SWTEST

PROBE TODAY, FOR TOMORROW

2024 CONFERENCE

Advanced Probe Card Solutions to Address HBM Wafer and Stack Die Test Challenges

FormFactor™

Speaker: David Cooke
Sr. Product Marketing Manager

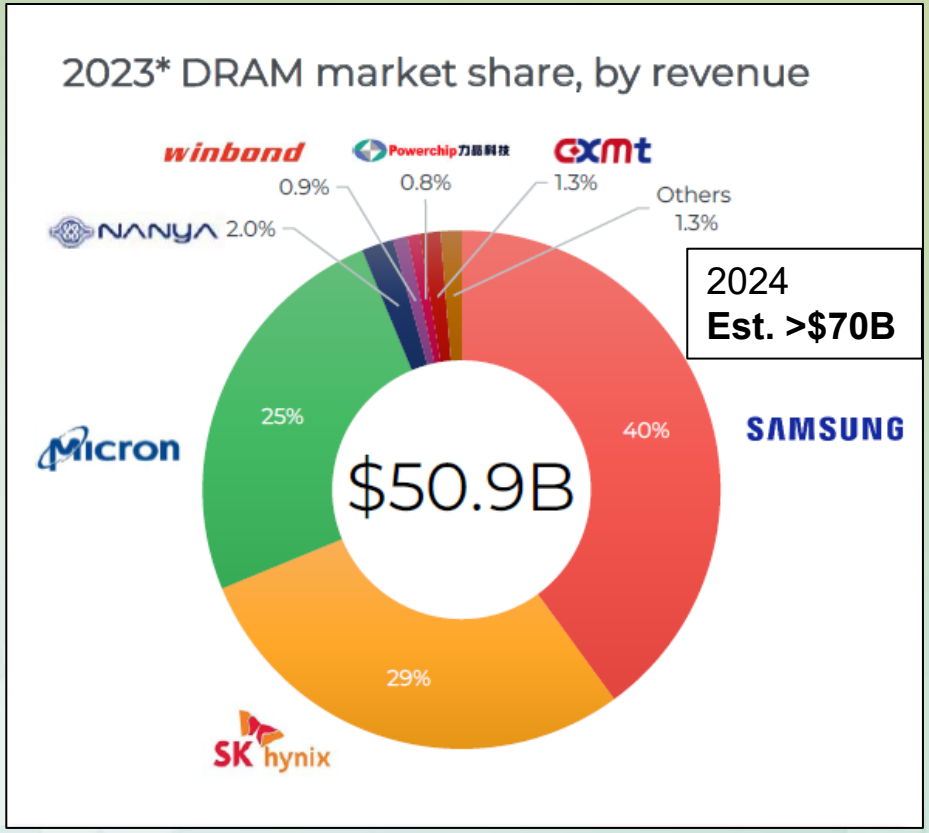
Co-Author: Kalyanjit Ghosh
Sr. Staff Mechanical Design Engineer

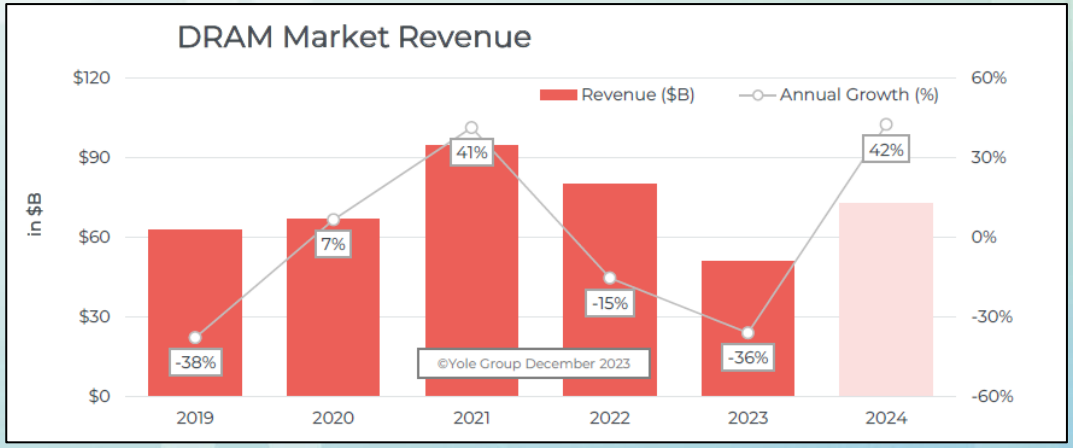
Overview

- DRAM Memory Market Overview
- HBM (High Bandwidth Memory) Market Review
- Advanced Packaging
- HBM – Processing In Memory (PIM)
- Advanced Packaging Increasing with Rise in AI
- Test Challenges & Yield Impact
- Thermal Management / Scaling Problem
- How HBM Impacts Test Speed
- Alternatives to Wafer Test
- Conclusions & Summary

DRAM Market Overview

- For the last 8 quarters the memory market has experienced the worst downturn in 15 years
  - Largely caused by excessive capex and wafer fab equipment spending and weak demand post COVID
- Major production cuts implemented by SK Hynix, Samsung, and Micron are bringing the market back to equilibrium (20-30%)'23
- The top 3 suppliers are 94% of a $50.9B DRAM total market '23
  - SK Hynix, Samsung, and Micron moving capacity to HBM
  - 2024 DRAM Revenue expected to reach >$70B in revenue
  - 2024 HBM Revenue expected to reach $14B (20% of overall DRAM revenue)
- During this memory market "winter" DRAM demand was weak, except for **AI server apps** (and automotive)
  - Generative AI has boosted the demand for high-speed memory
  - Data Center driving growth for storage and analytics
  - Applications requiring high bandwidth include: AI Servers ChatGPT, Supercomputers, 8K video, VR, Cloud, etc.

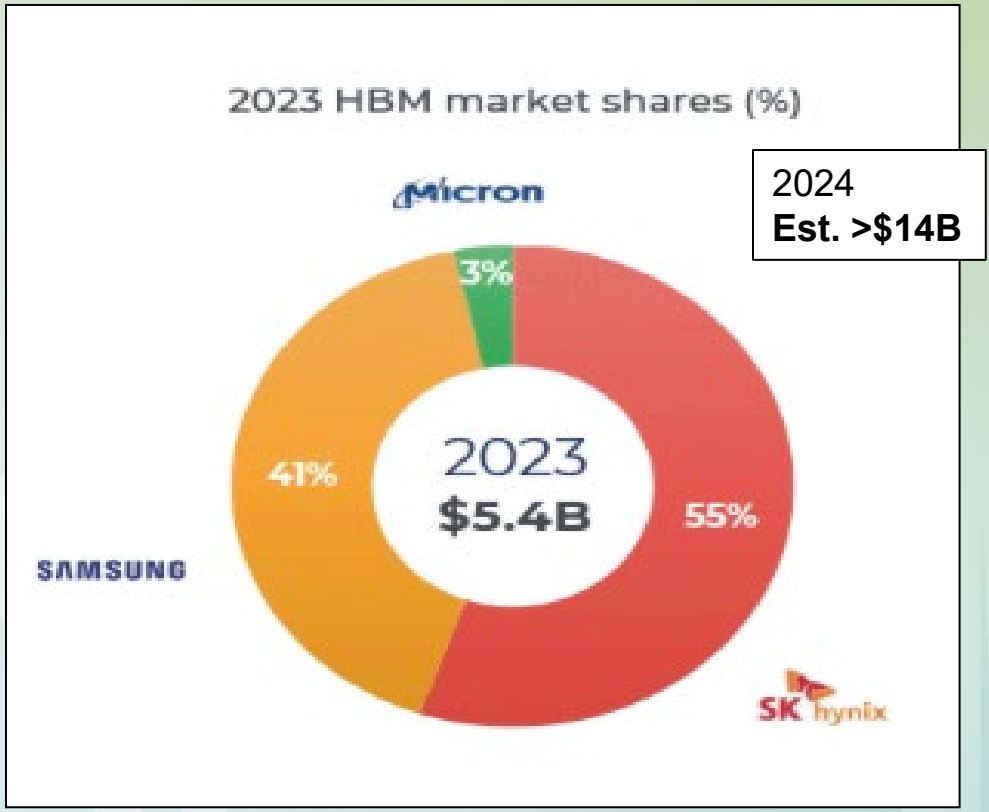




Source: Trendforce, Yole

HBM Market Share 2023-2024 – Market Growing Fast

- **SK Hynix: leading supplier having pioneered the standard in conjunction with AMD**
  - As of 2023, SKH has ~55% of HBM market share
- **Samsung has ~41% market share (but lead DRAM overall)**
  - Their goal is to double their HBM capacity in 2024
- **SKH and SEC, each investing $750M in 2024 for HBM3**
- **Micron smaller market share thus far**
  - Behind in HBM MS – but planning to catch up rapidly
  - Introduced 12High stacked dies (HBM3) in 2023
  - Ramp expected in 2024



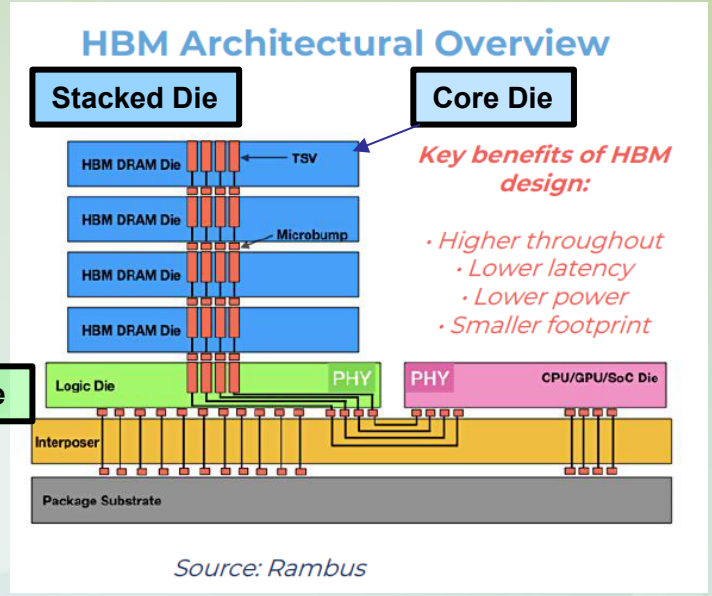
2023: HBM $5.4B (10% of total market)
2024: HBM $14B (20% of total market)

Source: Trendforce, Yole

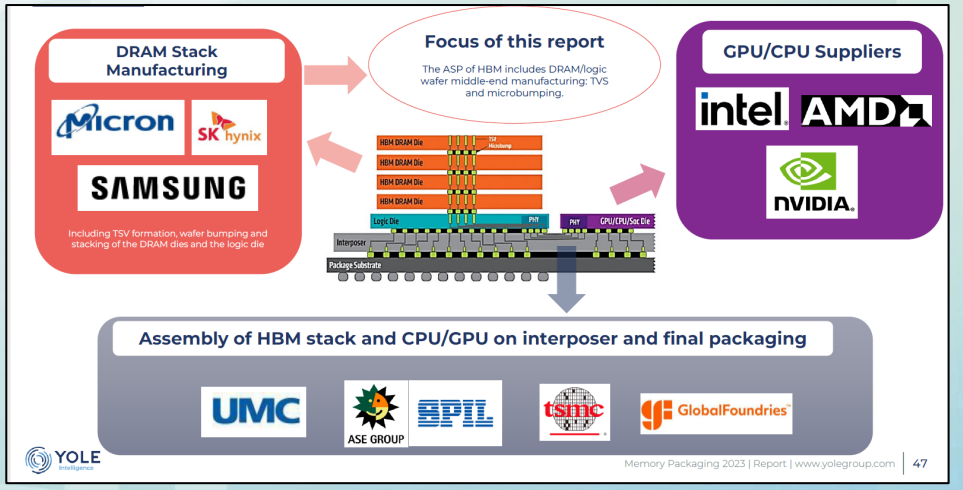
Advanced Packaging

HBM Module Overview

- **High Bandwidth Memory – DRAM die connected vertically - TSVs**
- **HBM3 enables fast data transfer due to its wide I/O interface with 1,024 data bits (vs. DDR5 has 64 data bits)**
  - HBM4 could feature 2,048 data bits – double the interconnect density
  - Referred to as Processing In Memory (PIM)
- **HBM applications are driven by demand for Generative AI (E.g. ChatGPT and other applications)**
  - The big 3 - shifting production to HBM.
  - Next 5 years expect a 38-45% CAGR (Source: Trendforce).
  - 2024 is expected to be the come back year for DRAM.



Source: Rambus



HBM Supply Chain

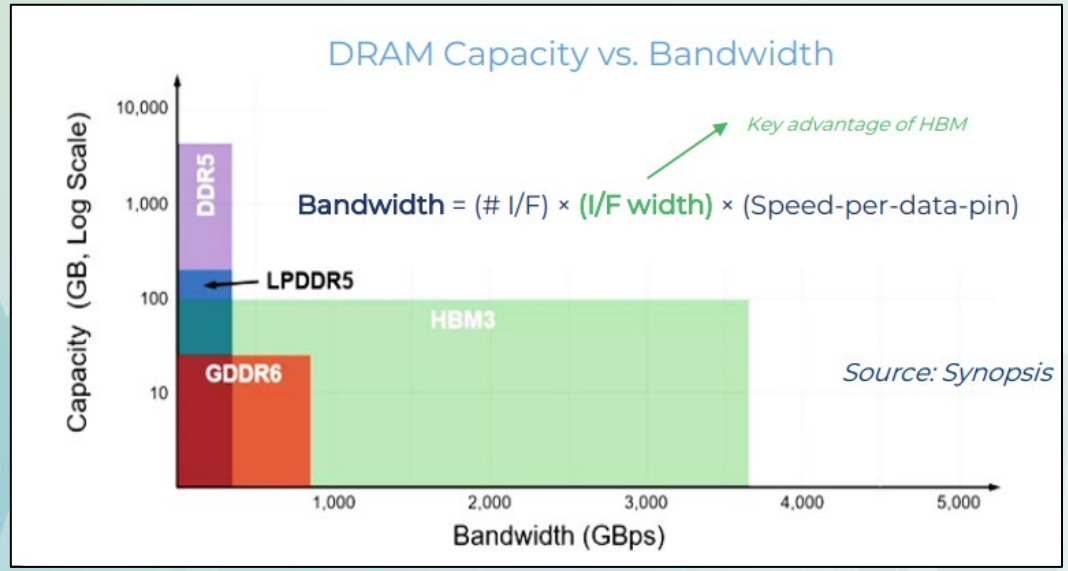
Nvidia H200

High Bandwidth Memory – Processing In Memory

Why is Bandwidth so much higher in HBM?

SoundHound AI

- **HBM is a high value high performance memory product**
- **Providing faster data processing than traditional DRAM – Ideal for Generative AI**
  - That's where advanced packaging comes in – stacking multiple chips with a GPU or CPU



Example of bandwidth & capacity calculations

	# I/F*	I/F Width (DQ bits)	Max. capacity per I/F (GB)	Speed per data pin (Mbps)	Capacity (GB)	Bandwidth (GB/s)	Comments
DDR5	8	64	512	5,600	4,096	350	*Dual 256GB LRDIMMs*
LPDDR5	12	32	16	7,500	192	352	*16Gb dies 8 dies/package*
GDDR6	12	32	2	18,000	24	844	*16Gb dies*
HBM3	6*	1,024	16	4,800	96	3,600	16Gb dies, 8 dies/stack

*I/F = memory Interface channel

Bandwidth = Number of memory interface channels x Interface width in bits x Speed per data pin

Bandwidth Example HBM3 = [6 x 1,024 (Gb) x 4,800 (Mbps)]/ 8 die per stack = ~ 3,600 GB/s

Source: Yole 2024

Bottom Line: Data in and out a lot faster – which reduces processing time of data

Advanced Packaging Demand Drives Testing Demand

- **Advanced Packaging Demand Taking-off**
  - Beyond 2025 50% of IC's are forecasted to be Advanced Packaging
- **Advanced Packaging Complexity Trend**
  - HBM DRAM stack die increasing
  - Package size is also growing
- **DRAM KGDS Test Help Reduce Risk and Cost for Advanced Packaging HBM Modules**
  - Higher complexity -> Lower Yield
  - Higher Complexity -> Higher Packaging Cost
  - Earlier defect detection helps save package cost

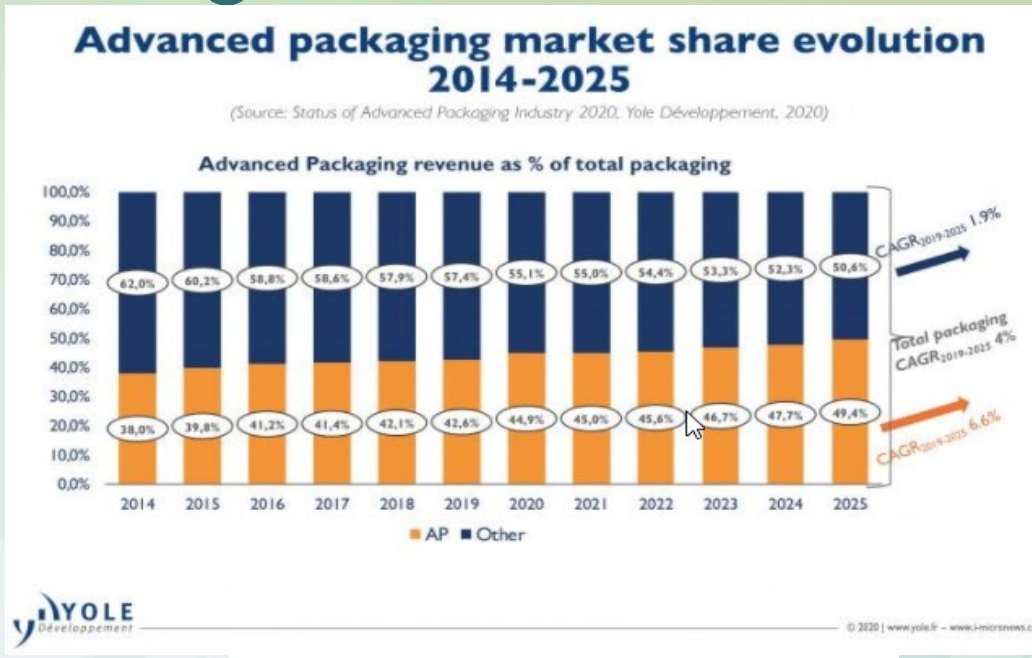
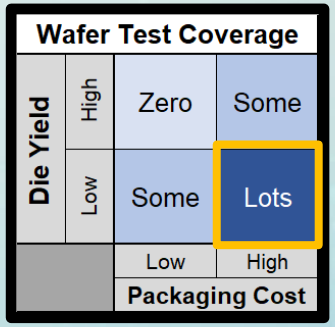


Figure 2. Advanced Packaging market share evolution 2014-2025.



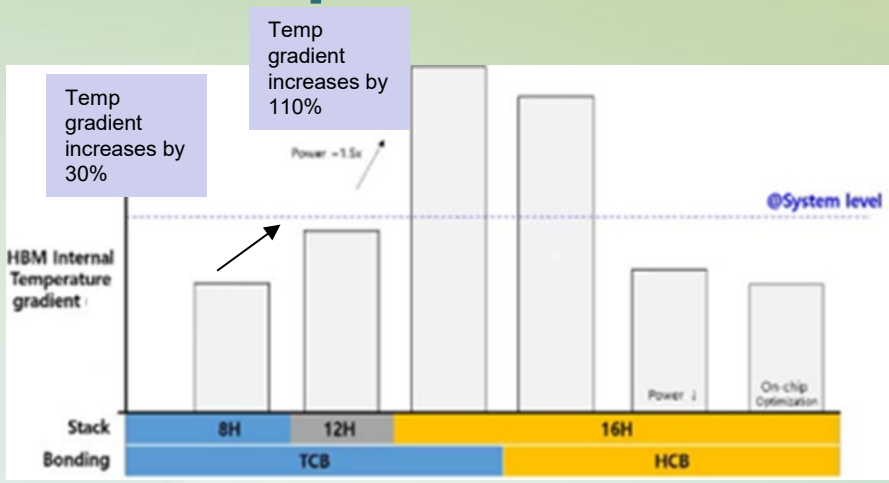
https://www.swtest.org/swtw_library/2020proc/pdf/00pm_SWTest_Untethered_Keynote_Slessor_FormFactor.pdf

Low yield & high package cost – probe is a good idea
High yield & low package cost – probe is a bad idea

Test Challenges and Advanced Packaging Yield Impact

Test Challenges

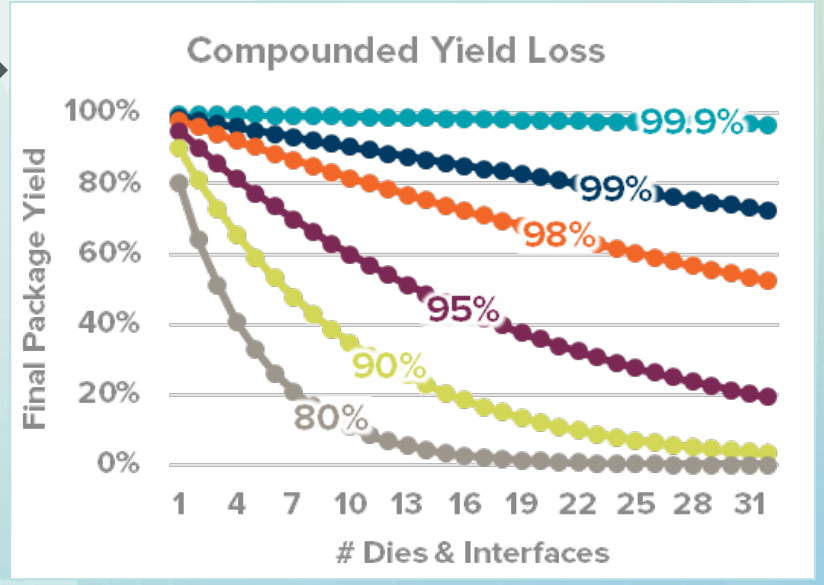
- **HBM wafers CTE varies in X and Y Axes (HBM Stack Die Wafer have anisotropic CTE (different in X and Y)**
  - Dependent upon number of stacked DRAM layers
  - Will cause misalignment between center of pads to center of probe tips
  - Requires custom scaling target and WSS material for ***multiple*** temperatures
- **Wafer and probe card heat-up during test due to power / current through devices**
  - Joule heating of the stacked dies above and beyond the chuck temperature can result in a scaling mismatch
- **High Speed**
  - Being able to test at high speed (>4GHz) for a stacked die compared to a conventional wafer



Source: Article from Samsung Notes 2023 test paper on HBM technology

Yield Impact

- **Final test of assembled package is necessary to improve performance and yield**
  - Wafer test provides valuable yield learning on component die and ensures the final stacked assembly does not get scrapped because of one bad die
- **Economics may dictate something finding other ways to test to ensure KGD**
  - Pre-package wafer test is fundamentally scrap-cost avoidance
  - Final-test and system-test opportunities to prevent escapes
- **Cost vs. coverage optimization comes down to math**
  - Must have KGD – must have highest quality – end customers will not accept anything less than 100%
  - Hedge bets – e.g., design interposers/ bridges with redundant vias, and build repairability into each HBM sub-die
  - Balance test coverage to catch higher-probability/impact issues, while accepting risk of lesser issues slipping through to final test at probe – final test must ensure 100% KGD



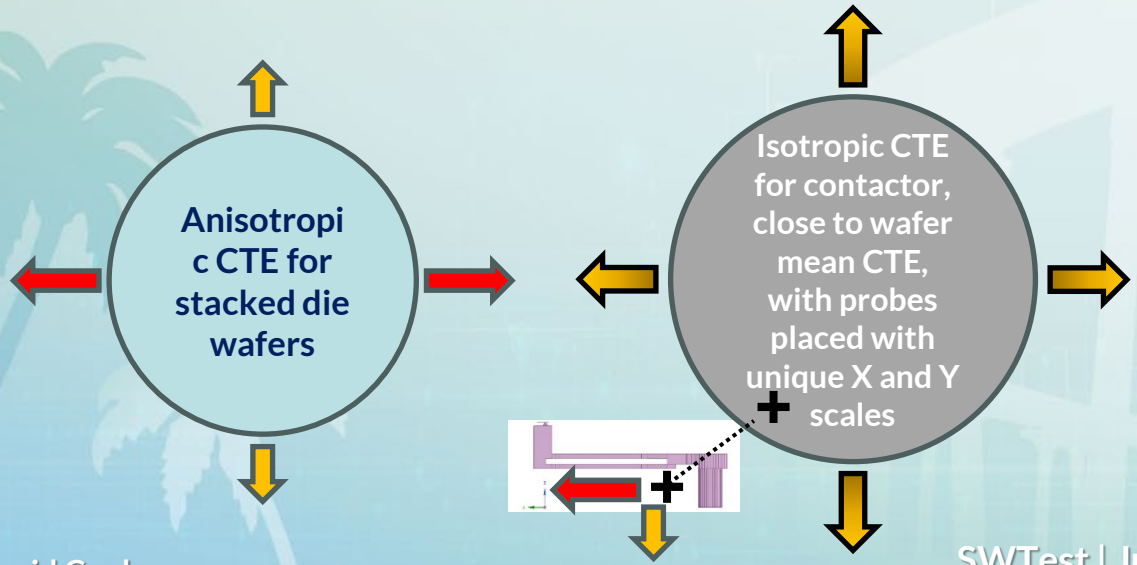
Source: Semicon Korea, Quay Nhin, Achieve the balance of test cost. Feb 2020

Thermal Management / Scaling Problem

Pain points: 1) composite/anisotropic wafer CTE, and 2) power dissipation during test

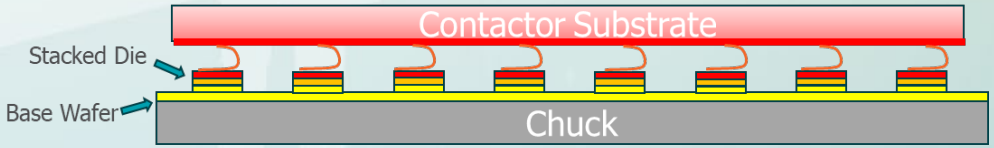
1) Unlike a standard Si wafer, HBM stacked die wafers thermal movement differs in X and Y - and the probe card must follow

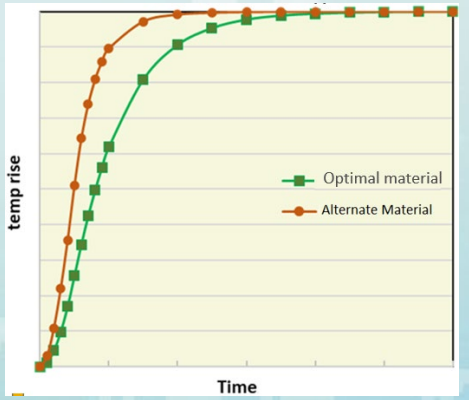
- **To follow the wafer's thermal movement, you need two knobs to turn on the probe card:**
  - **Selection of material with a tailored CTE** – to modulate overall probe card thermal movement close to wafer
  - **Customizable build scales** – place probes using build scales that differ between X and Y



2) HBM die stacks can exceed chuck temperature due to joule heating, especially with long test times or high power test steps

- **The probe card should not react quickly to this *transient* excess heating or there may be a scaling mismatch – it needs to respond slowly and minimally to maintain a tight probe to pad alignment.**





Source: FFI R&D team 2024

Normalized probe card temperature rise due to transient heating – **slower/less steep** is the desired response to maintain tight probe to pad alignment.

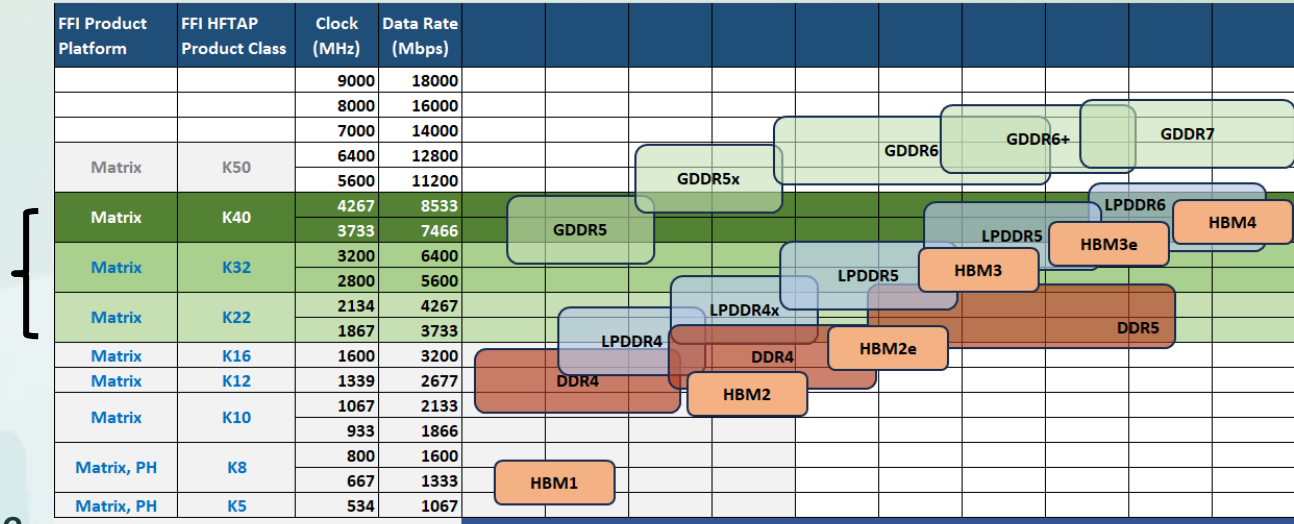
How Does Advanced Packaging Impact Test Speed for HBM

Challenges for High Speed testing (>4GHz)

- **Pre-singulated or post singulated test via sacrificial Pads**
  - Base logic die, core die, and stacked die
  - Test at native HBM operating speed
- **Challenges**
  - Probe design and layout, typically very dense designs
  - Signal routing for high-speed performance
  - Stacked wafer thermal expansion and warpage
- **Solutions and considerations**
  - Electrical: need a probe card that has superior electrical performance, with low insertion and return loss. Enabling higher clock speeds, nearing RF frequencies.
  - PDN is critical with HBM – typically higher current per contact, with higher power hungry designs.
  - Throughput: higher parallelism – better test efficiency, in lower speed test use x2 signal splitting to increase parallelism with T5503 native 64// to 128// or higher, balanced with SI and PI performance.
  - Mechanical: with sacrificial pad test, avoid micron-bump damage using probes with narrow probes.
  - Coupled with dual temp capability using thermally match CTE.

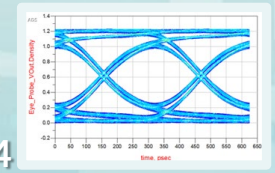
Green band is sweet spot for optimized at speed test

FFI Product Platform	FFI HFTAP Product Class	Clock (MHz)	Data Rate (Mbps)
		9000	18000
		8000	16000
		7000	14000
Matrix	K50	6400	12800
		5600	11200
Matrix	K40	4267	8533
		3733	7466
Matrix	K32	3200	6400
		2800	5600
Matrix	K22	2134	4267
		1867	3733
Matrix	K16	1600	3200
Matrix	K12	1339	2677
Matrix	K10	1067	2133
		933	1866
Matrix, PH	K8	800	1600
		667	1333
Matrix, PH	K5	534	1067

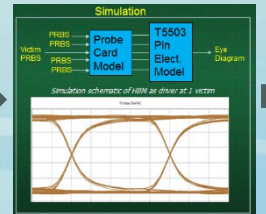


Source: FFI Marketing 2024

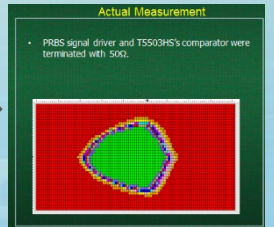
Simulation



Measurement



Device validation



Alternatives to Wafer Test

Test Alternative	Cost of Test	Comments
Individual Die Test	High	Available with single site probe card or socket
Vertical probe card x2 or x4 //	Medium	Available with probe card, but higher TCoO, due to low efficiency and number of touches required
Individual die test in an array	Low - medium	Available near future – customer evaluations ongoing

Customer test flows are still evolving

Ask to industry:

Collaboration with probe card suppliers for optimized lowest cost of test that will ensure 100% KGD – solutions are evolving even as we speak here today…

Conclusions and Summary

- Challenges for Stack Die Test
  - Pad Size and Pitch Shrinking (current versus future sizes)
  - Wafer temperature increases with each new version of HBM
  - Yield drop with advanced packaging

- Solutions
  - Composite CTE matching!
  - Thermally matched to KGSD wafer for tight alignment – probe card to wafer - performance.
  - As pad size and pitch reduce
  - High Frequency test or at speed test
  - >K16 use High Frequency Test at Probe
  - <K16, use x2 signal splitting module to increase parallelism from T5503 native 64 // to 128 //.
  - Test at Wafer level or die level to improve yielded packages

Thank you for help with this presentation: Patrick Rhodes, Yole Research, Kalyanjit Gosh, Mark Ojeda & **John Muir**

Thank you