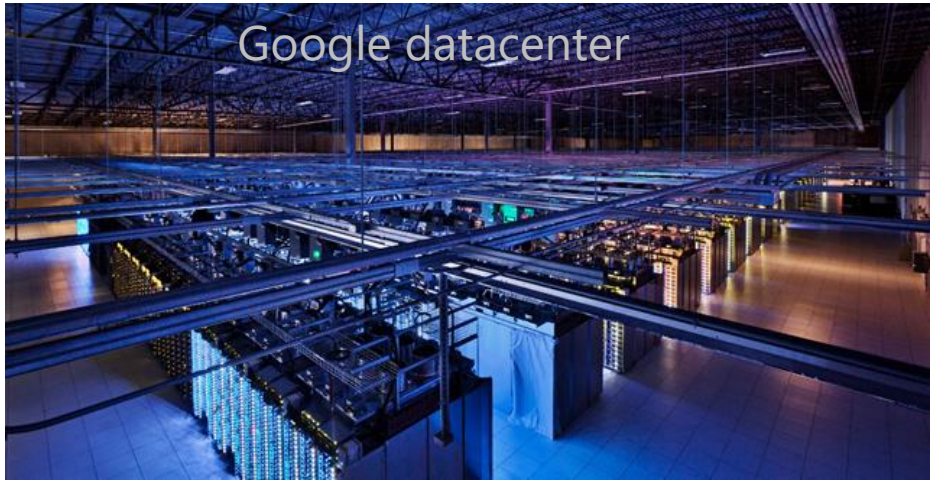
정의 : 연면적 2만2500㎡ 수준의 규모에 최소 10만대 이상의 서버를 갖춘 데이터센터

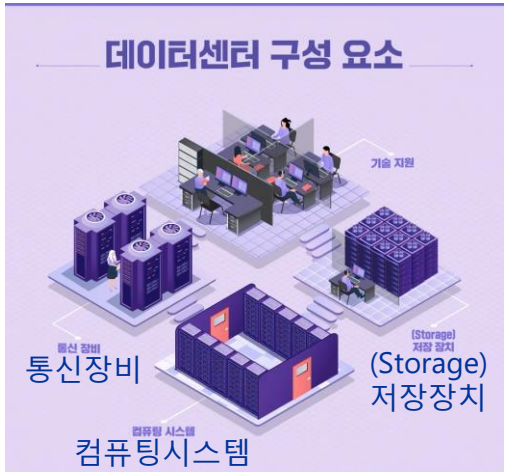
- 2021년 ~700 → 2023년 ~900개,

Server 랙(Rack)

※ 서버랙 캐비넷

중앙처리장치(CPU)나 램(RAM), 스토리지와 이들 장비가 장착될 메인보드, 전원 공급 장치(파워) 등, 하나의 캐비닛(랙, Rack)에 서버를 모으는 캐비넷





SK-hynix

HBM 정의

HBM
- High Bandwidth Memory

"고대역폭 메모리, 광대역폭 메모리"

HBM은 DRAM 칩을 **실리콘관통전극(TSV) 기술**을 적용해 **수직**으로 쌓아
데이터 처리 속도를 높인 메모리 반도체로,
AI 연산을 위한 **그래픽처리장치(GPU)에 탑재**

메모리(DRAM)와 처리장치(CPU, GPU 등) 간에
데이터가 오고 가는 **통로(BUS)의 「폭(bandwidth)」**을 넓힘

- **Bandwidth : 프로세서가** 솔리드 스테이트
  메모리에서 데이터를 **읽거나** 데이터를 **저장할** 수 있는 **속도**

높은 대역폭 및 속도 확보방법:
- 기존 D램보다 더 많은 **데이터 전송 통로(I/O)를 확보**해 한번에 많은 양의 데이터를 전송
- **TSV** (Through Silicon Via, 실리콘 관통 전극) **배선이 데이터 입출입 통로 역할**

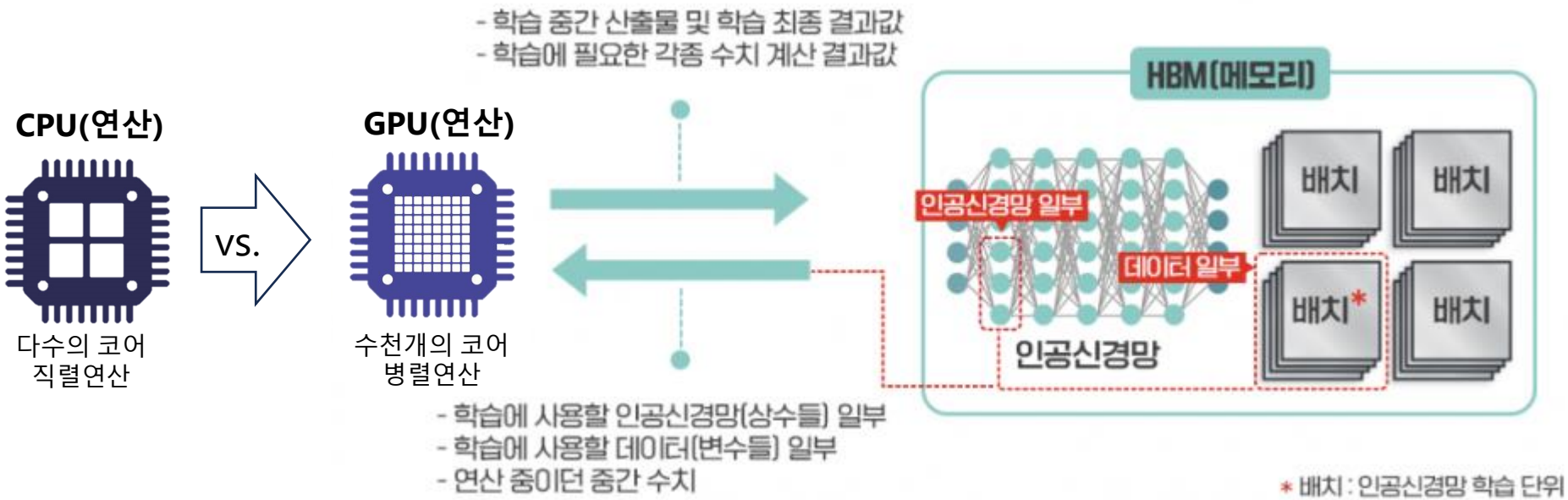
HBM 정의

HBM & GPU (AI 가속기) 역할
- High Bandwidth Memory

"**고대역폭** 메모리, **광대역폭** 메모리"

-GPU는 메모리에 저장된 인공신경망 및 데이터 일부를 지속적으로 가져와 연산(학습 및 추론)하고,

-중간 산출물과 최종 결과 등을 **HBM 메모리**에 저장하는 과정을 반복한다.



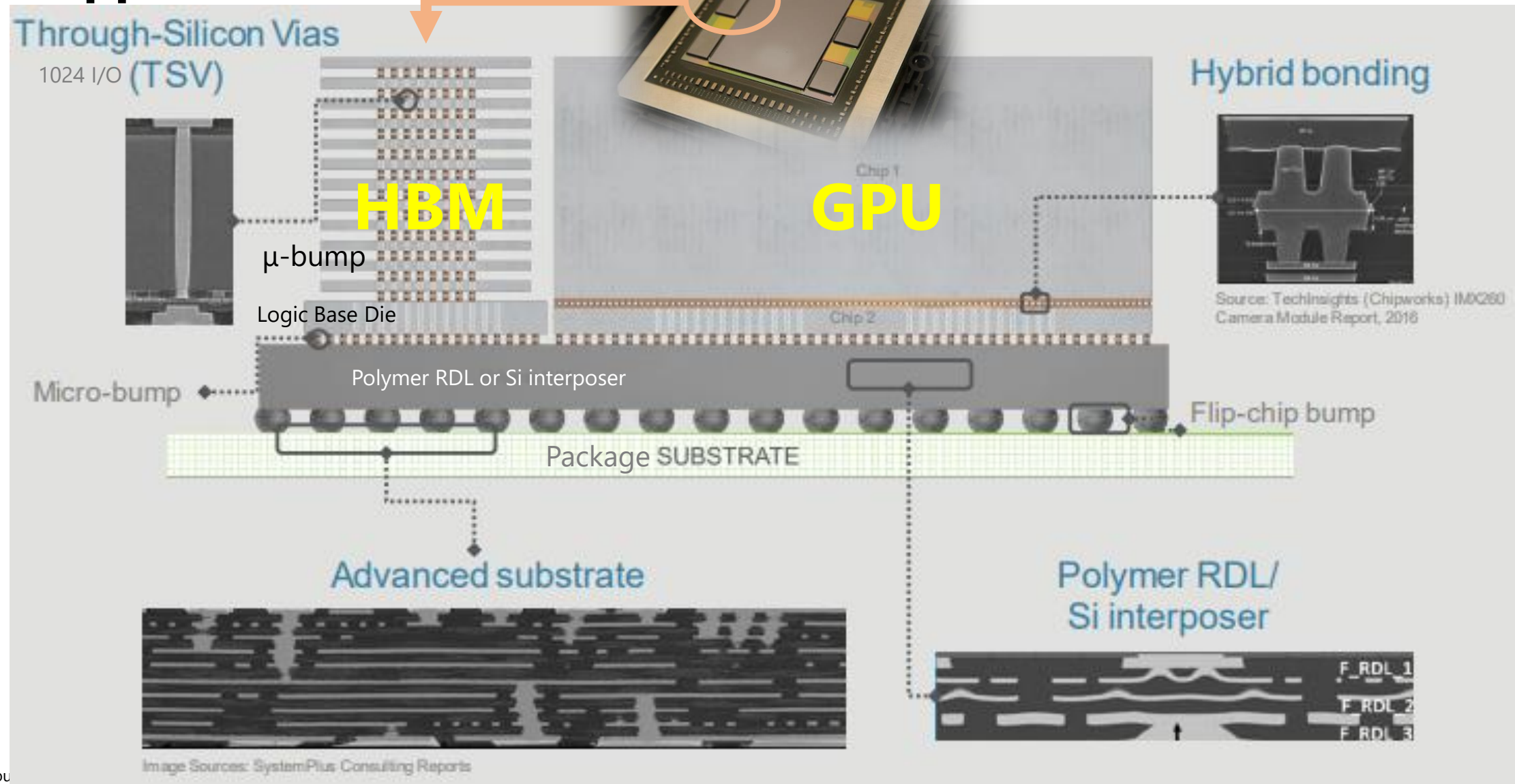
* SK-hynix 자료

* 배치 : 인공신경망 학습 단위
학습 과정의 데이터에 사용되는 메모리

* 추론 : 인공신경망이 실제로 문제를 푸는 행위

HBM 정의

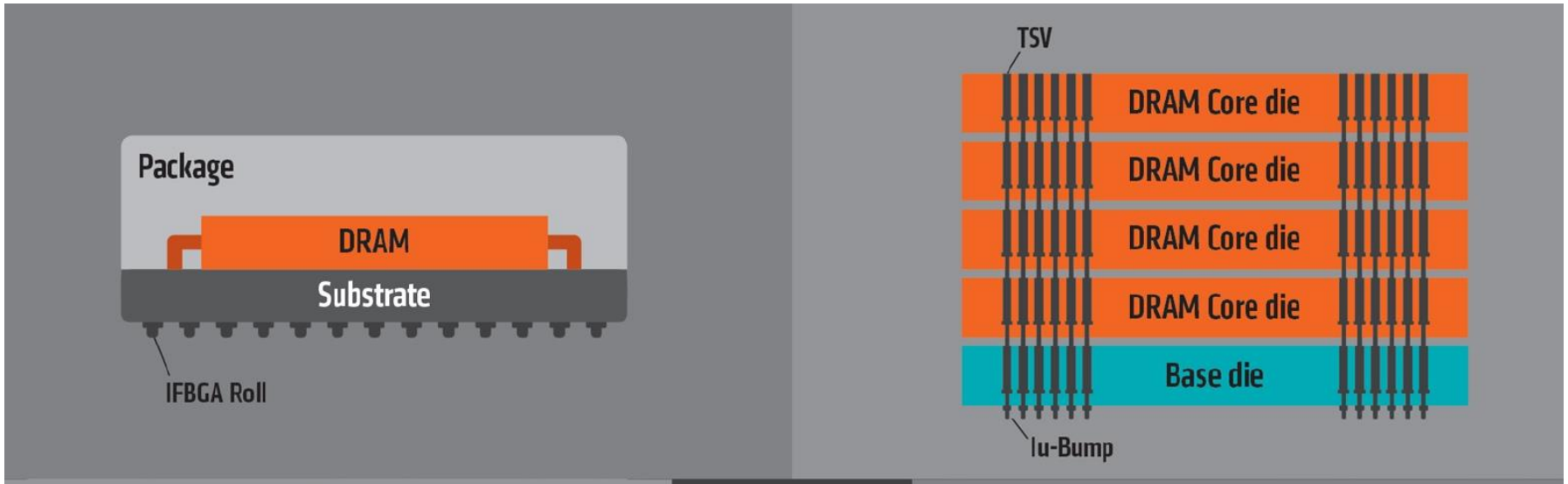
HBM Application, GPU



HBM 정의

장점(1): High Bandwidth

- HBM은 스택모듈 하나당 **1024-bit**의 데이터 입출력이 가능(TSV기술)
- 기존 GDDR5 메모리 모듈이 하나당 **32-bit** 입출력(wire bonding)을 담당



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

자료: AMD

HBM 정의

장점(1): High Bandwidth

Samsung HBM Memory Generations					1st HBM	compare
	HBM3E (Shinebolt)	**HBM3** (Icebolt)	**HBM2E** (Flashbolt)	**HBM2** (Aquabolt)	4Hi-HBM1 (SK-Hynix)	**GDDR5** (SK-Hynix)
Max Capacity	36GB	24 GB	16 GB	8 GB	1GB	
Max Bandwidth Per Pin *(Speed)*	**9.8 Gb/s**	**6.4 Gb/s**	**3.6 Gb/s**	**2.0 Gb/s**	**1.0 Gb/s**	**8.0 Gb/s**
Number of DRAM ICs per Stack	12	12	8	8	4	
Effective Bus Width (I/O)	1024-bit				1024-bit	32
Voltage	?	1.1 V	1.2 V	1.2 V	1.2V	1.5V
Bandwidth per Stack	**1.225 TB/s**	**819.2 GB/s**	**460.8 GB/s**	**256 GB/s**	**128 GB/s**	**32 GB/s**

D1a (14nm) node.

D1z (16nm) node.

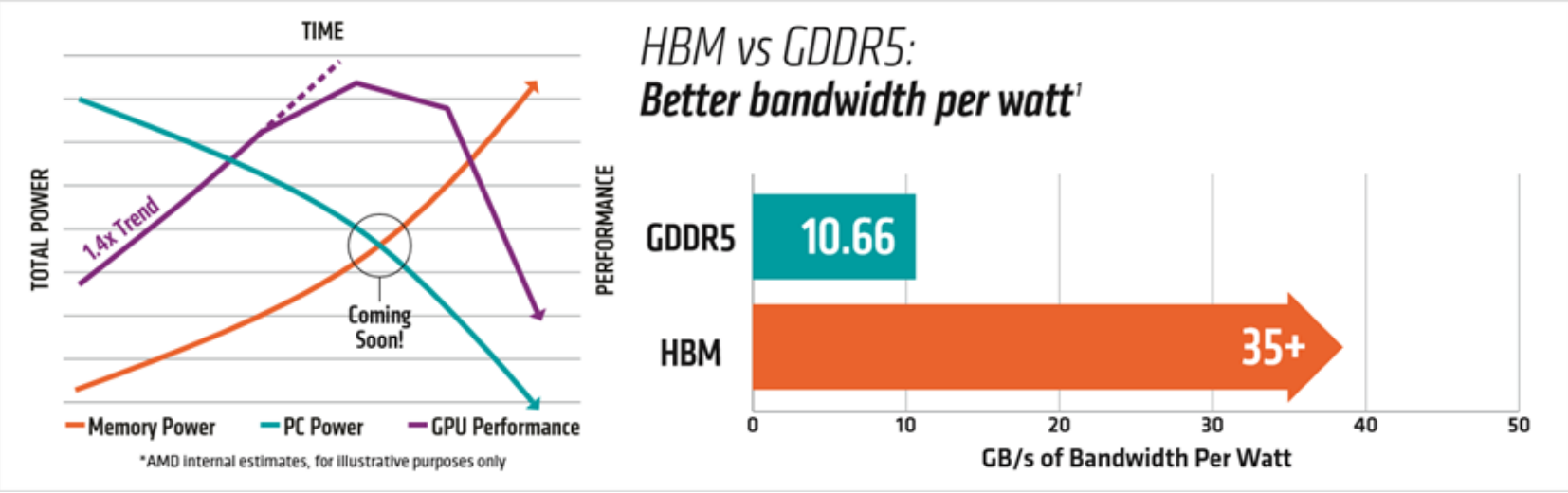
ULVAC

HBM 정의

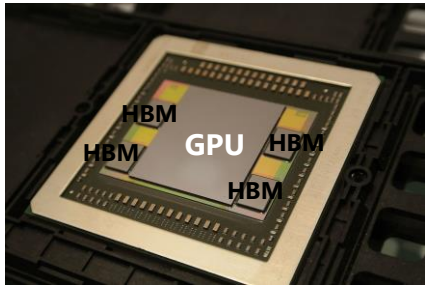
장점(2): 전력 효율 극대화

- **GDDR5** 메모리: 모듈 하나당 **32-bit 입출력(Pin수)** 담당
- GPU 코어의 메모리 컨트롤러에 따라 최소 **4개**(128-bit), **8개**(256-bit), **16개**(512-bit) 모듈 탑재
- **모듈마다 1.5V** 전압 필요, 개수가 늘어날수록 전력소비량 비례 증가

- **HBM**: 스택 모듈당 **1024-bit 입출력** 가능, 스택 모듈 **4개**를 GPU에 연결시 **4096-bit**
- 전압 **1.3V**로 낮음. 그래픽 카드 전력소비량 감소 가능



엔비디아 '볼타' GPU와 주변에 장착된 12개 GDDR 메모리

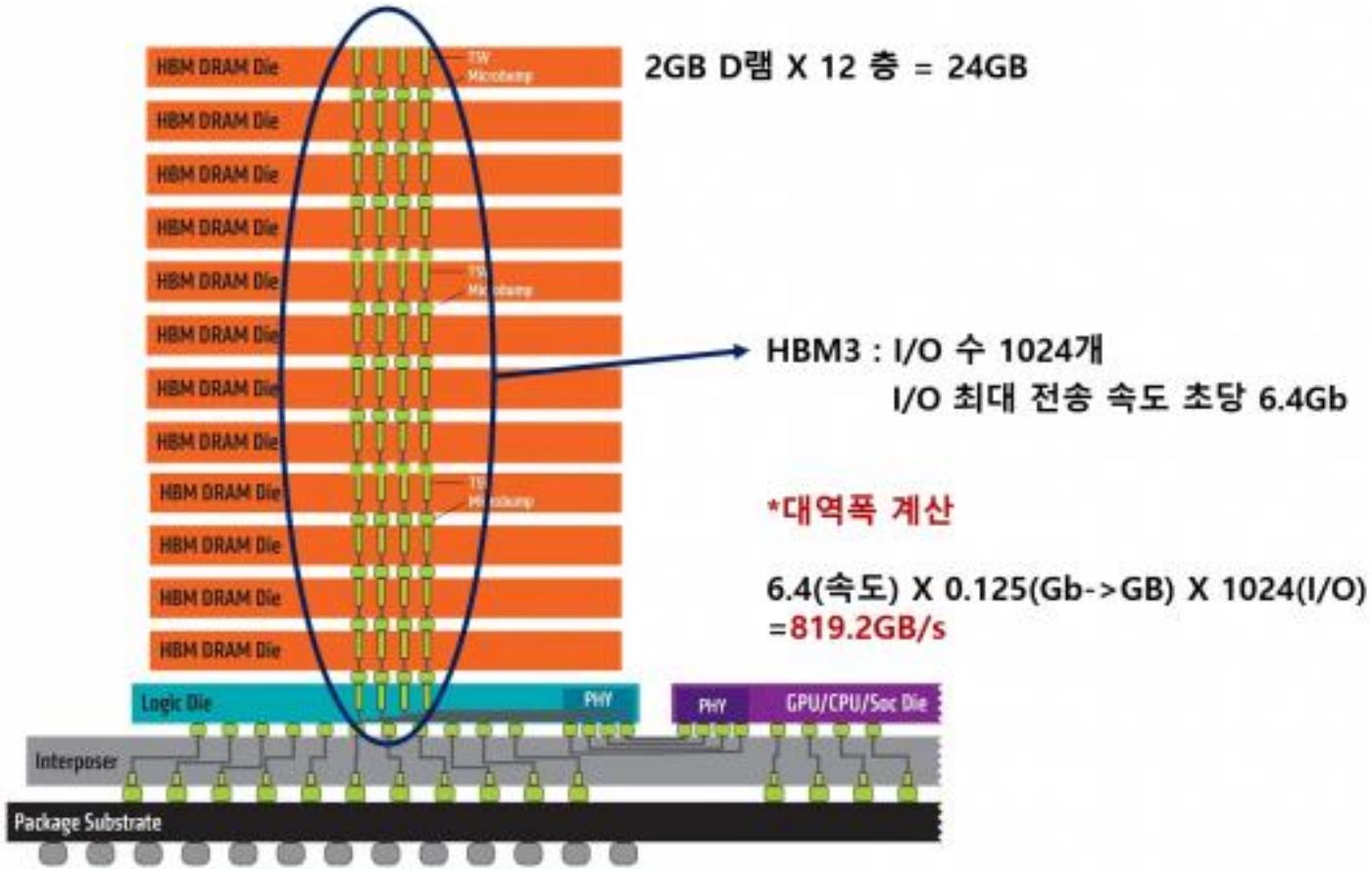


ULVAC

HBM 정의

"HBM3 12-Hi stack“ @SK-Hynix

("Hi"는 "High"의 축약어, HBM 스택의 높이 또는 레이어 수)

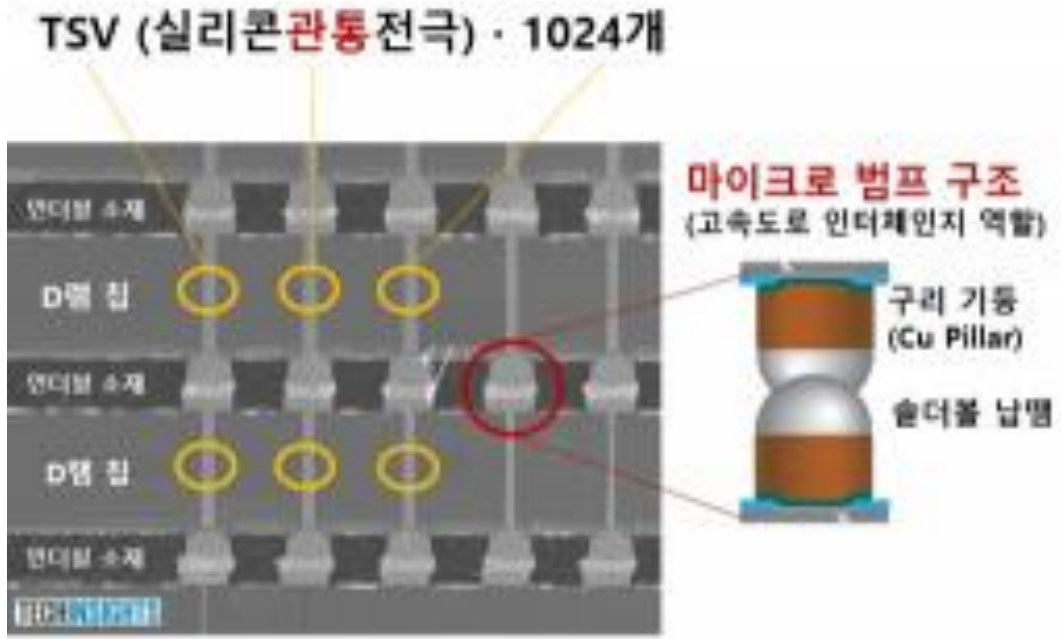


2GB D램 X 12 층 = 24GB

HBM3 : I/O 수 1024개
I/O 최대 전송 속도 초당 6.4Gb

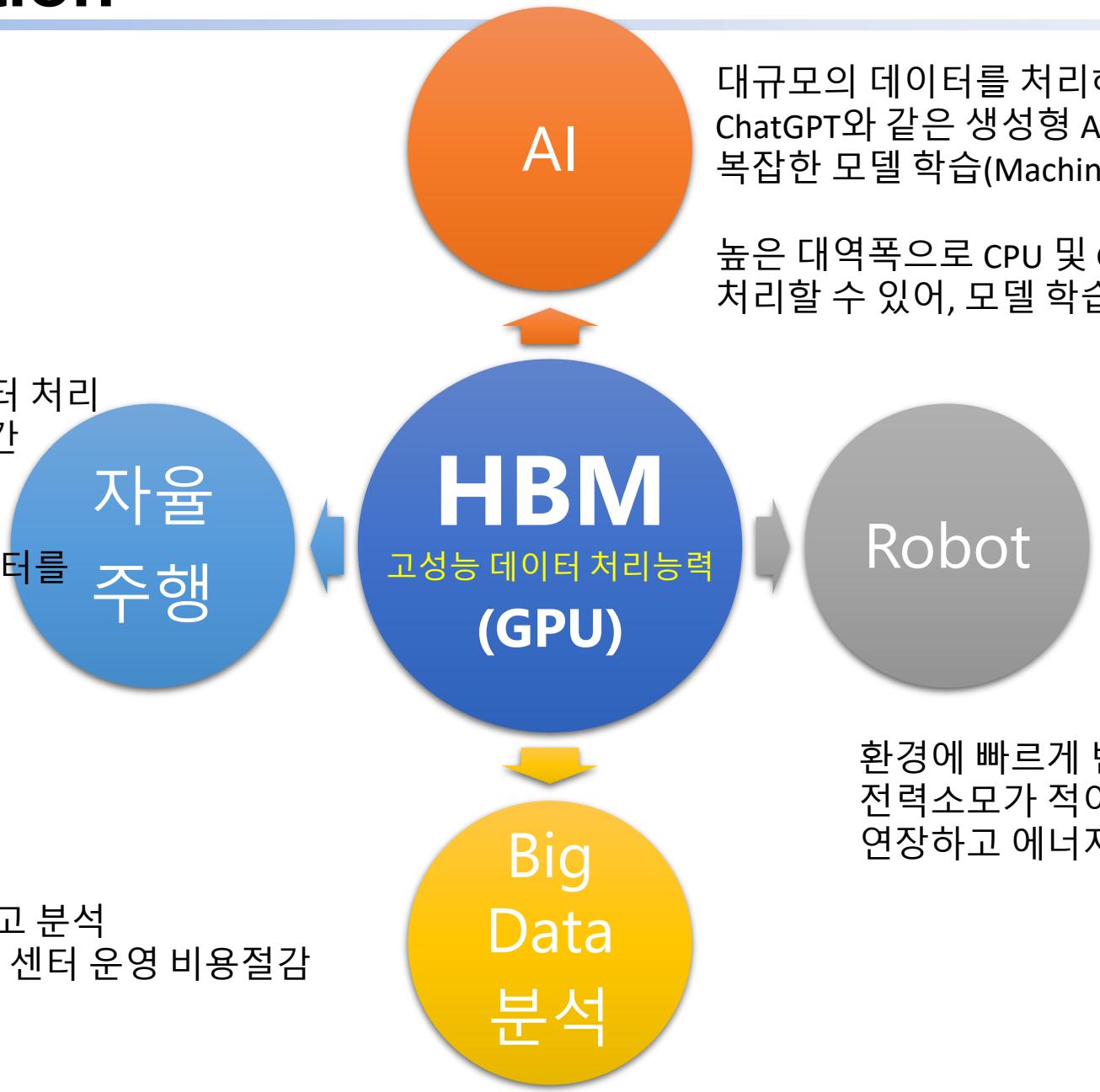
*대역폭 계산

6.4(속도) X 0.125(Gb->GB) X 1024(I/O)
=819.2GB/s



HBM Application

HBM 응용



대규모의 데이터를 처리해야 하는 AI 산업
ChatGPT와 같은 생성형 AI를 개발
복잡한 모델 학습(Machine learning)

높은 대역폭으로 CPU 및 GPU 간의 데이터 이동을 빠르게 처리할 수 있어, 모델 학습 시간을 단축 및 성능 향상

카메라, 레이더 등 센서 데이터 처리 및 위치 인식 등을 통한 실시간 의사결정(AI Inference)

자율주행에 탑재된 AI는 데이터를 빠르게 처리하여 실시간으로 추론해야

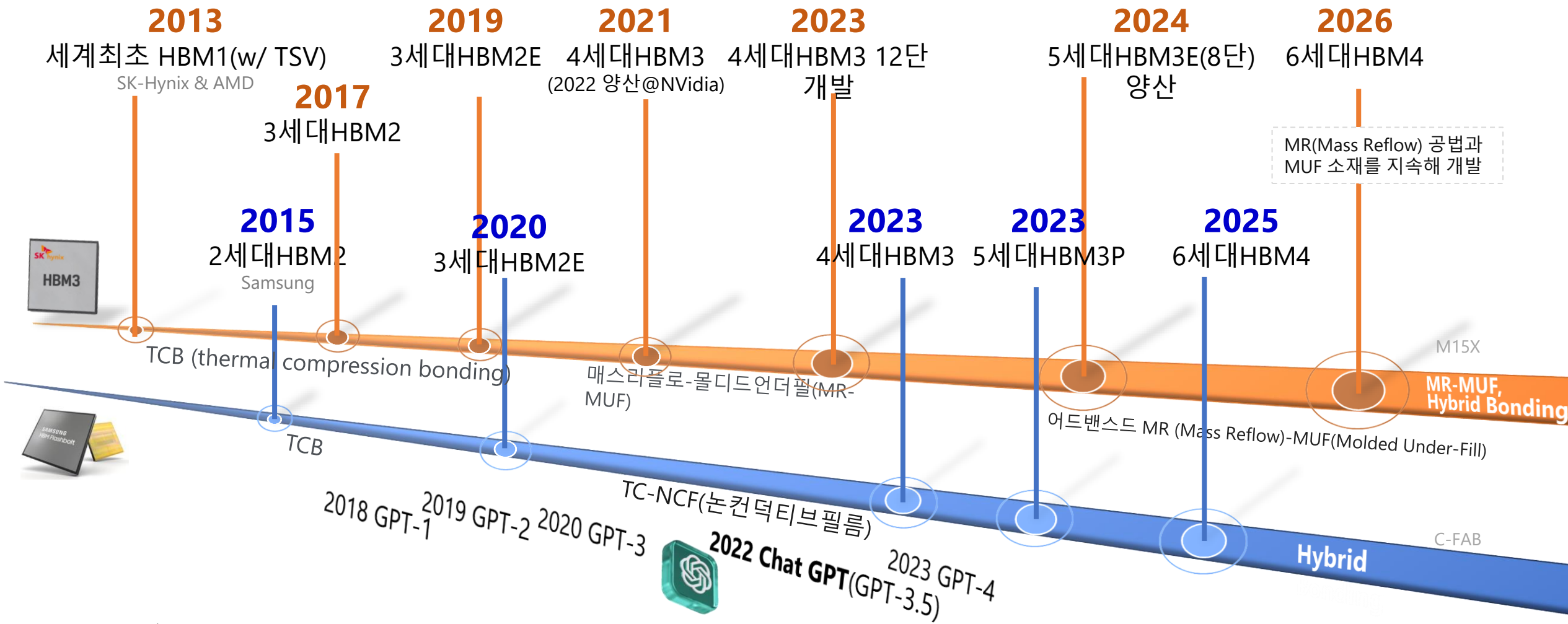
환경에 빠르게 반응하며 작업을 수행
전력소모가 적어 로봇의 작업시간을 연장하고 에너지 효율성 향상

대량의 데이터를 처리하고 분석
낮은 전력사용. 빅데이터 센터 운영 비용절감

HBM History

HBM 개발/양산 History

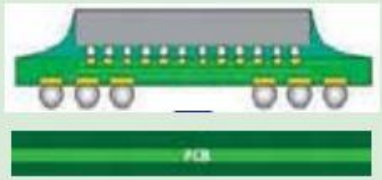
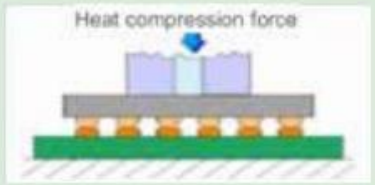
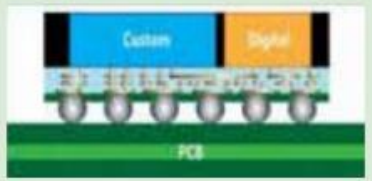
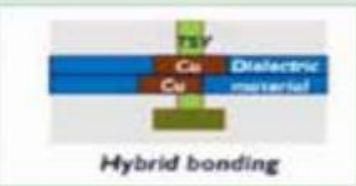
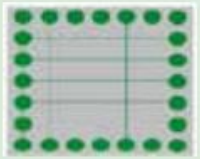
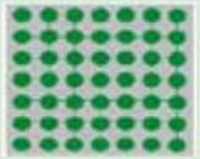
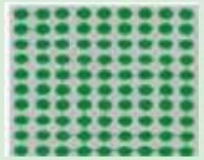
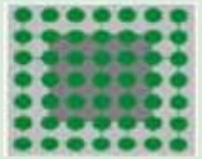
고용량 데이터 구현 → 코어 다이(Core Die) 용량 증대
→ 적층 수/높이 확장 방법 연구



HBM 제작 Fab 기술

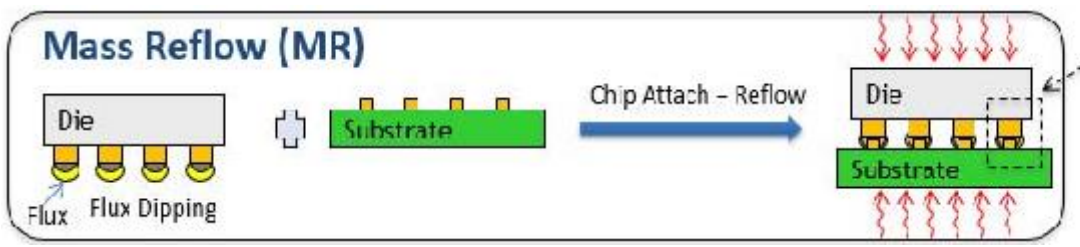
Bonding 기술 History

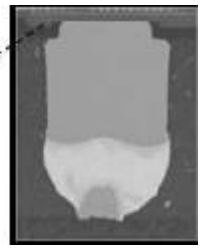
DIE BONDING PROCESS EVOLUTION

	Wire Bond (1975)	Flip Chip (1995)	TCB Bonding (2012)	HD Fan Out (2015)	Hybrid Bonding (2018)
Architecture					
Contact Type	 Wire	 Solder ball or copper pillar	 Copper pillar	 RDL or copper pillar	 Copper-to-copper
Contact Density	 5-10/mm²	 25-400/mm²	 156-625/mm²	 500+/mm²	 10K-1MM/mm²
Substrate	Organic/leadframe	Organic/leadframe	Organic/Silicon	None	None
Accuracy	20-10μm	10-5μm	5-1μm	5-1μm	0.5-0.1μm
Energy/Bit	10pJ/bit	0.5pJ/bit	0.1pJ/bit	0.5pJ/bit	<.05pJ/bit

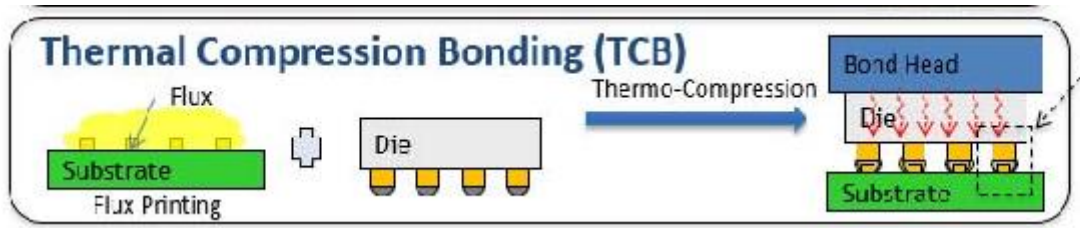
HBM 제작 Fab 기술

Bonding 종류

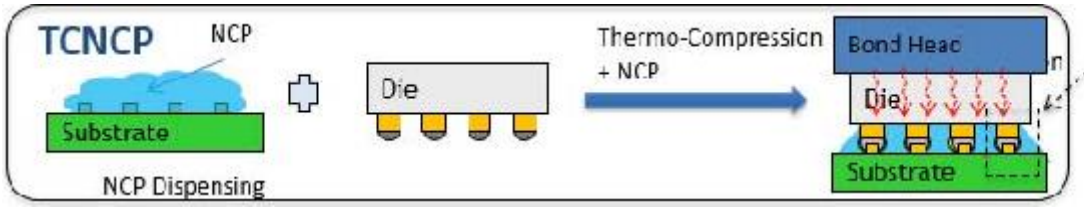




- **MR**: 플립칩이 컨베이어 벨트로 리플로우 장비를 지나가면서 본딩, 대량 본딩방식,
  (장점) 속도가 빠름,
  (단점) 마이크로 범프엔 부적합
- **MR-MUF: SK하이닉스 HBM2E부터 적용(세계 최초공법)**
  여러 칩을 리플로우를 통해 한 번에 접합 후 칩 사이를 금형재료 (실리카+에폭시)로 동시에 채우는 방식



- **TC본딩:** 열과 압력을 이용해 본딩
  범프소재별 열과 압력 달리 적용
  1) 본딩을 먼저 진행 후 언더필을 하는 기본방법
  2) NCP를 활용한 TC-NCP 공법



- **TC-NCP 본딩**: NCP (Non-Conductive Paste) 활용 TC본딩
  NCP는 접착제와 몰딩역할을 하는 소재
  NCP를 먼저 도포하고 열과 압력을 가하면 본딩 및 언더필이 생성되는 원리,
  NCP는 Paste다 보니 두께 조절이 어려워

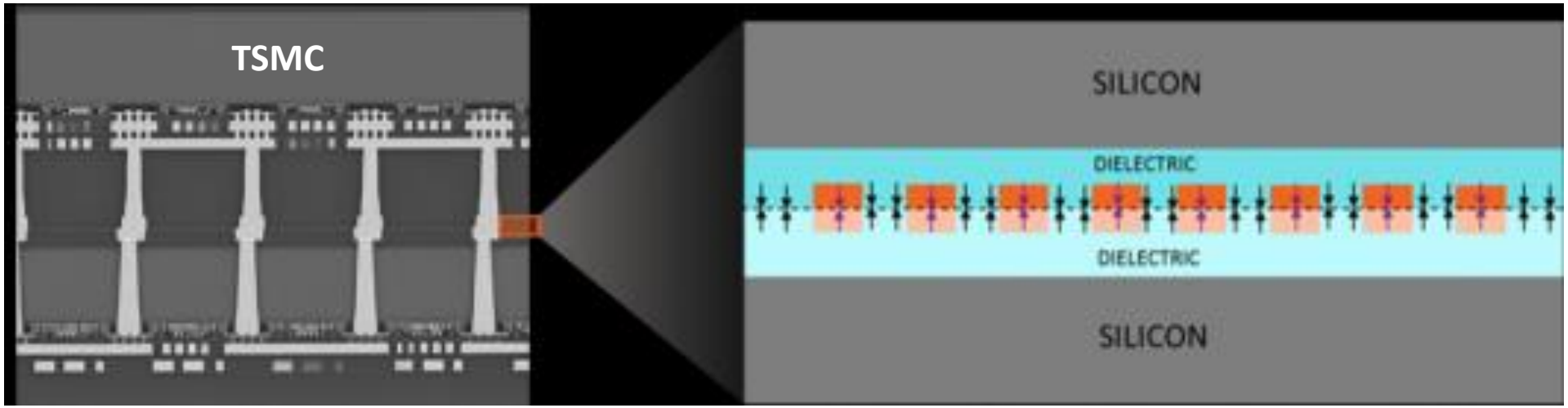
- **TC-NCF 본딩**: NCP를 **NCF (Film) 필름 형태로 바뀌어 사용(Samsung 공법)**

ULVAC

HBM 제작 Fab 기술

Next Level Tech.

하이브리드 본딩(Hybrid bonding): 범프 없이 칩과 칩을 접착하고, 데이터 통로를 곧바로 연결하는 고도화된 본딩 기술



Die to Wafer Hybird Bonding, D2W 본딩
→ Now, Scaling,
D2W 하이브리드 본딩이 10μm 상호연결 피치(Interconnect Pitch) 이하에서 2μm 피치까지

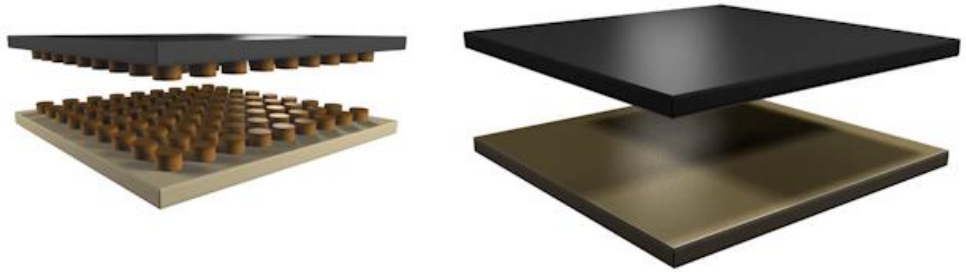
ULVAC

HBM 제작 Fab 기술

HCB : Hybrid Cu Bonding

Challenges

- Cu pad surface control
- Particle, cleanliness involved void
- Pad alignment accuracy
- Bonding temperature
- Metrology/inspection
- Integrated bonding/assembly system

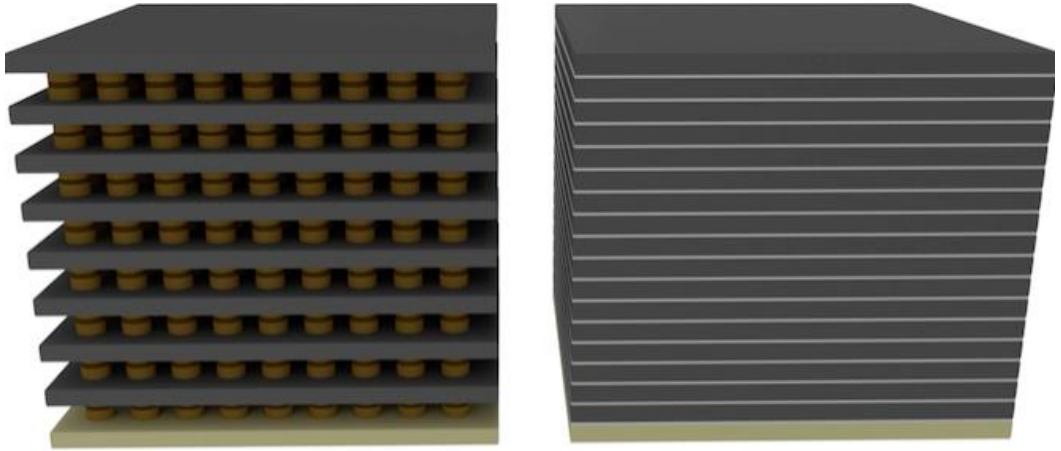


COPPER PILLAR
INTERCONNECT TECHNOLOGY

Only 625 interconnects / mm2

DBI® Ultra

100,000 TO 1,000,000
interconnects / mm2



COPPER PILLAR
INTERCONNECT TECHNOLOGY

8 HIGH

DBI® Ultra

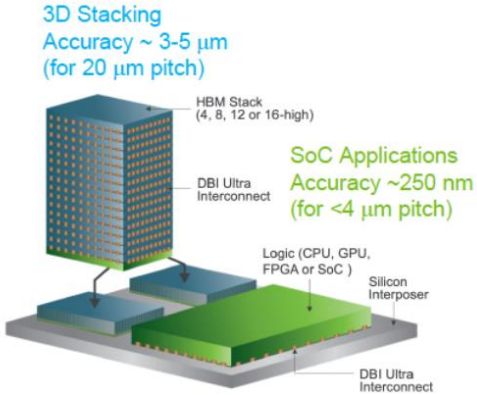
16 HIGH

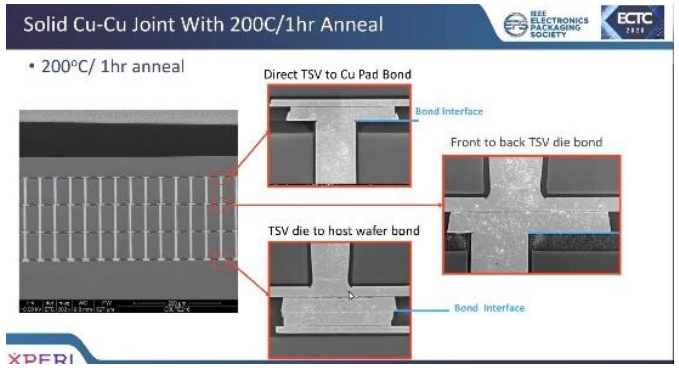
ULVAC

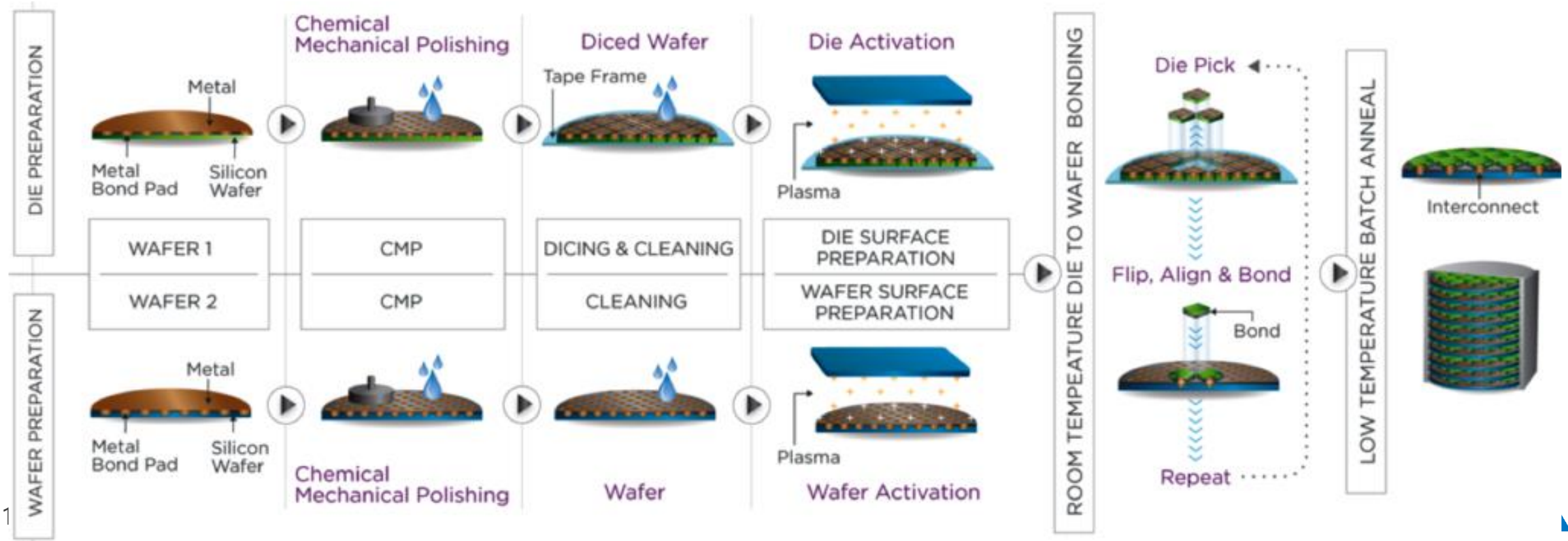
HBM 제작 Fab 기술

HCB@HBM : Hybrid Cu Bonding

1단계 : 실온에서 SiO2-SiO2 결합형성 (Cu 산화 최소화)
2단계 : 150~300C batch annealing → Cu-Cu 결합.







HBM 제작 Fab 기술

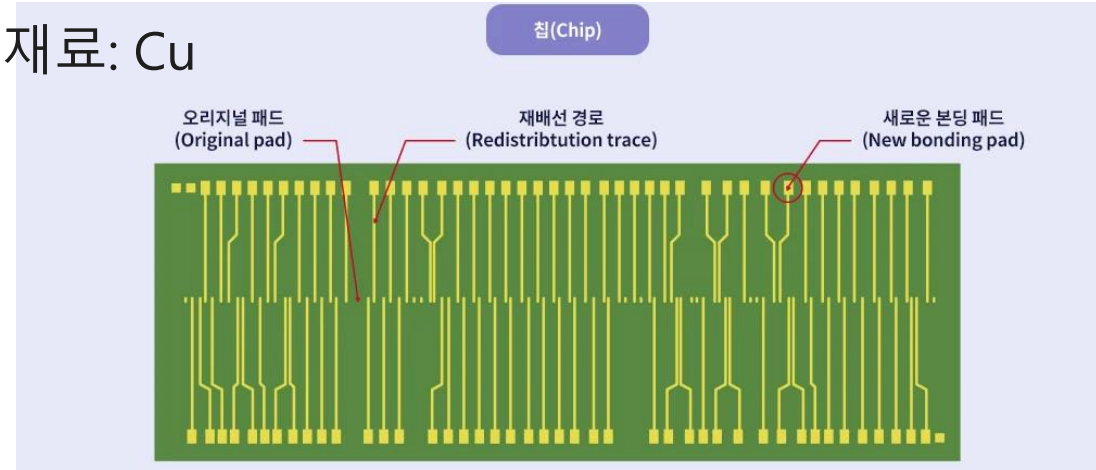
HBM RDL(Redistribution Layer, 재배선)

RDL은 실리콘 칩 위에 있는 고밀도 연결부를 PCB(Printed Circuit Board)의 저밀도 연결부와 서로 접속시키는 역할.

본딩 패드를 엣지(Edge)로 재배열하는 공정
→ 구조변경없이 Chip 적층 가능

PCB로 신호를 접속하게 하려면 여러 RDL 층이 필요

-재료: Cu



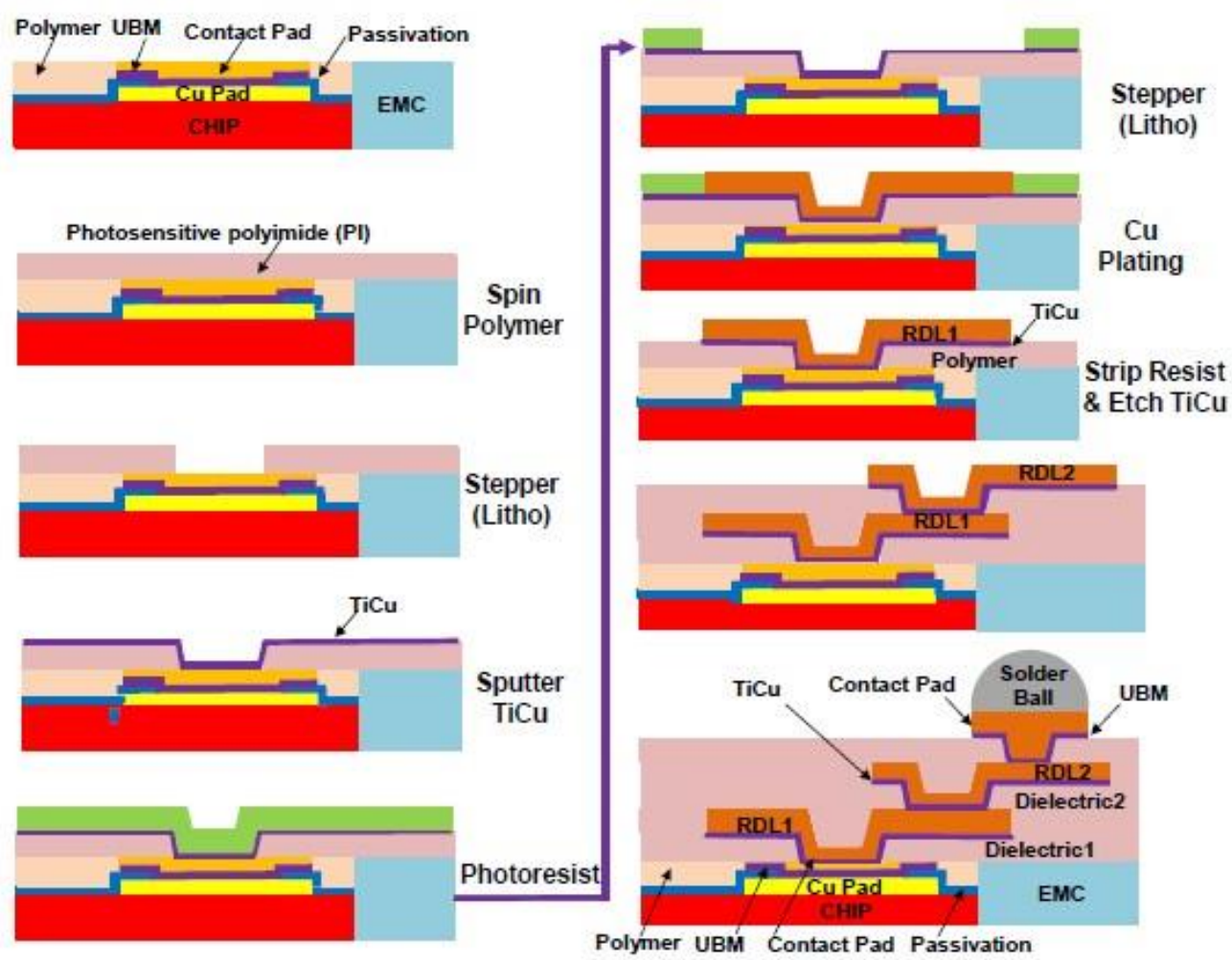


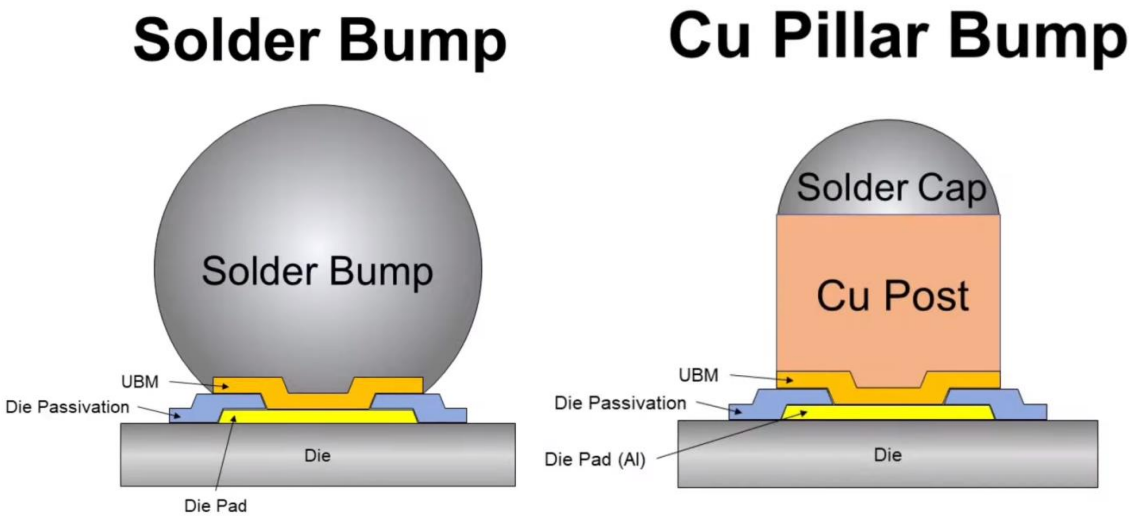
Fig. 14 Key process steps in fabricating the RDLs of FOWLP with conventional UBM/Cu-pad for solder ball

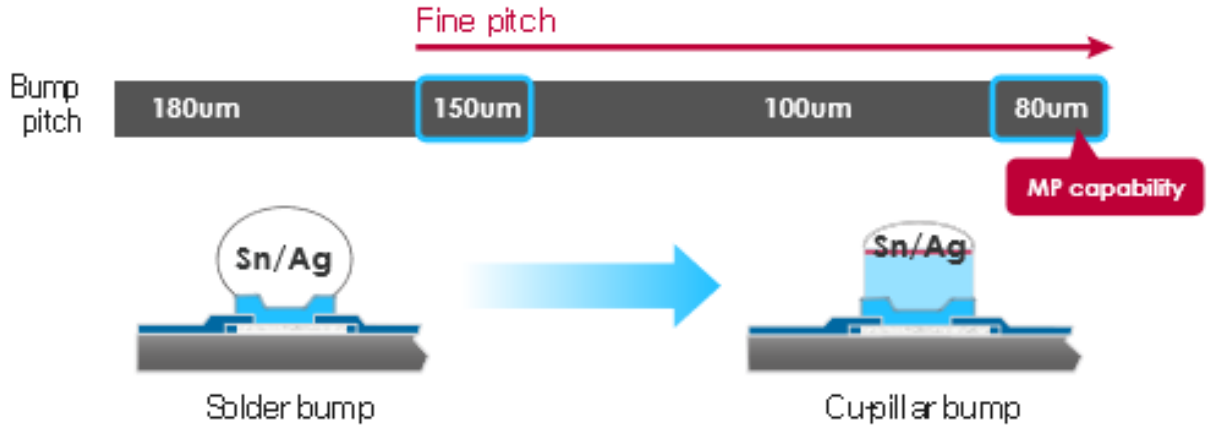
HBM 제작 Fab 기술

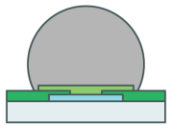
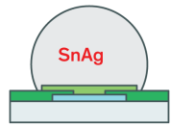
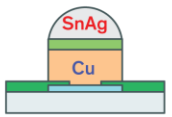
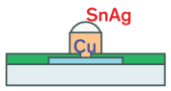
HBM Cu-UBM & Bump

Cu-솔더범프 구조는 Cu Post = Cu Pillar가 Reflow 공정 시 녹지 않아 미세피치 범핑에서 적용된다.

Solder Bump & Cu-pillar Bump



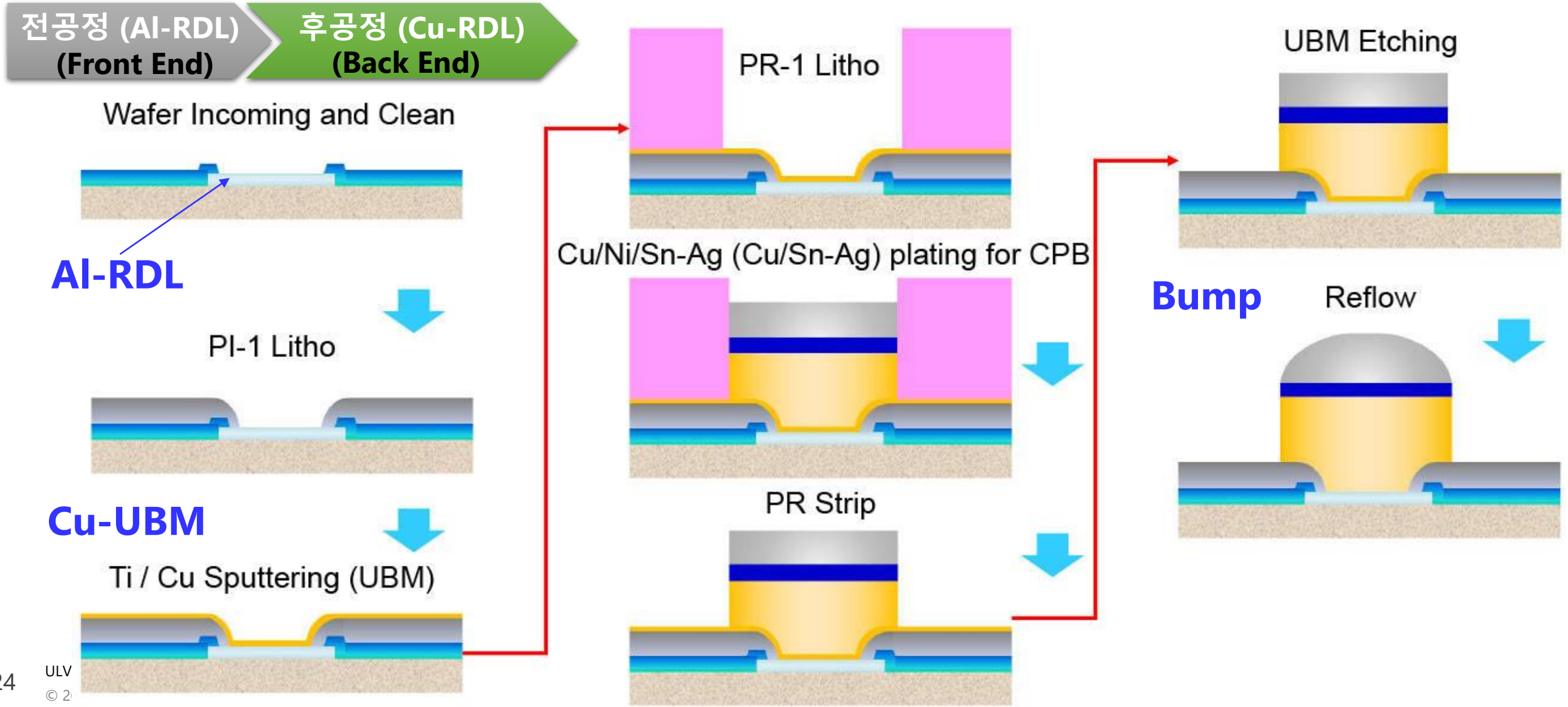


	SnPb C4 Bump	Pb-Free C4 Bump	Cu Pillar + Pb-free Cap	Cu μ-Pillar + Pb-free Cap
Structure				
Diameter	75 – 200 μm	75 – 150 μm	50 – 100 μm	10 – 30 μm

Old Technology → Current Technology → Next-Generation Technology

HBM 제작 Fab 기술

HBM RDL (재배선) : Al-RDL, Cu-UBM & Bump

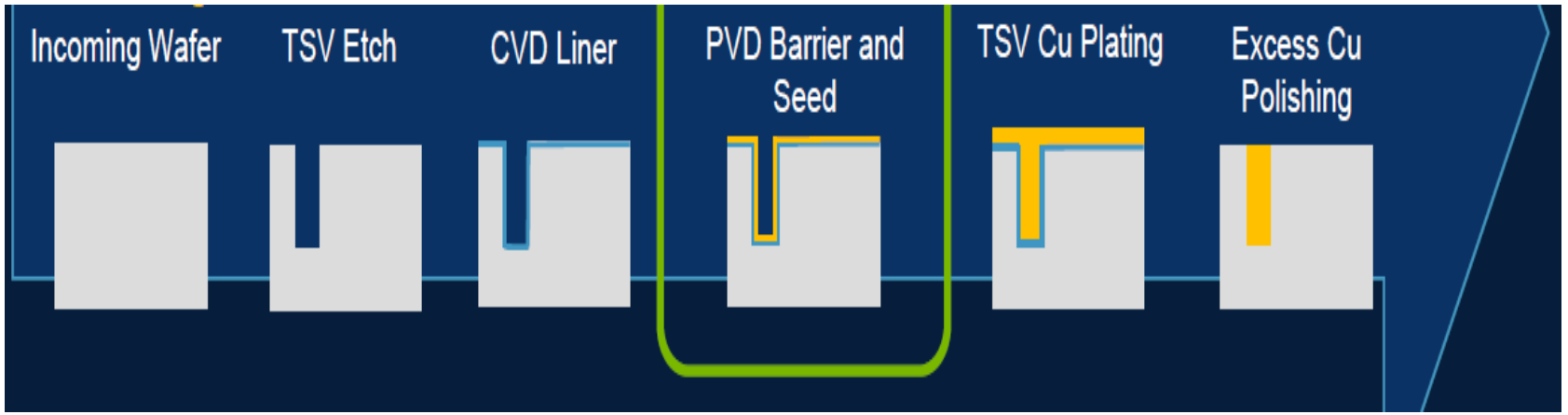


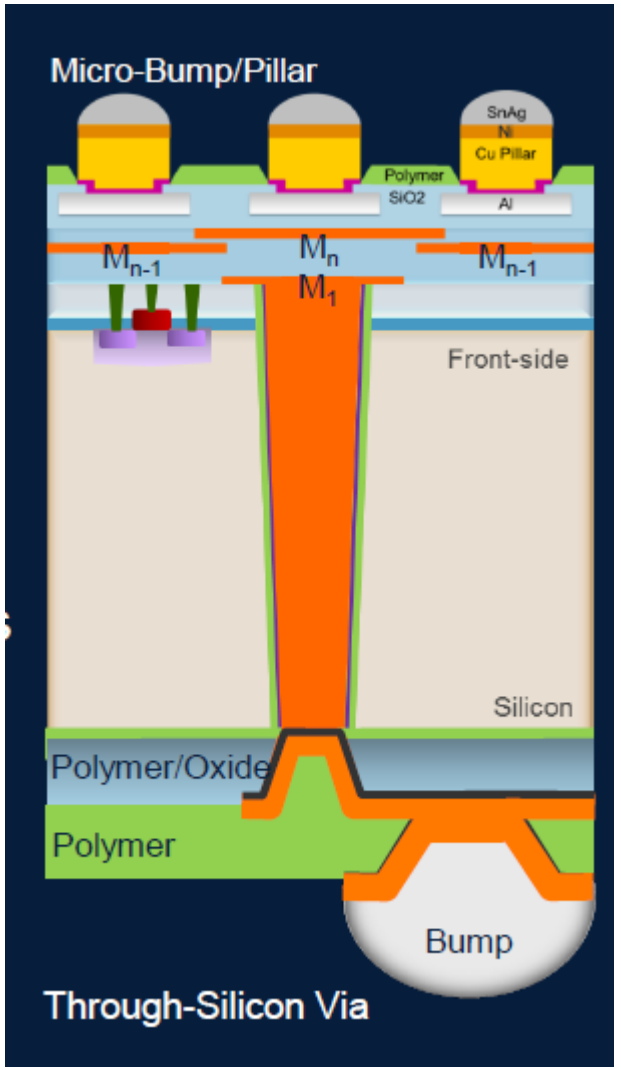
HBM 제작 Fab 기술

TSV(Through-Si Via, 실리콘 관통전극) Process

- 와이어로 칩 연결하는 와이어 본딩(wire bonding) 대체
- 칩에 미세한 구멍(via)을 뚫어 상단 칩과 하단 칩을 전극으로 연결하는 패키징(packaging) 기술 (backside CMP 기술)
- 빠른 신호전달, 고용량, 저전력에 유리

- 2010년 삼성전사에서의 TSV 기반 D램 모듈 개발 발표
- 2013년 12월 26일 SK하이닉스에서 TSV 기술을 적용한 초고대역폭 메모리(HBM, high bandwidth memory) 개발



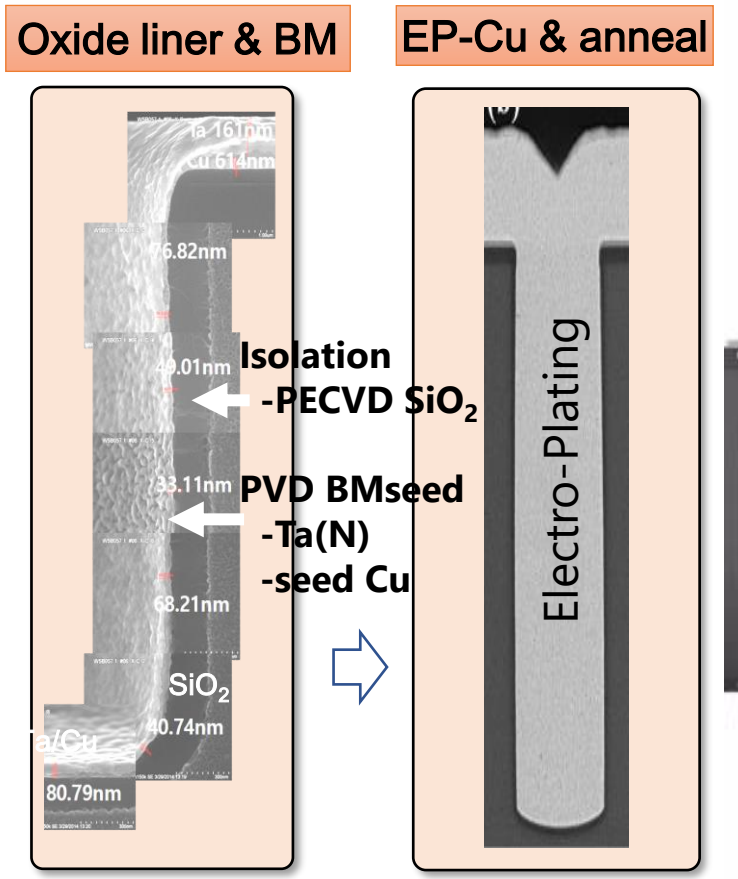


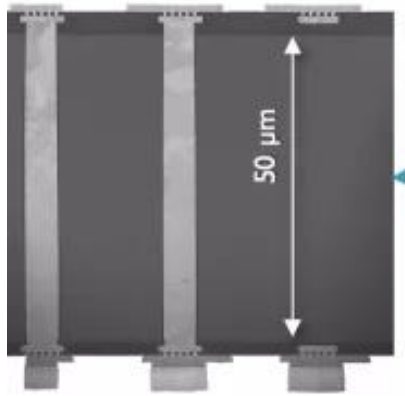
Source: applied materials

HBM 제작 Fab 기술

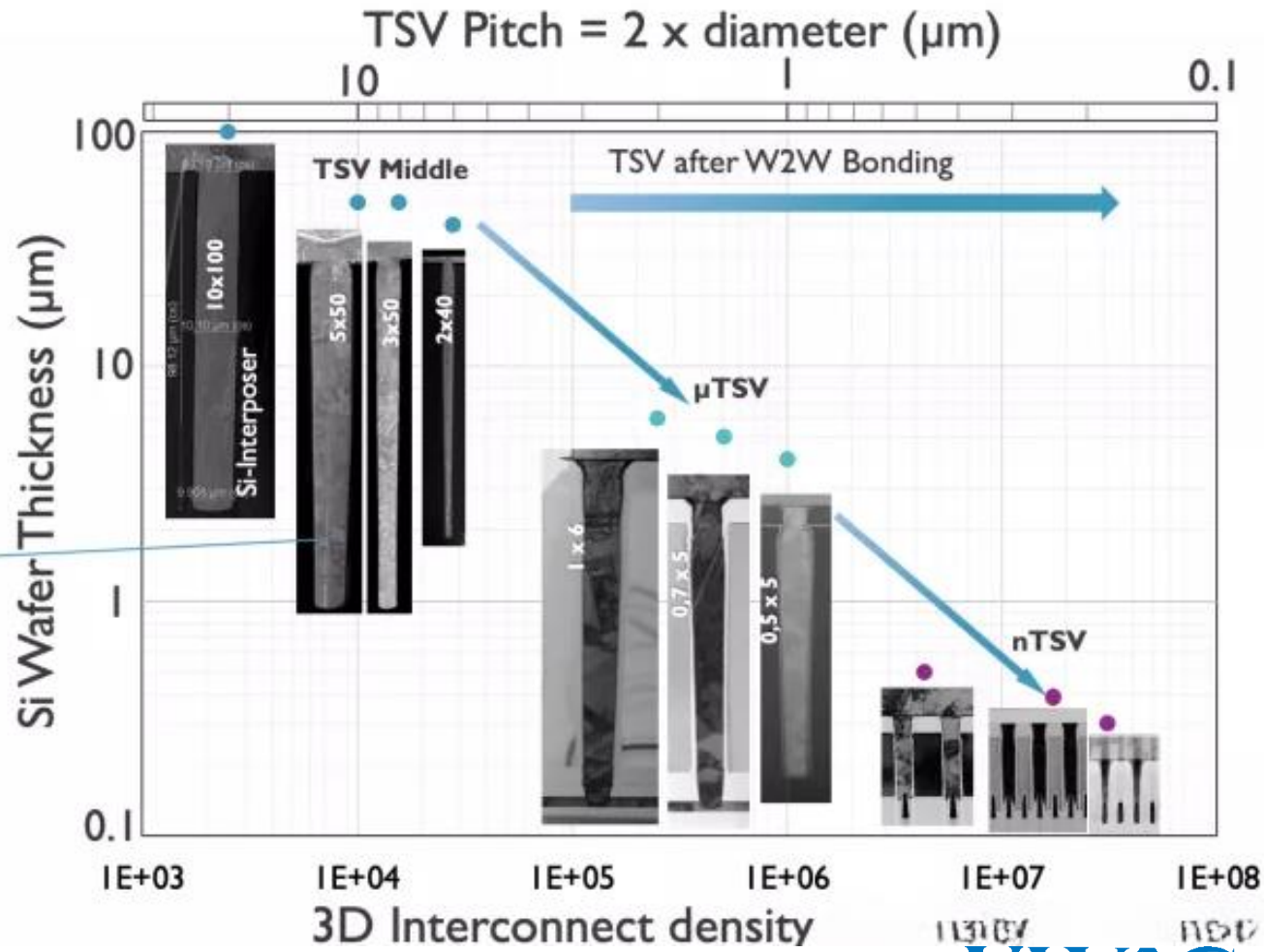
TSV(Through-Si Via, 실리콘 관통전극) Process

- Chip 또는 Si interposer에 Via Hole 형성, Metal 증착 및 Cu 전기도금으로 Via-Fill 기술



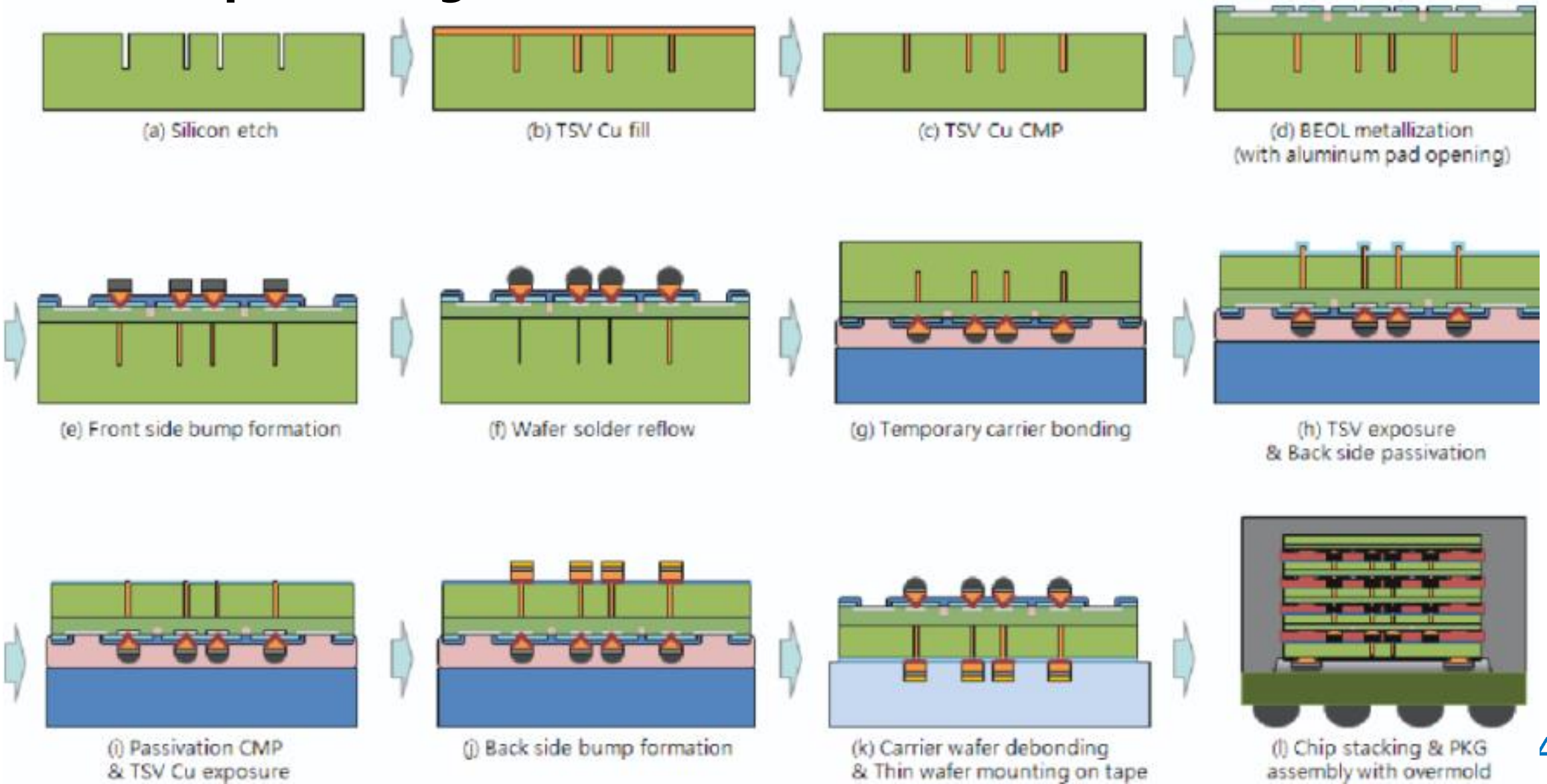


SCALING ROADMAP



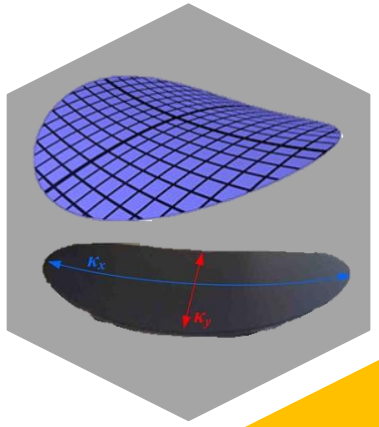
HBM 제작 Fab 기술

HBM Chip Stacking Procedure



HBM 제작 Fab 기술

진공기술 난제 Item



Warpage

Wafer에 chip 실장시,
Chip 두께 얇게(TSV back grinding)..
박막의 잔류 응력,
균일하지 않은 TSV의 배열
Chip 위/아래 배열 비대칭성 → C4 bump 변형유발

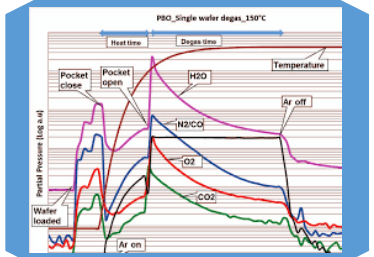
3차원 적층 구조

높은 열전도도 방열특성

저온공정기술(<200℃)

PID, Photo imageable dielectric

Outgas



초고진공

Vacuum chuck(hole)

ESC

Wafer handling vacuum chuck

Vacuum Pump

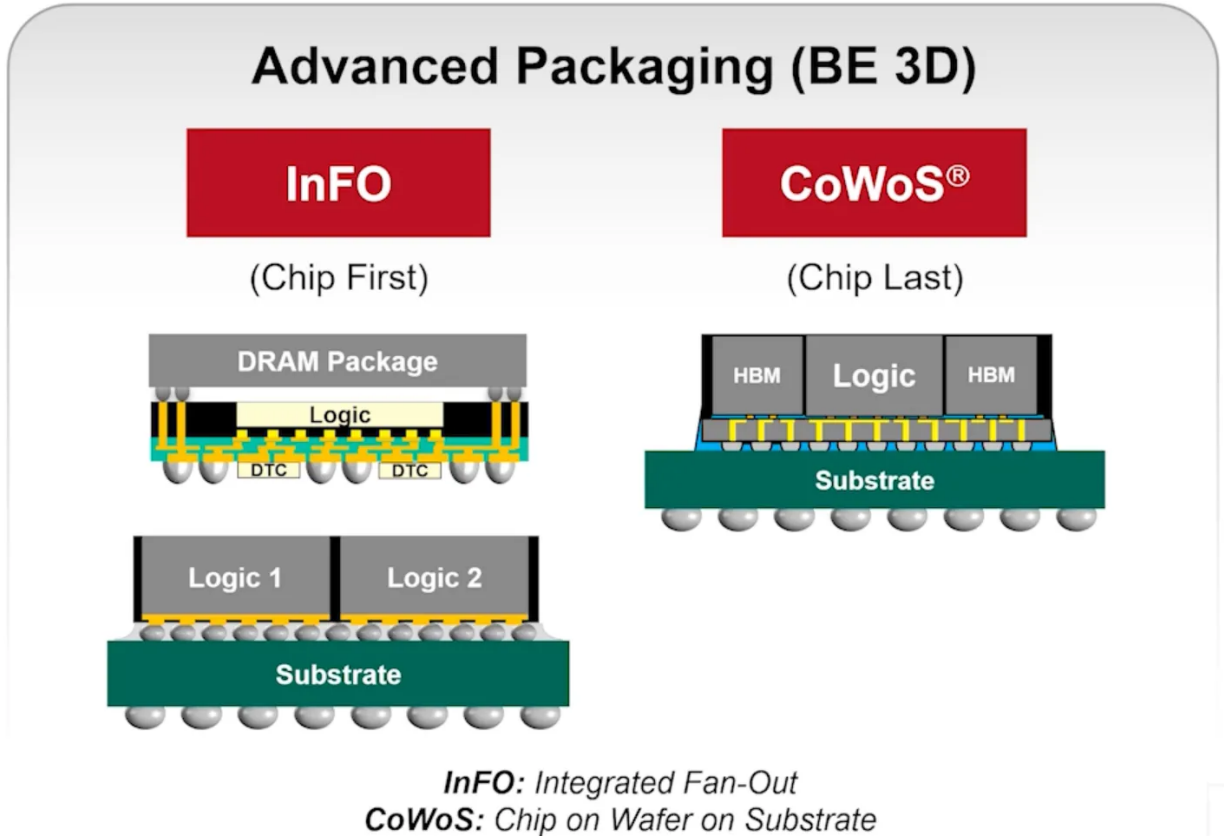
Vacuum Recovery

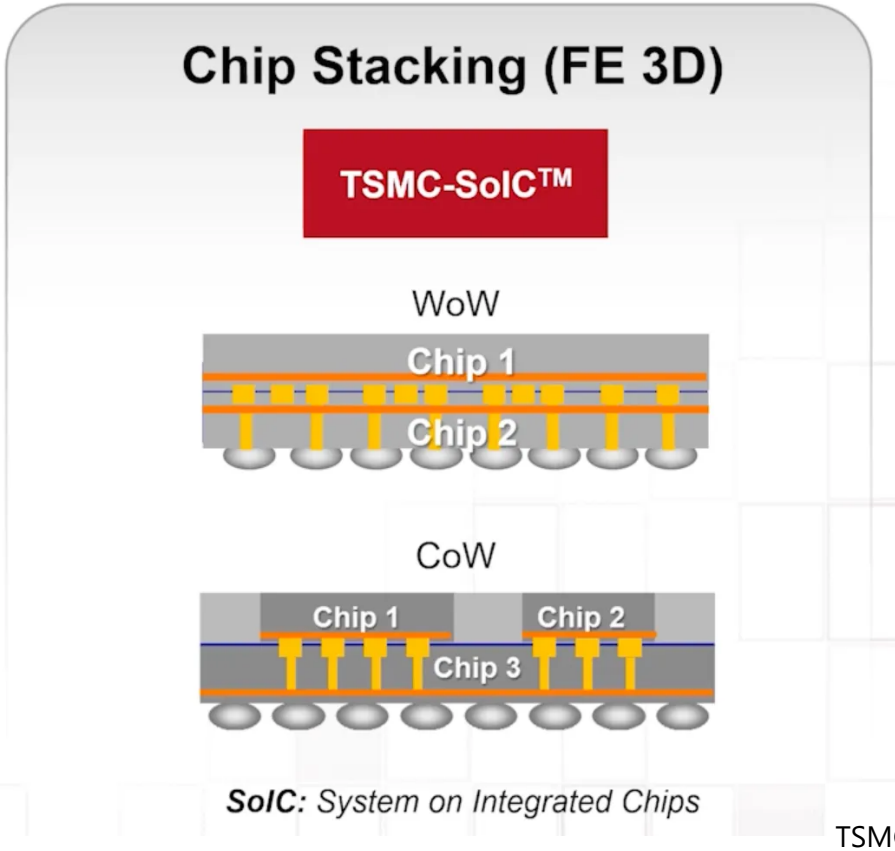
ULVAC

Advanced Packaging

Advanced Packaging 이란?

- 패키징에서의 연결 즉, '패키징 인터커넥션(Interconnection)' 기술, 칩을 패키징하는 특정 방법
- 사용기술: 실리콘 관통 전극(TSV), 브리지, 인터포저
- **이종 집적(Heterogeneous Integration)** 즉, 시스템 반도체와 메모리 반도체를 불문한 반도체 통합





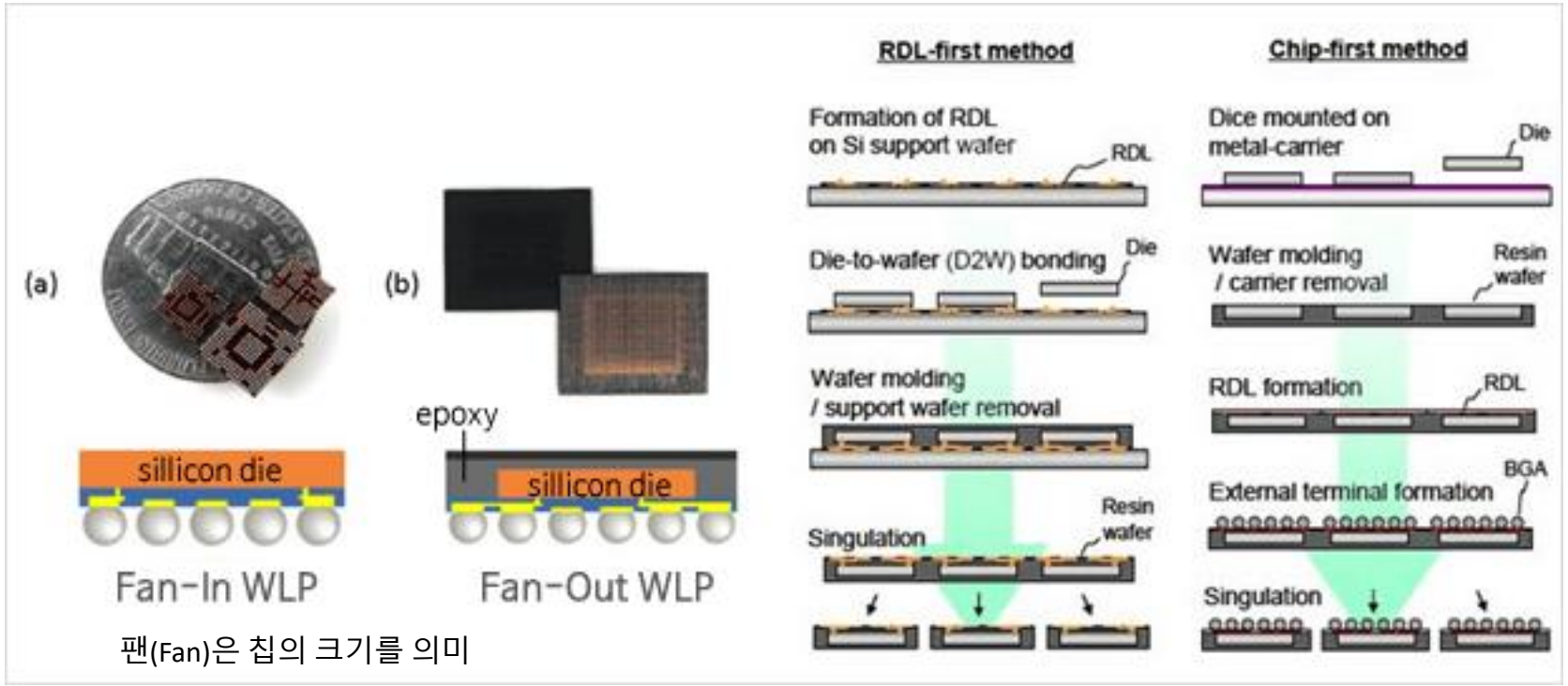
TSMC 3D Fabric

TSMC 자료

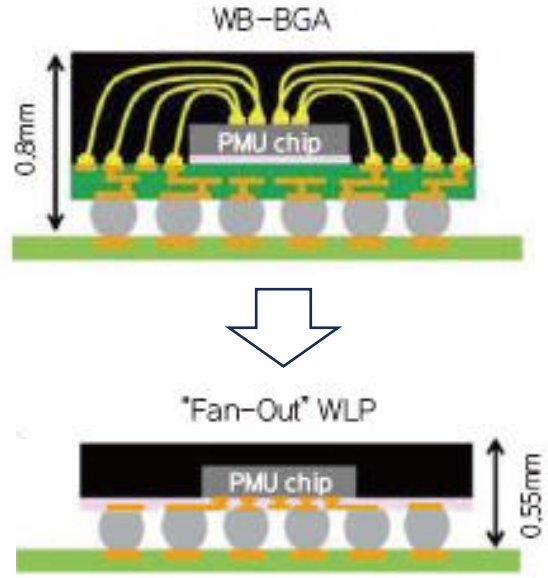
Advanced Packaging

● Fan Out Wafer Level Package(FOWLP)

WLP는 wafer에 직접 Chip을 실장하는 기술 → package 두께/부피 감소
기판과 같은 매개체 없이 솔더볼 (입출력 단자)을 칩 위에 바로 붙여 패키징하는 방식
Packaging 원가절감,
이종 칩과 수평 연결이 가능 → 고성능제품 창출 (ex. HBM + GPU / CPU 연결)
반도체 고성능화로 입출력(I/O) 단자수 증가 → Fan-Out(FO) + 재배선(RDL)



팬(Fan)은 칩의 크기를 의미



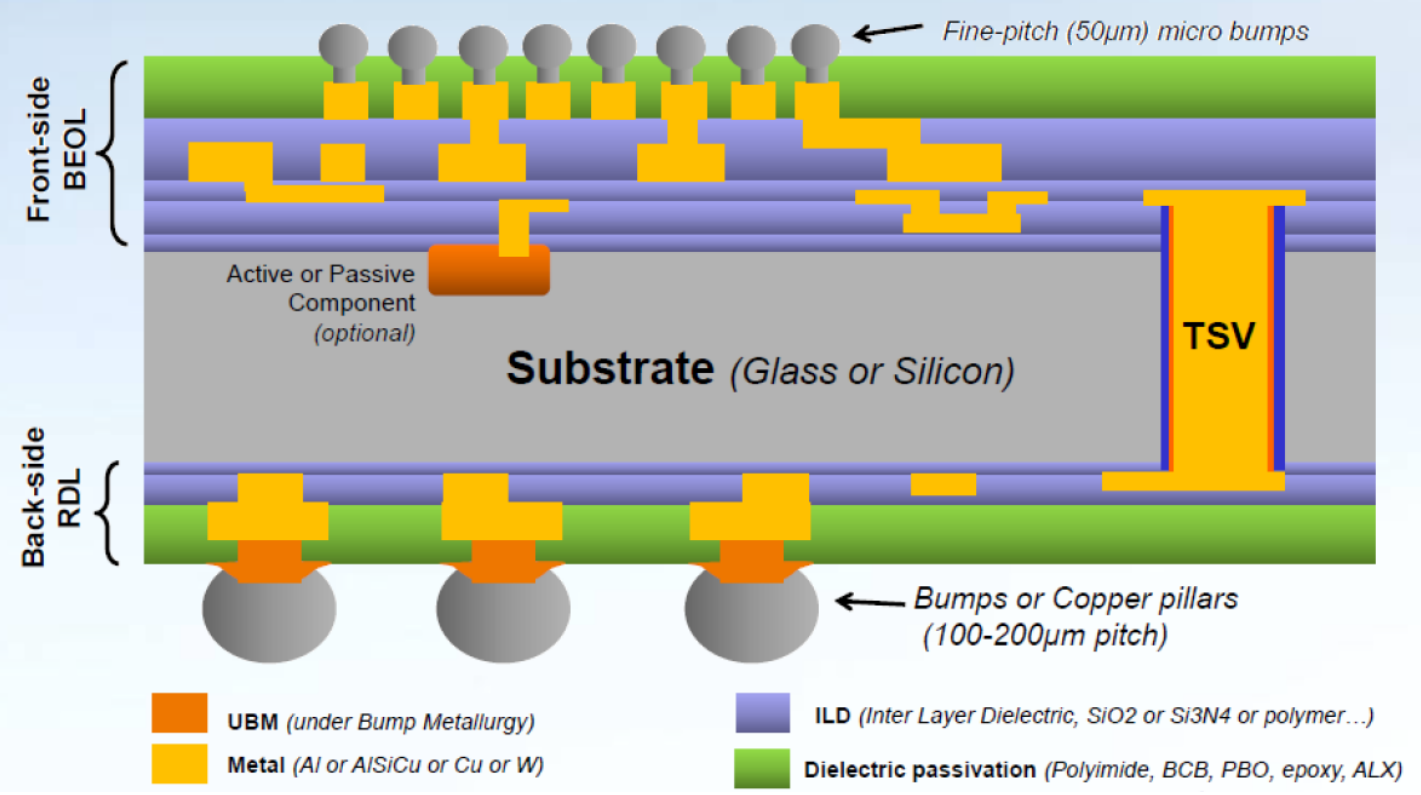
Fan-out WLP 장점

(좌) Fan-in vs Fan-out WLP 비교 / (우) FOWLP 제조과정

Advanced Packaging

- **Si Interposer의 용도**

  - Buffer(매개체) : IC와 PCB를 연결, 이종 chip 사이에 삽입되어 재배선 및 buffer 역할
    (IC의 단자 밀도와 PCB의 단자 밀도가 20배 정도)
  - Fine Pitch Si Substrate : 다수의 chip을 integration 하는데 사용되는 silicon substrate
  - 실리콘 인터포저를 통해 집적도를 높인 후, 를 Cu-C4로 범핑한 후 기판과 신호를 주고받음



Advanced Packaging

Samsung's advanced packaging Options - 2.5D, 3D

Next Roadmap

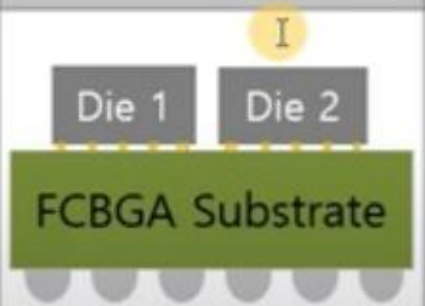
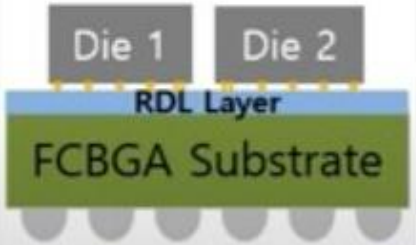
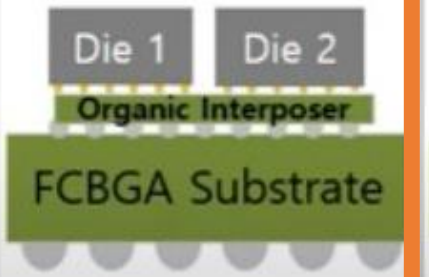
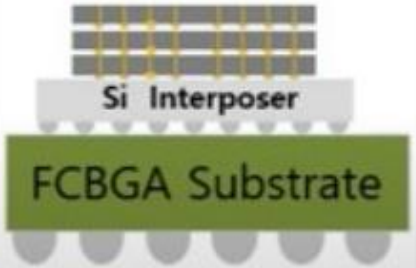
Package Type	2D	2.1D	2.3D	2.5D	3D
Schematic Image					
L/S (㎛)	9/12	~2/2	~2/2	<1/1	<1/1
I/O Density	Low	Middle	Middle	High	High

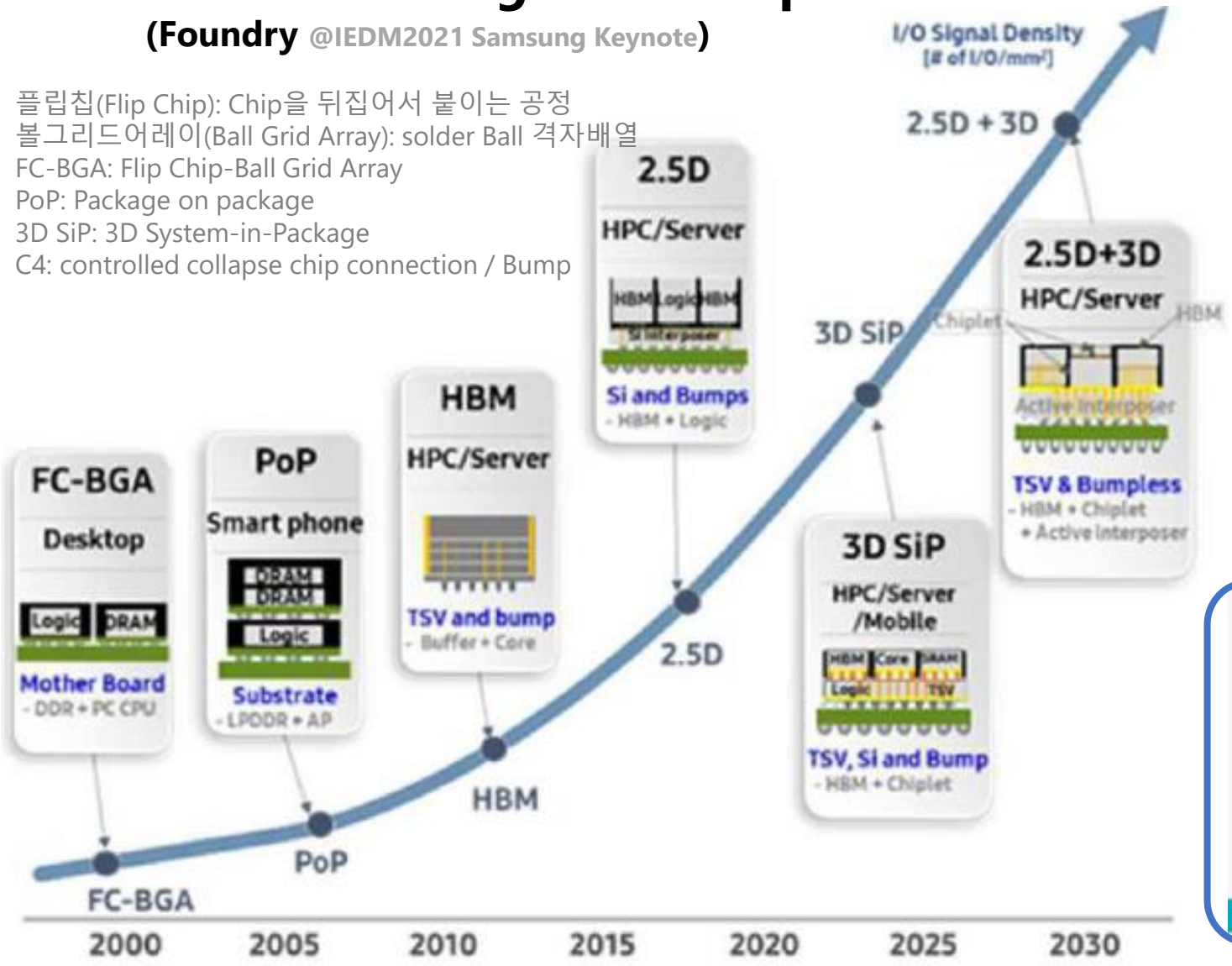
Figure 1: Samsung's packaging portfolio. (Source: Samsung TSP)

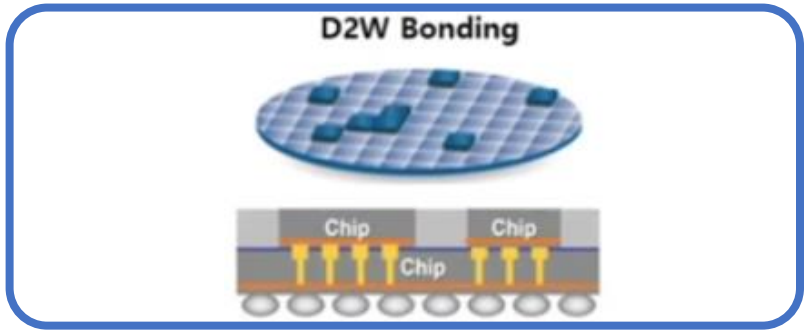
Future @Advanced Packaging

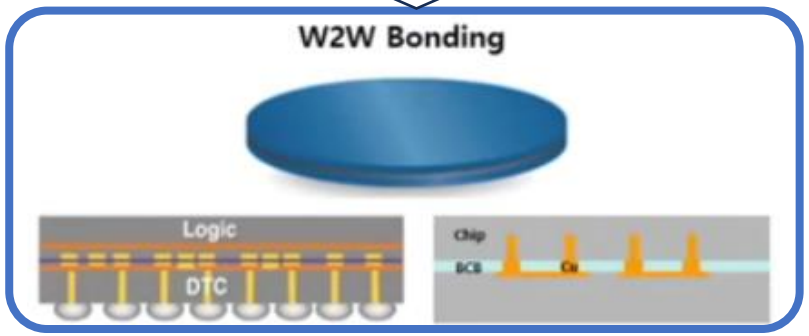
Advanced Package Roadmap

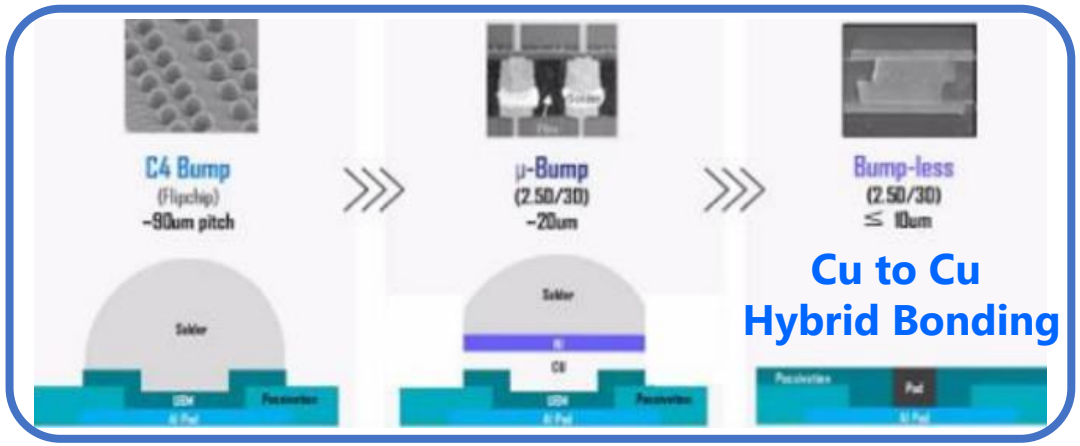
(Foundry @IEDM2021 Samsung Keynote**)**

플립칩(Flip Chip): Chip을 뒤집어서 붙이는 공정
볼그리드어레이(Ball Grid Array): solder Ball 격자배열
FC-BGA: Flip Chip-Ball Grid Array
PoP: Package on package
3D SiP: 3D System-in-Package
C4: controlled collapse chip connection / Bump









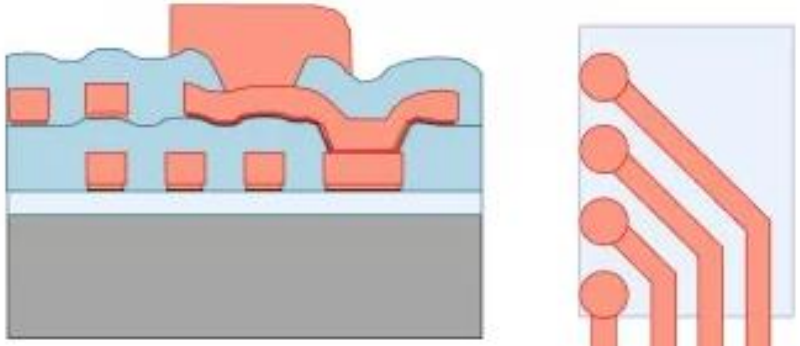
Source: SAFE2020, samsung

Future @Advanced Packaging

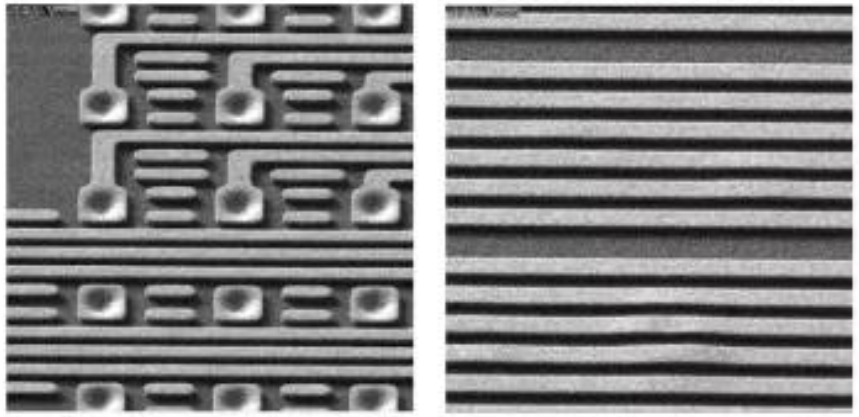
Next Fine Pitch RDL(Redistribution Layer, 재배선)

"Semi-Additive" RDL

2 ⇒ 1.6 µm Line/Space

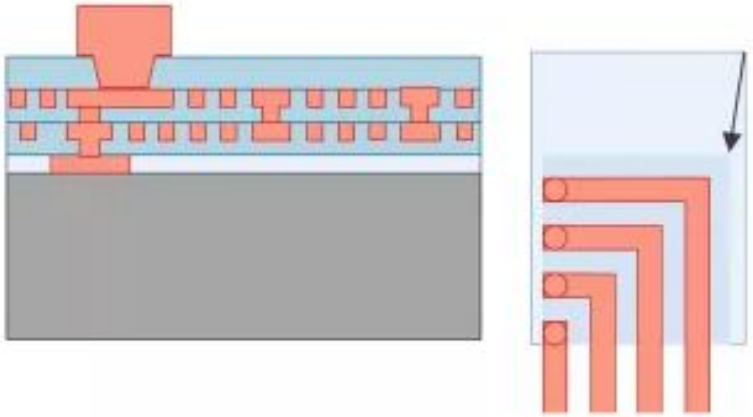


2 µm Line/Space



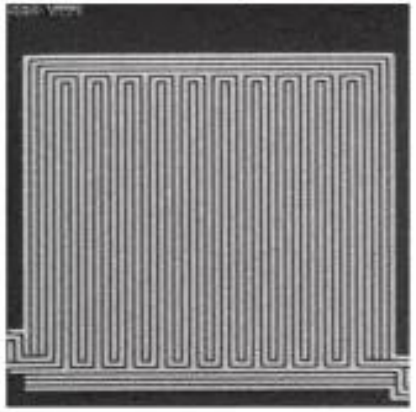
Polymer "Damascene" RDL

2 ⇒ 1 µm ⇒ 0.5µm Line/Space

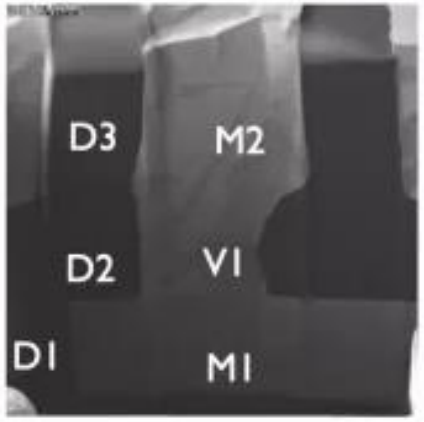


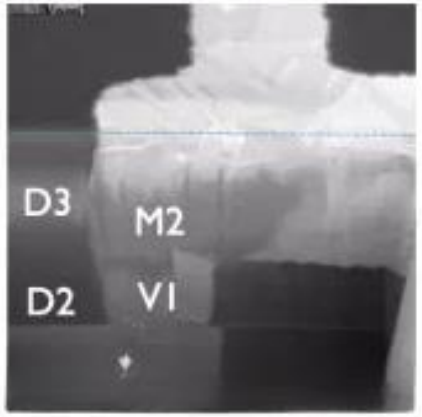
- Flat surface on every layer
- Small pitch on all layers
- Land-less vias < 3µm Ø
- Improved fan-in/out wiring

1µm Line/Space



Dual Damascene M1/V1/M2 Cu





imec

ITF WORLD

Leading the World

In Vacuum Technology

ULVAC