COMPASS

a **FormFactor** users' group conference

FormFactor™

Verification of Singulated HBM2 Stacks with Die Level Handler, and Review of Wafer Level Sort Challenges

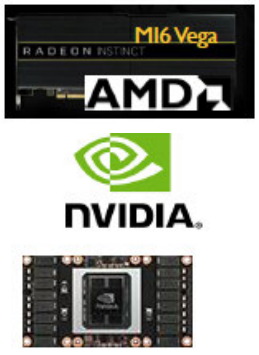
Alan Liao
Director, Probe BU Product Marketing

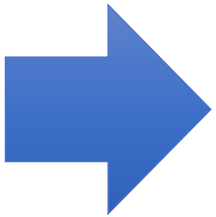
Agenda

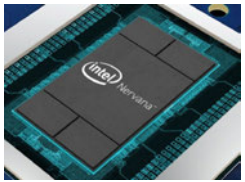
1. Aggressive Adoption of HBM Memory
2. HBM process flow and test insertion review
3. Singulated Stack Direct micro-bump probing challenges
4. Sacrificial DFT pad probing challenges
5. Future work

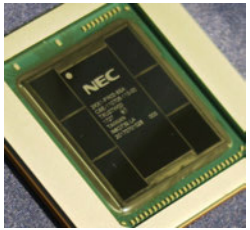
HBM Adoption Rate- More Bandwidth-Hungry Applications

- Strong Market Demand for High Bandwidth Memory – Just the beginning!
  - 2.5D/3D advanced packaging enabled a new generation products/solutions for graphic, AI, and deep learning – AMD, NVIDIA, Intel, Google, NEC, Fijitsu
    - Recent report from “ResearchAndMarket” with bold forecast for HBM & HMC
      - $1B in 2018 with path to $3.8B by 2023, a 33% CAGR
    - Applications expanding beyond Graphic into AI, Servers, and Supercomputer









AI Accelerators with integrated HBM

- Google Tensorflow
- NVIDIA Tesla
- Intel Nervana
- Intel Stratix 10
- Xilinx Virtex UltraScale
- NEC Supercomputer
- Baidu

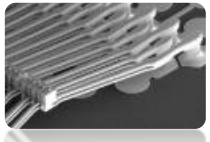
High Bandwidth Memory – Integrated SiP Test Insertions

Enabling HBM and Silicon Interposer Testing

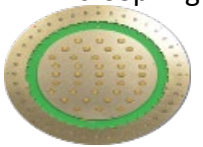
Pre-Singulation Al pad **PRODUCTION** test

Post-Singulation uBump **R&D** test

Matrix Probe Card



T11 MicroSpring



2.2GHz K22

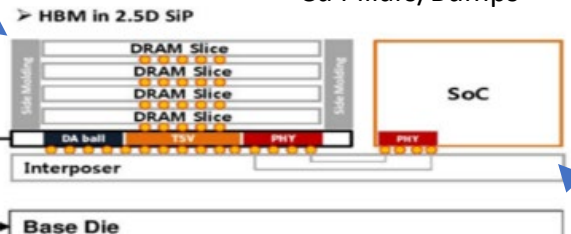
Vertical MEMS MF40

DRAM

Al/Cu Pads

SoC

Grid-array Sacrificial Pads
Cu Pillars/Bumps

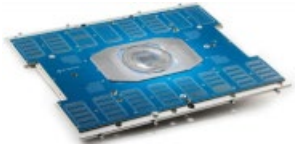


Base Die

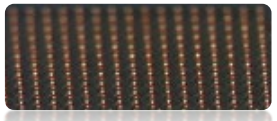
Direct Access Al/Cu Pad
Critical I/O uBumps

Si Interposer

Grid-array uBumps

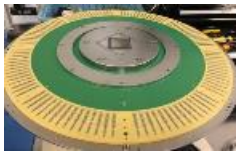


Apollo Probe Card



Vertical MEMS MF60 – MF130

Vertical MEMS MF40



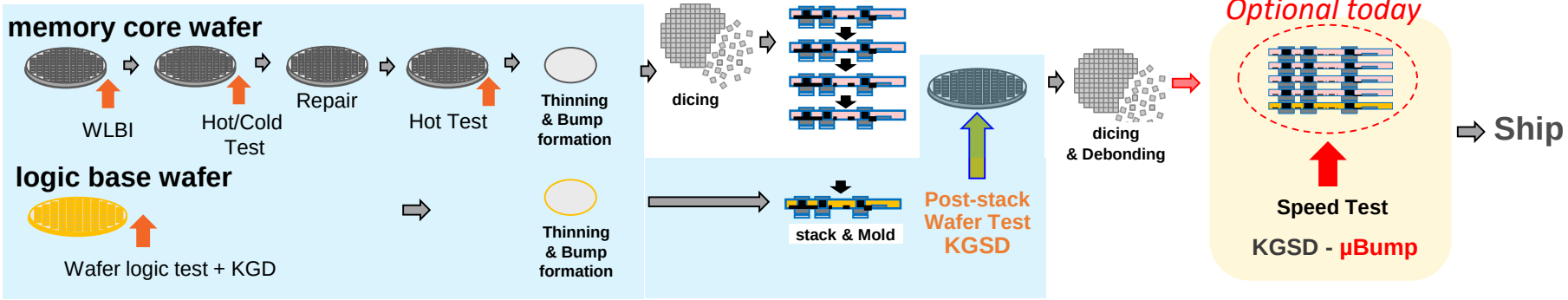
Altius Probe Card for Si Interposer

HBM Process Flow and Test Insertions

Pad-probing test insertions

µ-bump probing

Optional today



Key challenges – Pad insertions

- CTE variation of the stacked wafer
- CTE variation between various stack configuration
- Wafer shrinkage/pad location changes post stack

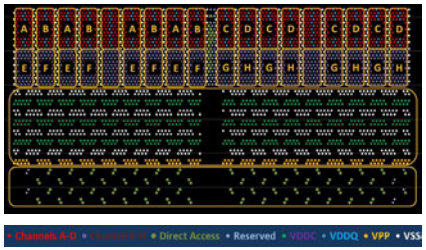
Key challenges - µ-Bump insertion

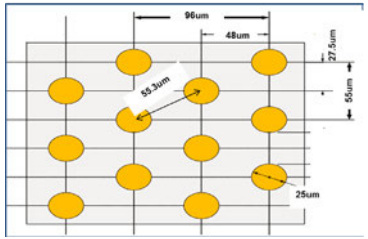
- *Handling of bare stack die*
- *Thermal movement*
- *Contact stability at elevated temperature*
- *Micro-bump "coining" behavior at high temp*

Direct Micro-bump Probing: Key Design Challenges

Probe Technology

- Sub-50um pitch MEMs Probe (48um x 55um)
- 10um tip XY error
- High CCC
- Low K constant
- CRES and Life time

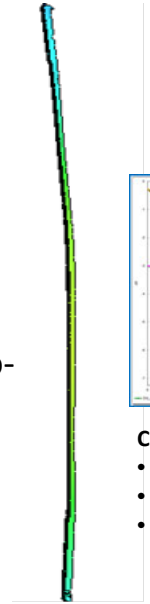


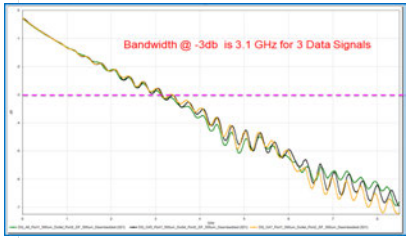


HBM JESD235A

Electrical, Signal & Power Integrity Requirement

- Support at-speed testing > 2.4Gbps
- ~6x3mm die size
- Reliable contact to ~5000 & ~10,000 micro-bumps
- STF design and manufacturing
- Impedance & X-talk optimization
- Maximize performance margin

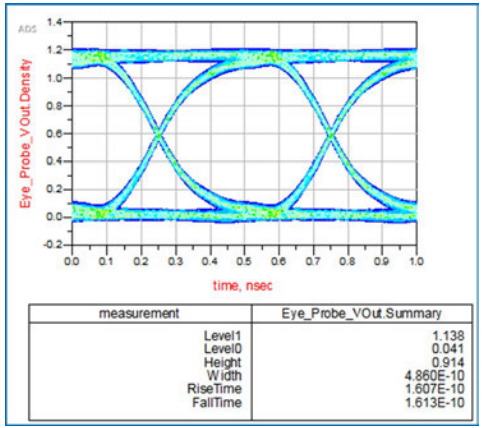




Condition

- Rise time 100pS, 20 to 80%
- 1V swing, no pre-emphasis
- VNA measured data used for eye-diagram simulation @BGA side with 1pF termination

1.0GHz/2133Gbps Measured

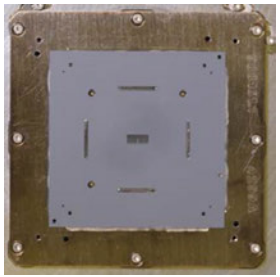


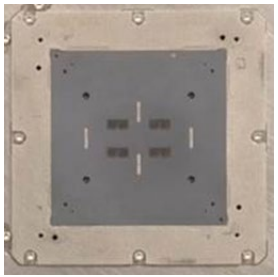
measurement	Eye_Probe_VOut.Summary
Level1	1.138
Level0	0.041
Height	0.914
Width	4.860E-10
RiseTime	1.607E-10
FallTime	1.613E-10

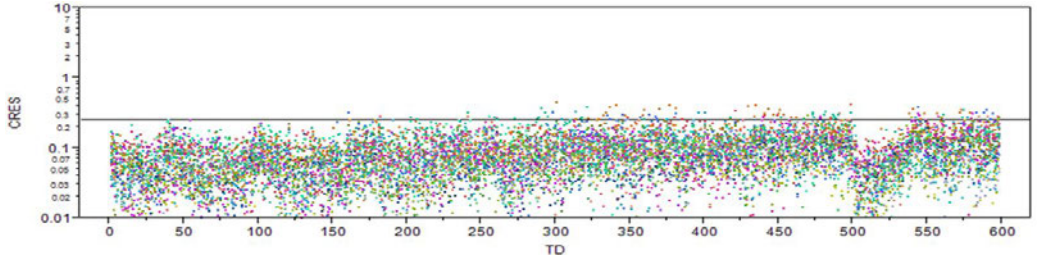
Direct Micro-bump Probing: Application Challenges

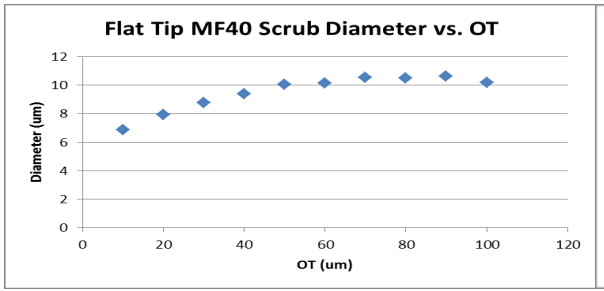
- **Parallelism**
  - X1 and X4 Configuration
- **µBump "coining" d/D behavior**
  - Spring K designed to minimize bump damage across OT range
- **Low K Stable CRES MEMS spring – minimum K to achieve good CRES**

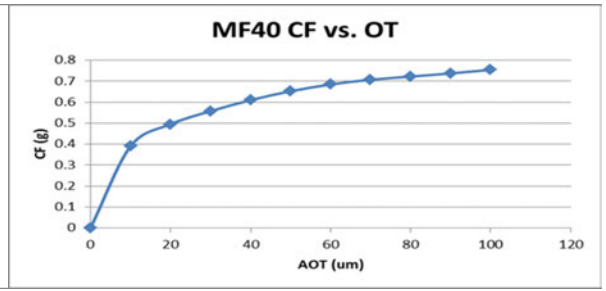
Cleaning Media	ITS PL-1AH (1um grit)
Cleaning motion	Z-only
Cleaning Over Drive	75µm
TD at each cleaning	10
Device TD between cleaning	20

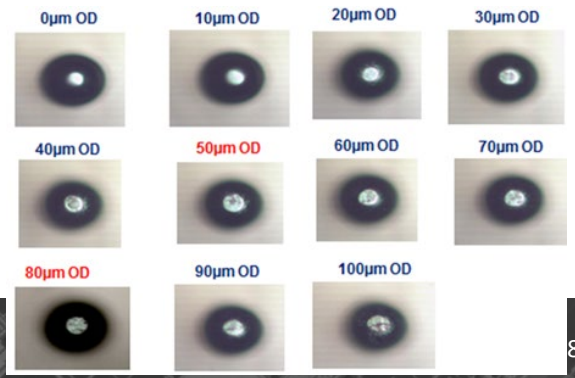












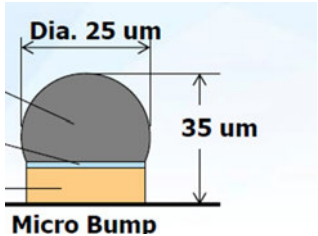
Direct Micro-bump Probing – HBM2 KGDS Test Result

- **We succeeded in contacting all I/O pins**
- **Ambient scrub mark and result**
  - Contact Time:6sec, Contact : 1 time vs 2 times
  - Contact Time:600sec, Contact : 1 time vs 2 times

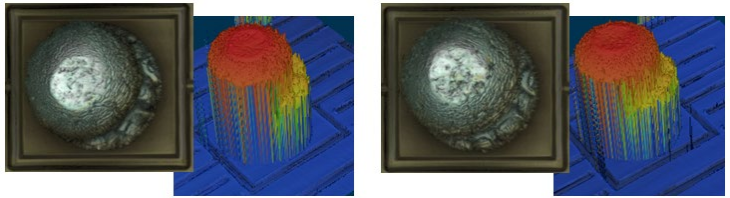
-The scrub becomes deeper as the number of contacts increases

-The scrub becomes deeper as the test time becomes longer

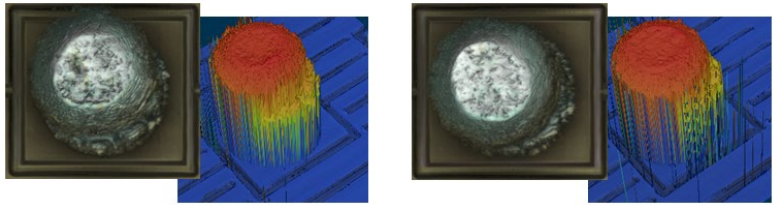
uBump Diameter : 25um
Over Drive : 60um
Temperature : Ambient



Condition	T.T:6sec 1 time	T.T:6sec 2 times
Scrub depth[um]	0.87	1.72
Scrub diameter[um]	10.86	10.86



Condition	T.T:600sec 1 time	T.T:600sec 2 times
Scrub depth[um]	2.61	2.99
Scrub diameter[um]	14.81	15.04

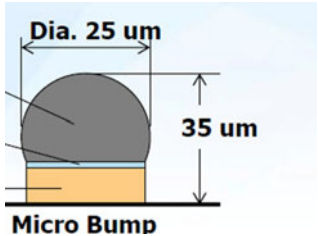


Direct Micro-bump Probing – HBM2 KGDS Test Result

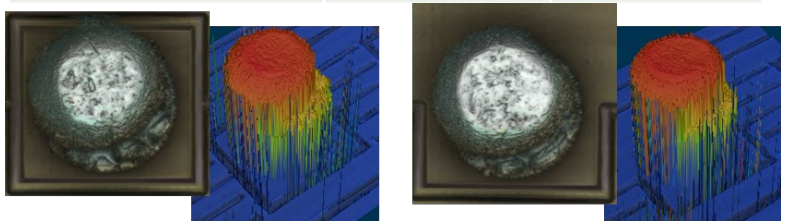
- **We succeeded in contacting all I/O pins**
- **High temperature scrub mark and result**
  - Contact Time:6sec, Contact : 1 time vs 2 times
  - Contact Time:600sec, Contact : 1 time vs 2 times

-The scrub becomes deeper as the temperature becomes higher (i.e., 85C, 95C, 105C)

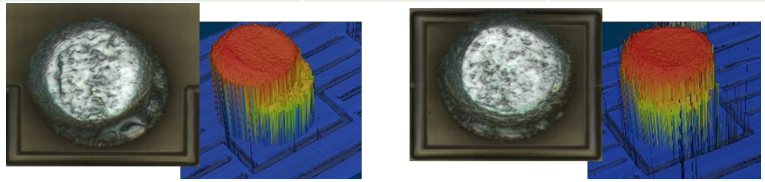
uBump Diameter : 25um
Over Drive : 60um
Temperature : **105degC**



Condition	T.T:6sec 1 time	T.T:6sec 2 times
Scrub depth[um]	1.66	1.84
Scrub diameter[um]	14.34	16.07

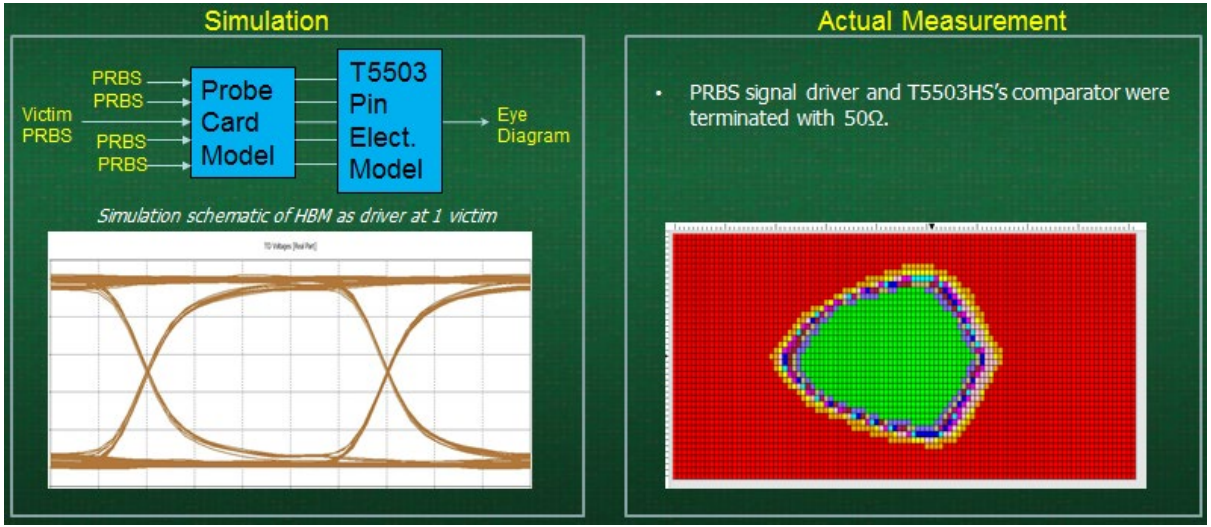


Condition	T.T:600sec 1 time	T.T:600sec 2 times
Scrub depth[um]	2.80	3.86
Scrub diameter[um]	17.06	18.71



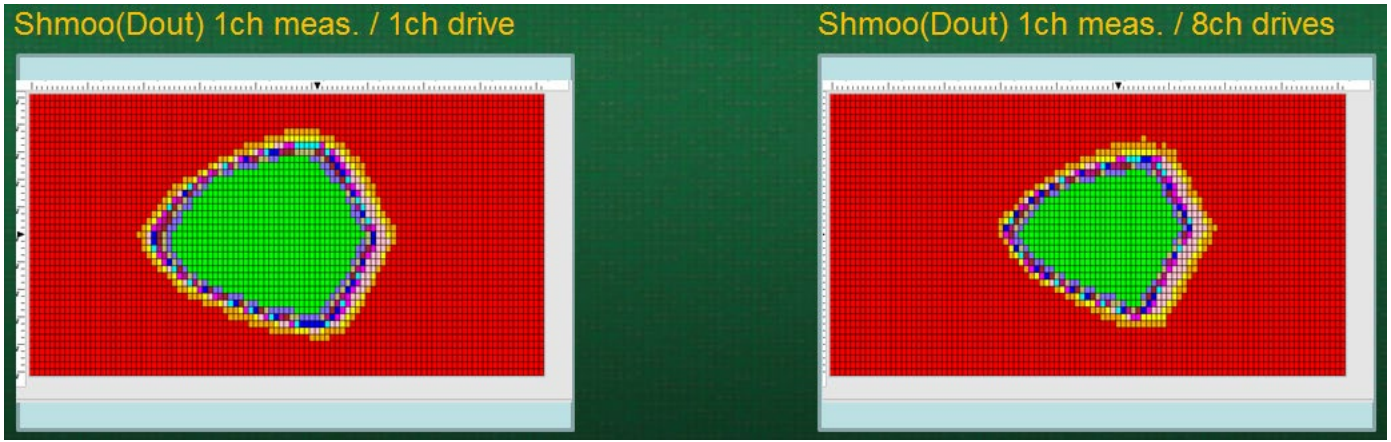
Signal Output/Input Performance on HBM2 Die

- **We Simulation vs Actual Measurement result @ 2Gbps**
  - The waveform is similar in simulation and actual measurement on HBM2 die
  - Strong eye-diagram performance correlation



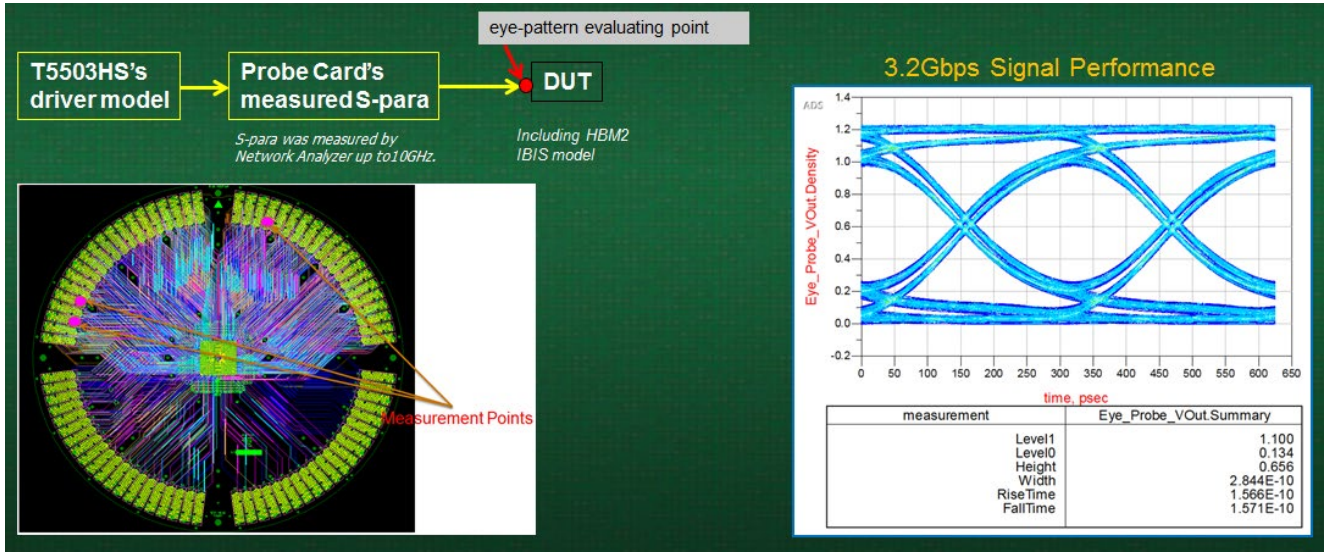
Signal Output/Input Performance on HBM2 Die

- **1ch drive vs 8ch simultaneous drive actual result @ 2Gbps**
  - With data activity on just one memory channel the output data eye width is quite large.
  - With data activity on all eight memory channels the output data eye shrinks.



Signal Output/Input at Higher Frequency

- 1.6GHz/3.2Gbps simulation result
  - MF40 technology supports operating speed to 3.2Gb/s with additional design rules optimization
  - Strong simulation versus actual measurement result as validated through ATE at 2Gbps



Micro-Bump Test Benefit Summary

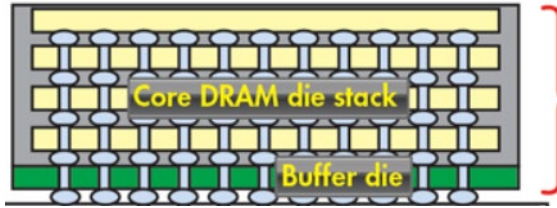
- Working together as a team Advantest together with FormFactor developed a production worthy tool for confirming Known-Good Memory Stacks with ~4,000 micro-bumps and < 60um bump pitch.
- The resulting design exceeded our design goals for probe force and CCC with a wide operational temperature range.
- The solution exceeded our high frequency goal demonstrating >3 Gbps performance.
- The solution contacts to all eight HBM channels simultaneously enabling native mode performance and functional testing of these complex devices.

Sacrificial Pad “DFT”– Pre-singulated Probing

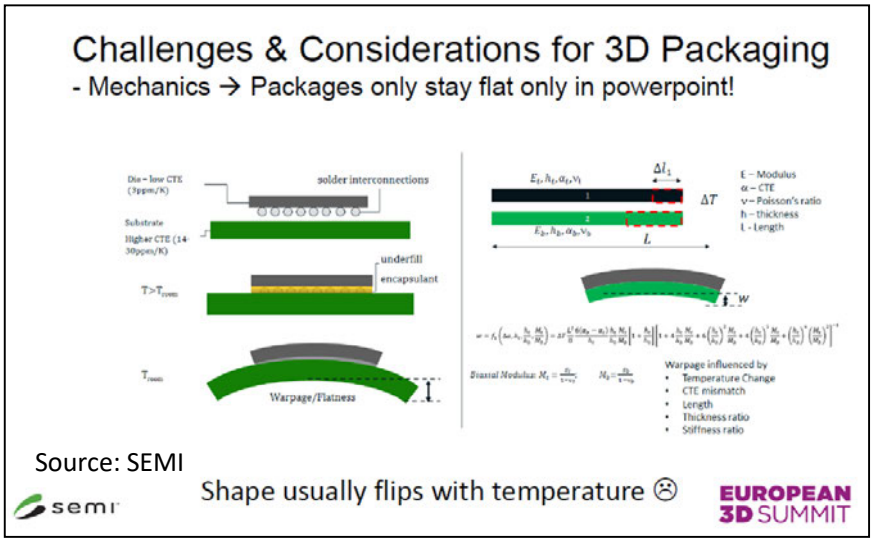
- FormFactor is also a leading probe card supplier for the sacrificial pad probing insertion on HBM pre-singulated wafer and stack wafer–
  - FFI SmartMatrix 1500XP and SmartMatrix 2000XP are the main products for probing HBM base die, HBM core die, and final HBM stacked wafer in 4Hi and 8Hi configurations.

“DFT” Pad Probing Challenges of HBM Stack Wafer

- Stacking Wafers and The Expected Thermal Induced Challenges of Composite Wafer

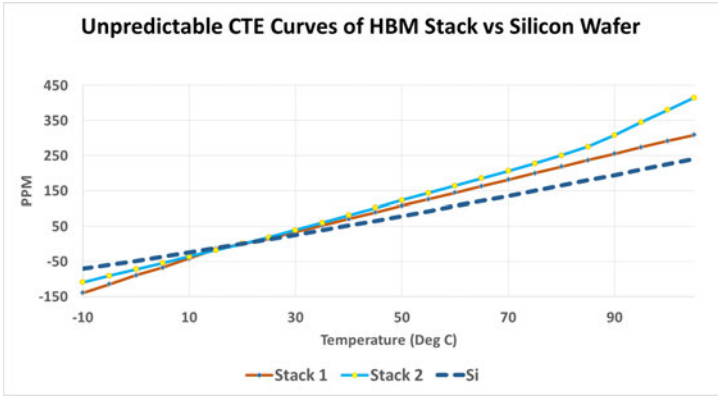


Under-fill & molding compound

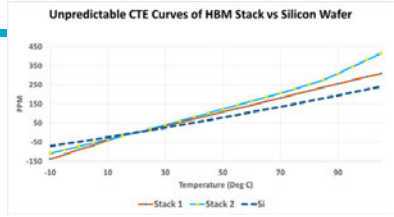


Key Challenges on Composite Wafer

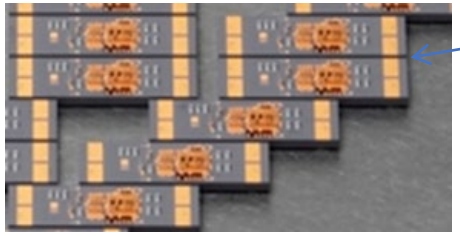
- *Wafer Warpage- pad XY coordinate changes*
- *Wafer CTE changes vs Silicon wafer*
- *Wafer CTE variations between stack configurations*
  - ➢ Basically a “moving target” from the probing perspective



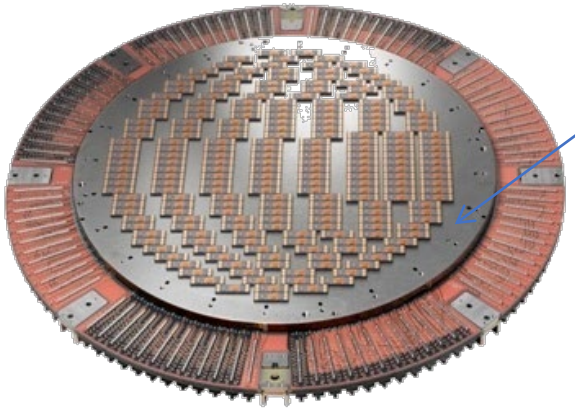
Addressing the “Moving Target” Concern of HBM Stack Wafer



- **SmartMatrix DUTLet based solution for probing HBM stack wafer**



- DUTLet positioning or placement flexibility
  - *DUTLet XY position scaled to match composite wafer CTE*
  - *Automatic pick-and-place for the needed precise accuracy*
  - *Like the individual spring/probe, the DUTLets are individually replaceable for best manufacturing yield*

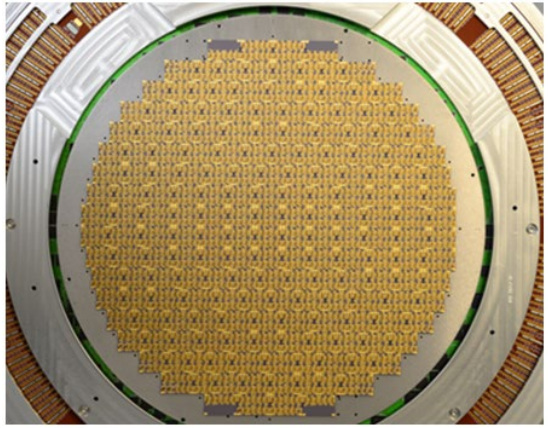


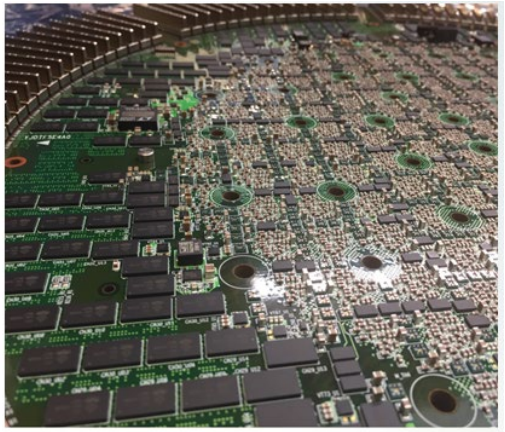
- **The wafer side stiffener (WSS) substrate material flexibility**
  - *DUTLet are attached to a WSS metal substrate*
  - *WSS substrate advantage – the ability to select desired material with CTE that matches the various composite wafer configuration*
  - *Enables probe card to precisely tracks wafer expansion at hot and cold temperature*
  - *Single temperature and Dual temperature operation*

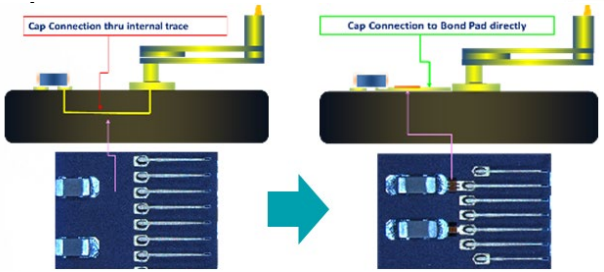
SmartMatrix 2000XP – for HBM Base, Core, and KGSD

1TD ~2000DPW, ~130,000 pins probe card

- Enabling Technologies
  - Micro-strip DUTLet ceramic
  - MW PCB with advanced routing
  - FFI ATRE : X16-X22TRE, XDC-Boost
  - Advanced capacitor attach process
    - 33% footprint reduction
    - DUTLet surface routing capability
      - PI enhancement
- Known-good-stack high frequency test probe (HFTAP)
  - Max parallelism limited by ATE
  - 1.6GHz (3.2Gbps) for HBM2
  - 2.2GHz (4.45Gbps) for mobile and commodity DRAM



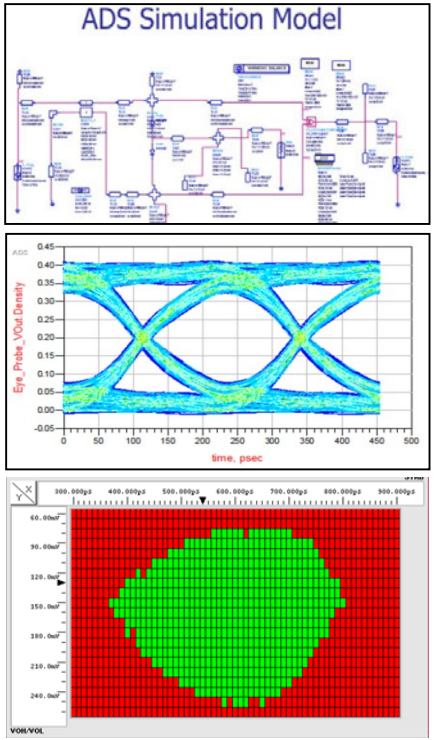




- 100k to 130K pin/PC
- 55 to 70 pin/DUT
- ~4.2 x 8mm die size
- 125MHz (200 with TTRE)
- ~30K total net count
- up to 400kg force nominal
- HFTAP parallelism limits by ATE

FFI High Frequency Probe Card Roadmap for HBM Memory

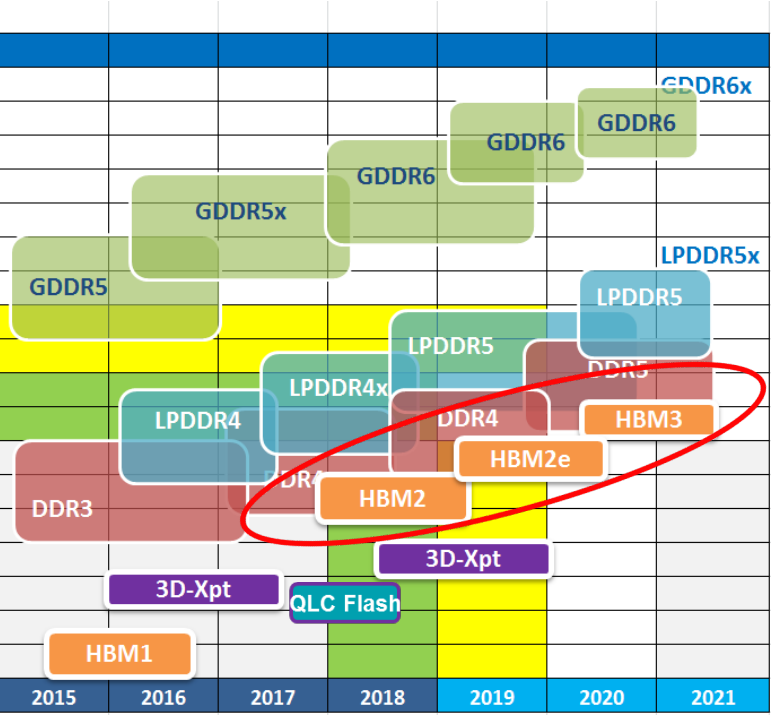
K22 Qualified & in Production



- Enabling high frequency test at probe on mainstream memory

		Target Test Speed	
Platform	FFI "HFTAP"	MHz	Mbps
		9000	18000
		8000	16000
		7000	14000
		6400	12800
		5600	11200
TBD	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

*K32 in feasibility study



Future Work Considerations

- Handler
  - Going to multi-die

- Probe Card
  - Micro-bump probing - going beyond X4 parallelism and at native speed
  - “DFT” pad probing – high frequency at probe up to 3.2GHz (6.45Gbps)