Hybrid Bonding (Bump-less) Enabled 3DIC: External Products Landscape

Prashant Majhi

Stephen Morein

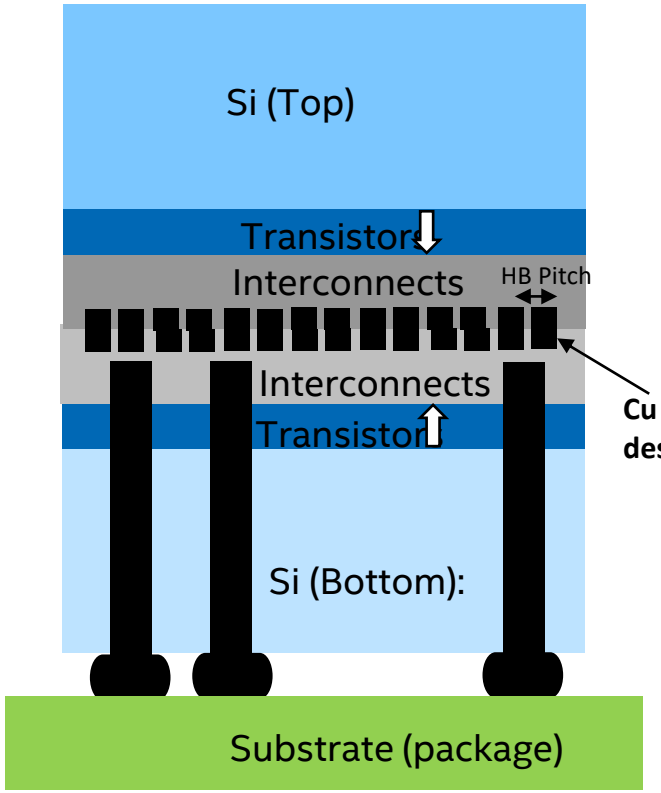
intel®

Exec Summary

- This presentation
  - Discusses External Product Landscape enabled by (or driving) Hybrid Bonding based 3DIC
  - Does not discuss internal product roadmap/activities exploring HB

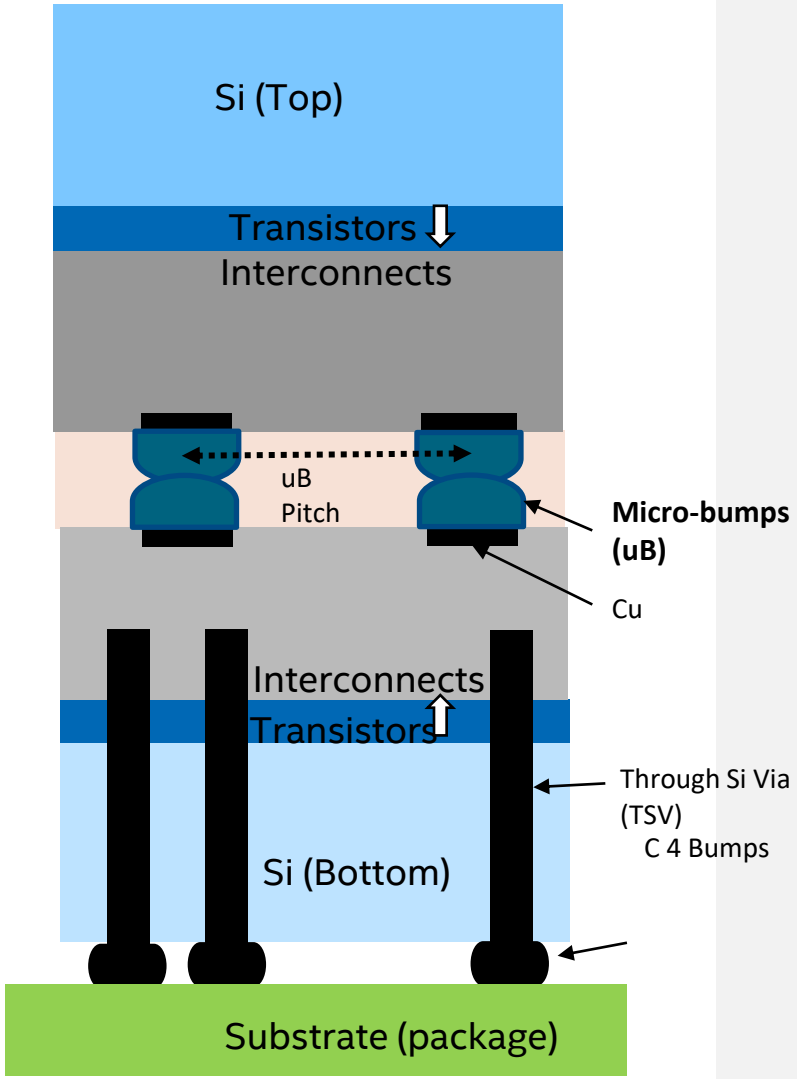
- Key Takeaways
  - Scaling 3DIC: Micro-bumps to bumpless Cu-Cu Hybrid Bonding for
    - Increased density of connections (#/mm2)
    - Improved D2D parasitics
  - Several external products use Hybrid Bonding @ HVM [Mostly WoW but CoW starting to mature as well]
  - Expect competition to start adopting HB enabled 3DIC for HPC, Graphics, FPGA and AI products
  - External eco-system enabling hybrid bonding (process technology, EDA, test/metrology …) developing actively
  - Achievable yield (and associated costs) for Hybrid Bonding continues to be strongly debated

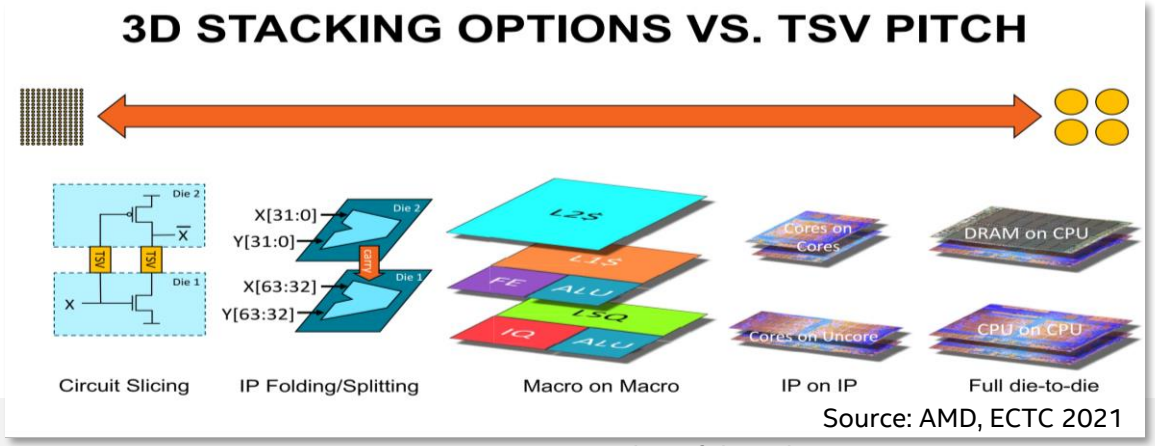
3DIC: Micro-Bump to Hybrid Bonding



Scaling Pitch for Increased D2D Signals and Lower Latency / Energy (I/O Wall)

Cu pads (std DD) designed for HB





Industry Landscape in Advanced Packaging

	Wire Bonding	Flip Chip Bonding	2.X, 2.5D (passive Interposer)	3DIC u-bump (Active Interposer)	3DIC-HB (CoW, CoC) [Hybrid Bonding]	3DIC-HB (W2W) [Hybrid Bonding]	3D-IC (M) [sequential /monolithic integration]
NAND			NAND stack TSV/ Wide IO		Excess CuA	CMOS/Array	CUA + Array Monolithic
DRAM			HBM	HBM over logic	DRAM stack	DRAM stack	3D DRAM
HPC/AI SoC			SoC/HBM	Logic/Logic; Logic/SRAM	Logic/Memory	Logic-L1/3 SRAM stack	CFET
FPGA/Network					SRAM/Logic	Fabric/Fabric	
GPU			GPU/HBM	Logic/SRAM	Logic/Memory	SRAM stack	
CMOS Image Sensor					SWIR/ROIC	IS/ISP	IS/DRAM/ISP
Si Photonics			SiP/SoC		Laser/Passive		Laser/Passive
SCM/3DXP					Exploratory Phase		

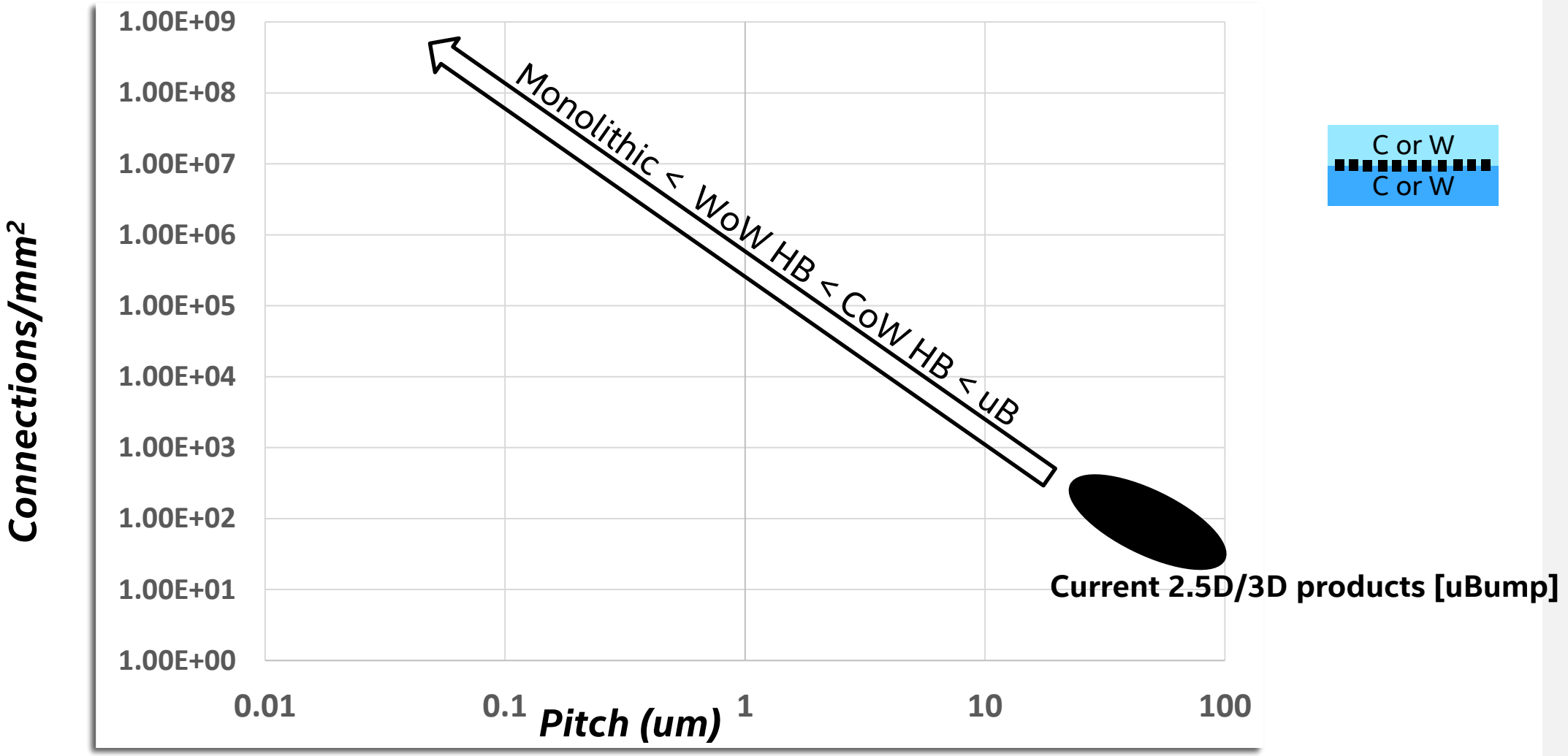
HVM

R&D

CoW: Chip on Wafer; CoC: Chip on Chip; WoW: Wafer on Wafer

No Bumps → Direct Cu-Cu or oxide/oxide + BE Metals

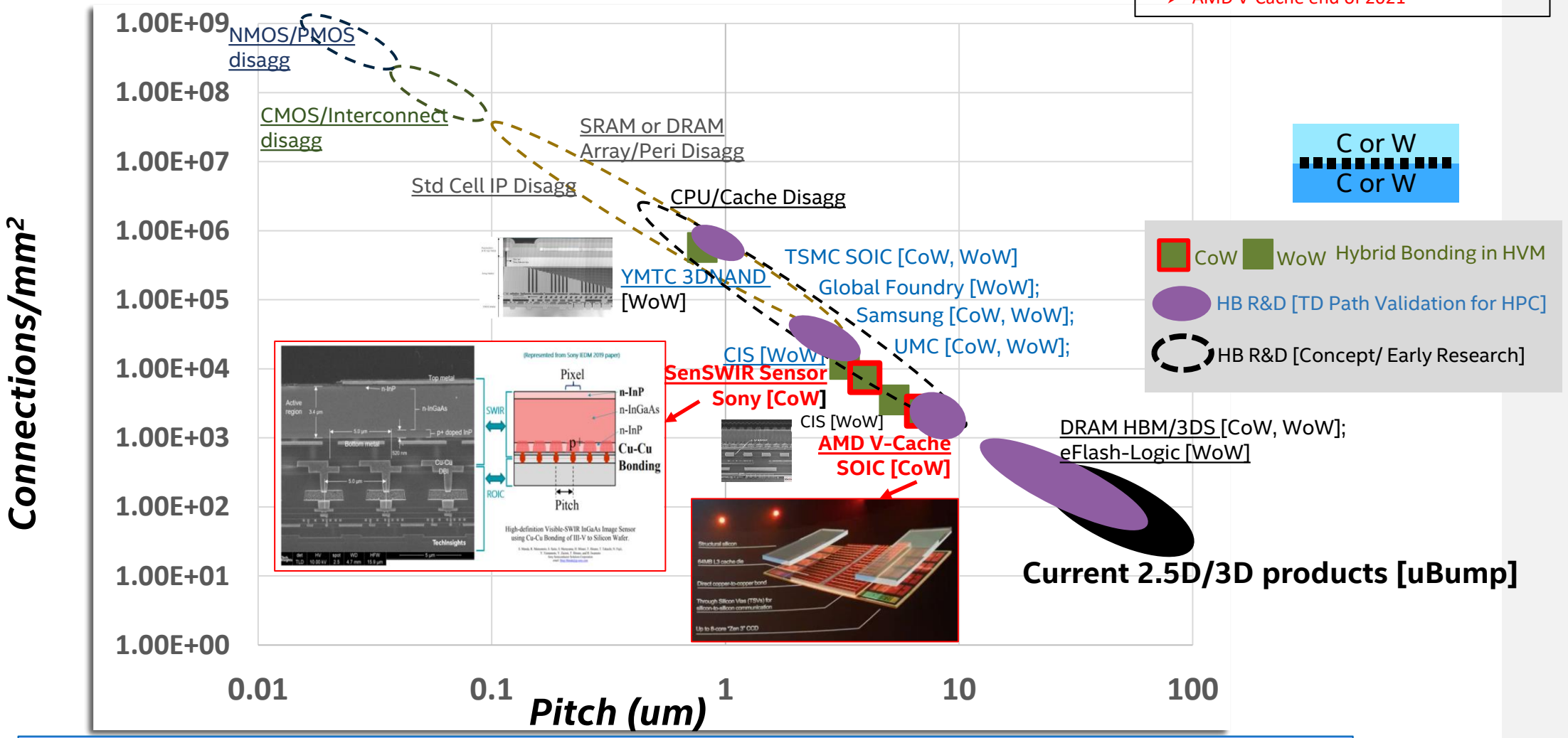
3DIC: Pitch Scaling Trajectory



Bonding Landscape: R&D and HVM [Q2'2021]

WoW Hybrid Bonding in Production since **2016**
- Sony image sensors since 2016
- YMTC 3D NAND since 2020

CoW Hybrid Bonding started Production **2020**
- Sony SWIR Sensor, 2020
- AMD V-Cache end of 2021



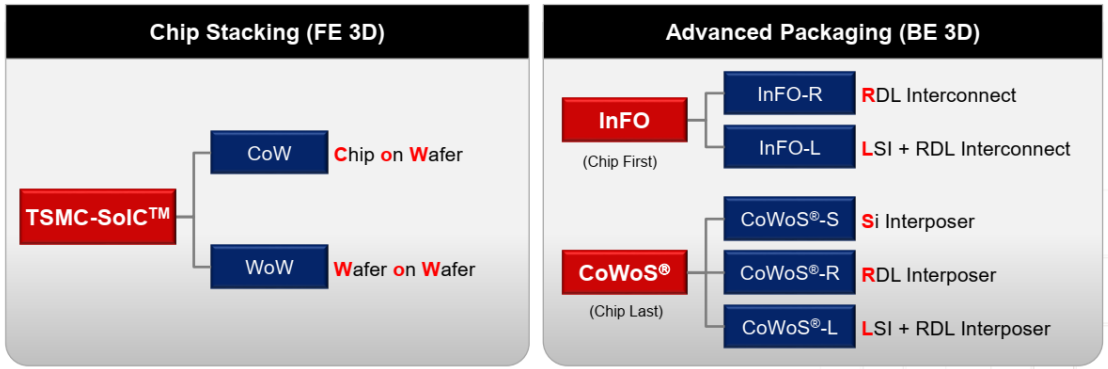
HB (WoW and CoW) mature for image sensors and 3DNAND, getting to HVM for HPC at TSMC

Other Bonding Applications [in Tech Development]

- TSMC: HV and LV Logic for Display Driver and other IOT ⇨
- TSMC/Samsung: DTC or ISC for tightly coupled high Density MIM Cap ⇨
- TSMC: Bonding of Corrugated Si for 3DIC immersion Cooling ⇨
- TSMC: COUPE for Tightly Coupled Si-Photonics [P/E to XPU] ⇨
- GF/Samsung HB for (III-V) Laser on (Si SOI) PIC ⇨
- TSMC/Industry/Academia: Tightly coupled NVM to Logic [inc MRAM] ⇨
- TSMC Immersion In Memory Compute ⇨
- …

TSMC: 3DIC with 3DID Scaling

TSMC 3DFabric™

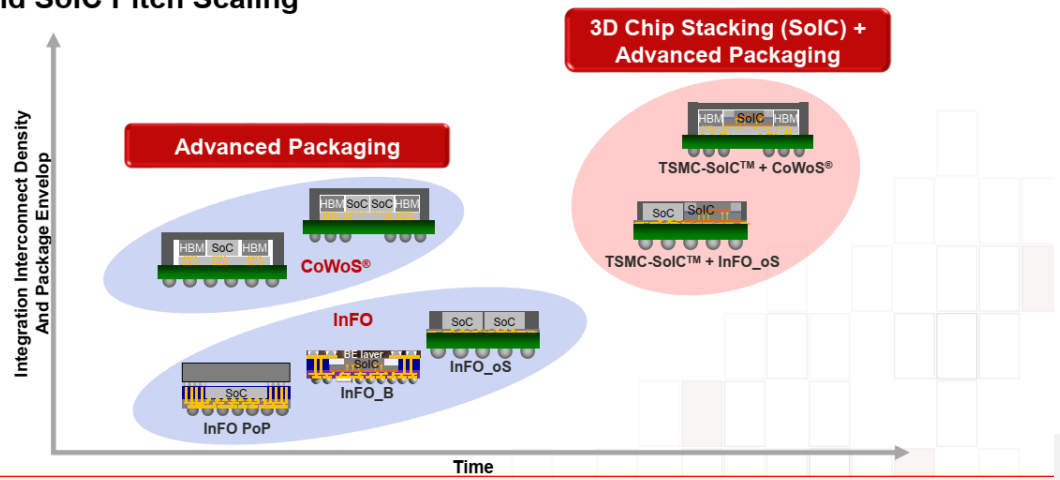


SoIC: System on Integrated Chips

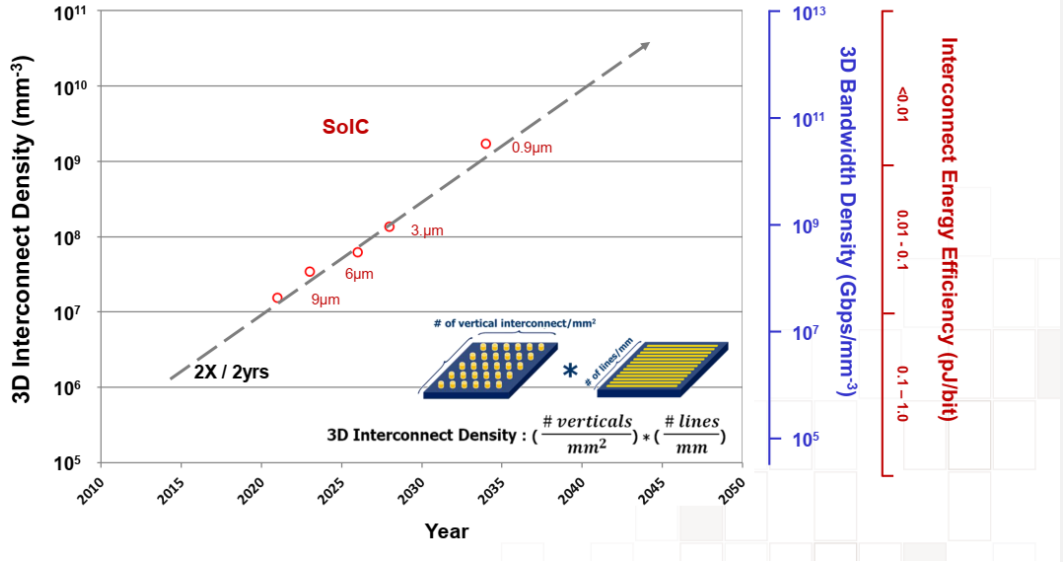
InFO: Integrated Fan-Out
CoWoS: Chip on Wafer on Substrate
RDL: Redistribution Layer
LSI: Local Si Interconnect

3DFabric™ Integration Technologies

- 3DFabrics updates- additional structures, Packaging Envelop Increase and SoIC Pitch Scaling

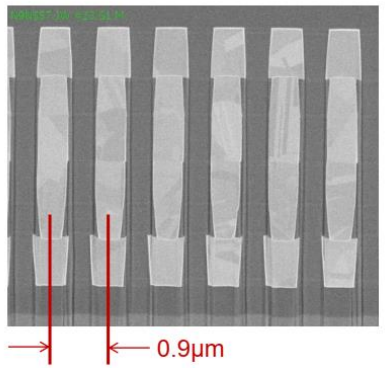


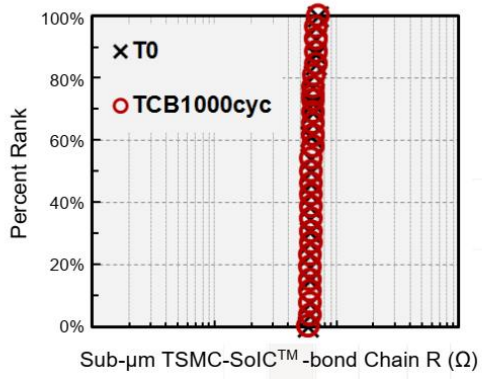
Inter-chip Interconnect Scaling Roadmap



Sub-µm CoW Interconnect Feasibility

- 0.9µm bond pitch stacking
- Highly reliable after TCB 1000 cycle
- Enable direct integration of SoIC/bonding and SoC/BEOL interconnect





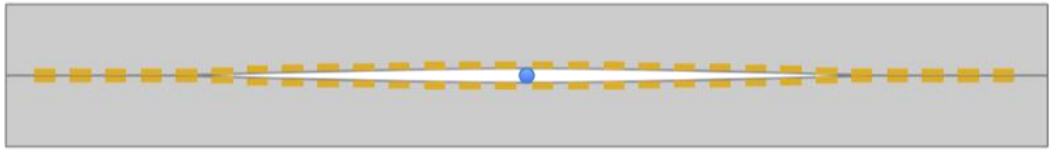
https://www.anandtech.com/show/16051/3dfabric-the-home-for-tsmc-2-5d-and-3d-stacking-roadmap

Commonly “debated” Topics in Hybrid Bonding

- HB yield for WoW and CoW
- HB Integrated Cost/Affordability
- HB Pitch Scaling “Limits”
- HB Test Methodology
- T,F,M readiness: EDA for HB enabled 3DIC designs
- HB re-work strategy?
- …

On Yield for Hybrid Bonding

- Hybrid bonding depends on contact between near perfect surfaces
- A particle can interfere and create a void
  - Void can be 10x to 1000x wider than the height of the particle
  - No connections made in void area
    - 100's of failed pads

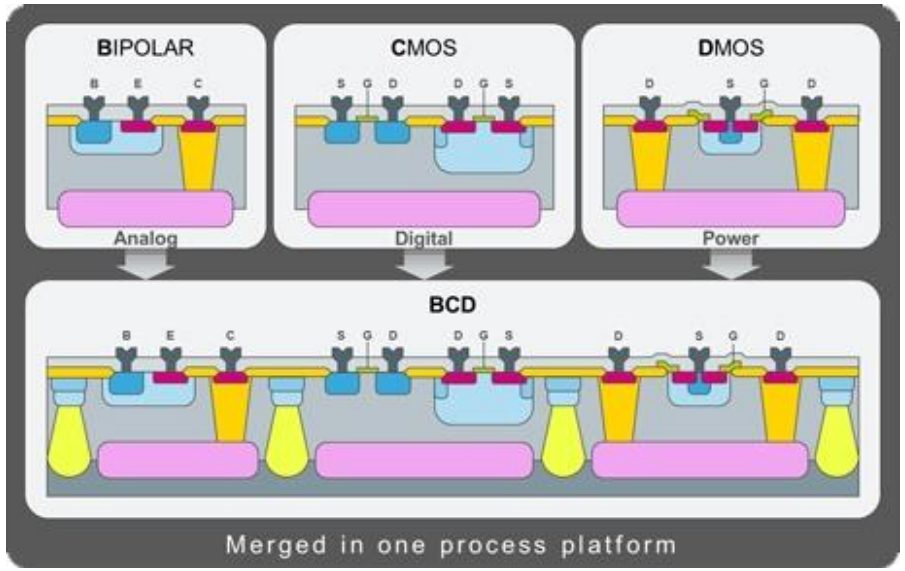


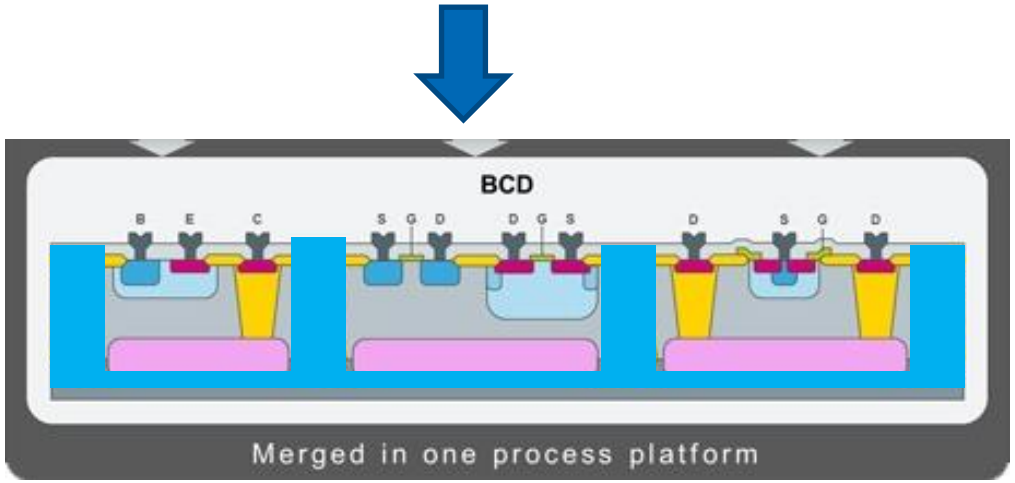
- WoW
  - Wafers kept in cleanroom
  - Clean, CMP, clean, plasma activate, bond
  - WoW is mature and high yield
- DoW
  - Multiple opportunities for contamination, handling activated die, etc.
- What is mature yield for DoW? 90% attach rate? 95% attach rate?

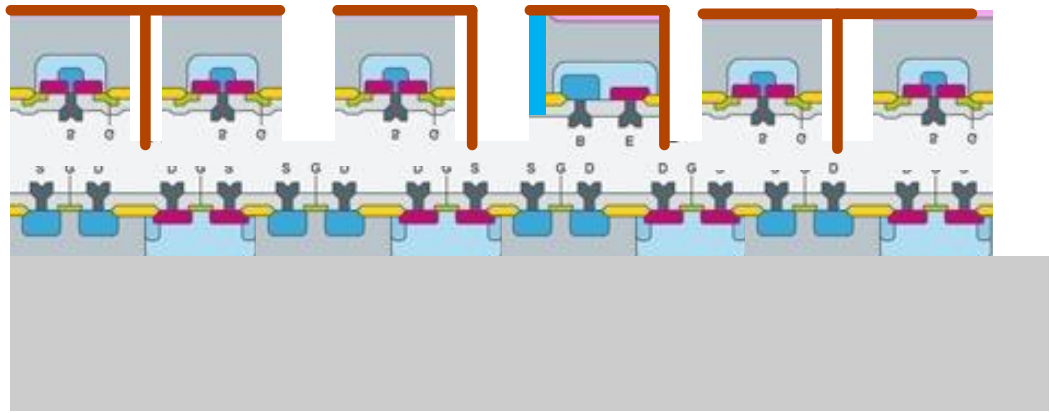
“secondary” uses of HBI

- Oxide/HBI bond is very strong
  - Can be handled like monolithic silicon after bonding
  - BSI image sensors:
    - Sensor wafer thinned < 5u after bonding
  - Dram
    - Wafers thinned without oxide/HBI show 40% decrease in retention time @ 20u wafer thickness
    - Wafers thinned after oxide/HBI showed no decrease in retention time down to 5u thickness
    - Sony in production with DRAM thinned to ~10u (5u silicon)
  - BCD on air….

BCD on air…


BIPOLAR
CMOS
DMOS
Analog
Digital
Power
BCD
Merged in one process platform


BCD
Merged in one process platform



Exec Summary

- This presentation
  - Discusses External Product Landscape enabled by (or driving) Hybrid Bonding based 3DIC
  - Does not discuss internal product roadmap

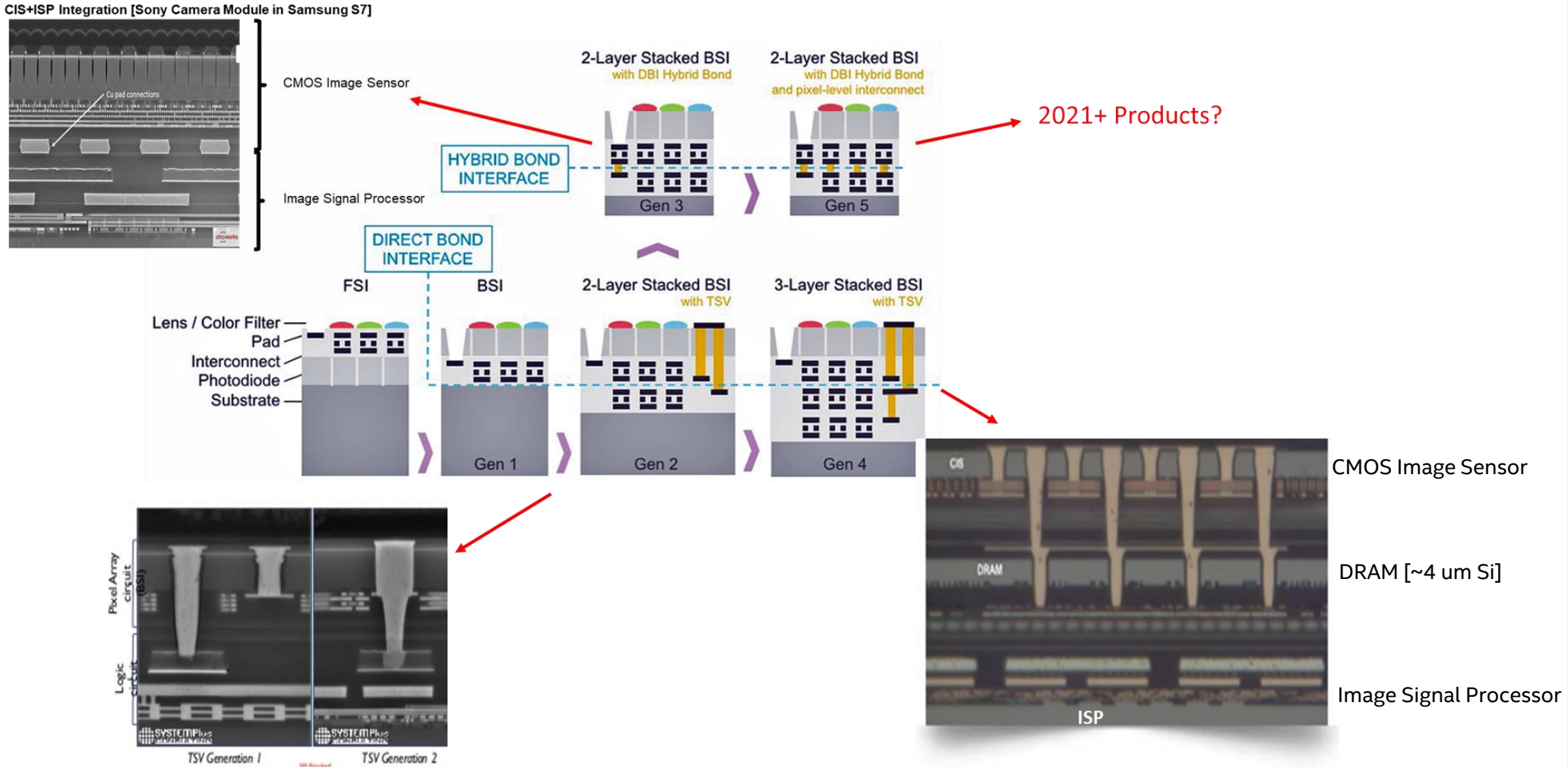
- Key Takeaways
  - Scaling 3DIC: Micro-bumps to bumpless Cu-Cu Hybrid Bonding for
    - Increased density of connections (#/mm2)
    - Improved D2D parasitics
  - Several products use Hybrid Bonding @ HVM [Mostly WoW but CoW starting to Mature as well]
  - Expect competition to start adopting HB enabled 3DIC for HPC, Graphics, FPGA and AI products
  - External eco-system enabling hybrid bonding (process technology, EDA, test/metrology …) developing actively
  - Achievable yield (and associated costs) for Hybrid Bonding continues to be strongly debated

Backup

- Links to Product 1 pagers

HB for CMOS Image Sensor ..... (1/3)

CMOS Image Sensor Roadmap



HB for CMOS Image Sensor ..... (2/3)

Sony: SenSWIR using CoW HB in HVM

Groundbreaking SenSWIR Sensor by Sony - **IMX990/IMX991**

Sony announced the IMX990 and IMX991 SenSWIR imagers in 2020, with a 1.34 MP and 0.34 MP resolution, respectively. By moving away from pixel-level bump bonds and taking advantage of greater miniaturization in Cu-Cu Direct Bond Interconnect (DBI), Sony was able to reduce the pixel size of the InGaAs/ROIC SWIR imagers down to 5.0 µm. This makes the IMX990/IMX991 the smallest pixel-pitch InGaAs-based SWIR image sensors commercially available on the market.

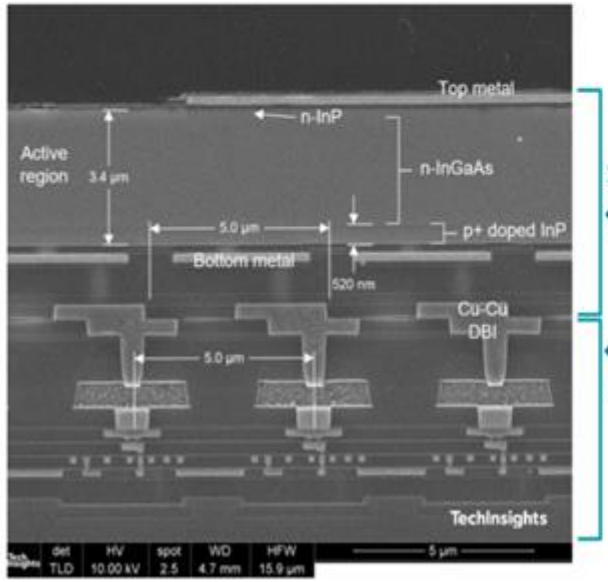
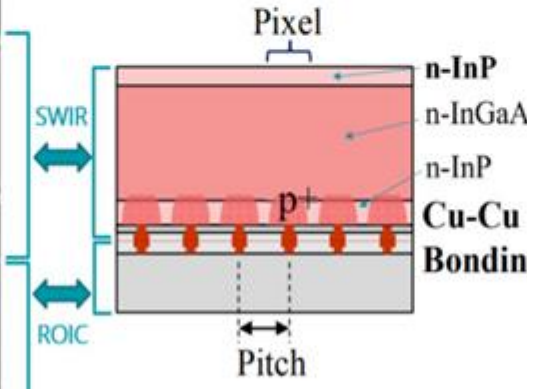


Figure 3 (a): Sony IMX990 SEM Cross Section

(Represented from Sony IEDM 2019 paper)



High-definition Visible-SWIR InGaAs Image Sensor using Cu-Cu Bonding of III-V to Silicon Wafer.

Figure 3 (b): Sony IMX990 Schematic Diagram

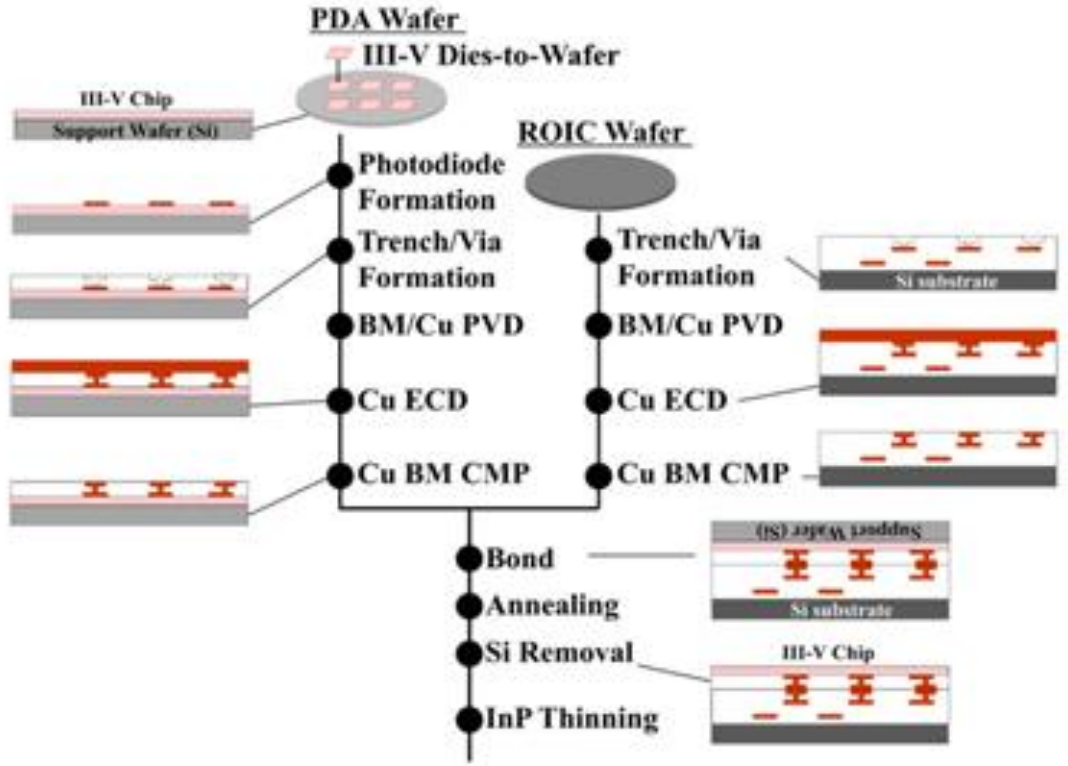


Fig. 6. Process flow of Cu-Cu bonding showing schematic diagrams of representative process steps.

Sony IEDM, 2019

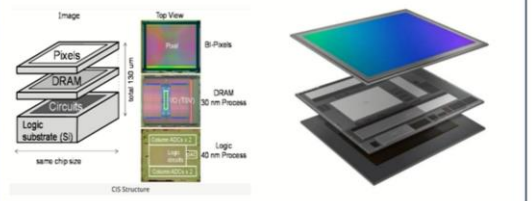
HB for 3 Wafer Stack CMOS Image Sensor …(3/3)

Samsung: On development of 3 Wafer HB

Introduction

- **Multi-Stack Wafer Bonding**
  - CIS market already have 3 stacking experience.
  - Expected market for CIS : **Pixel shrinkage** → Cu-Cu hybrid wafer bonding + 3 stack or more.

3 Stacked CMOS Image Sensor : Using DRAM for fast readout speed



Source : IEDM 2017, Sony

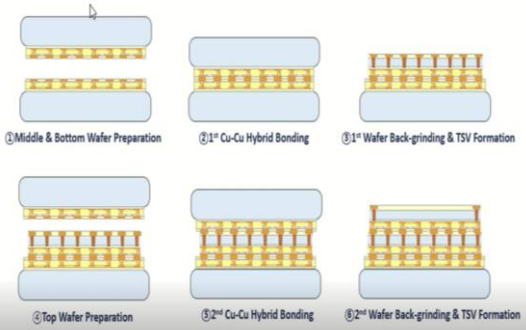
Source : www.samsung.com

SAMSUNG

2021 IEEE 71th Electronic Components and Technology Conference | Virtual | June 1 – July 4, 2021

Key Technologies : Process flow

- **3 Stack Wafer Bonding Process**



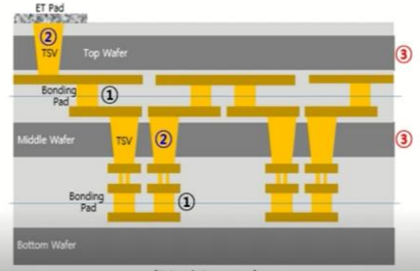
[3 Stack Process Flow]

SAMSUNG

2021 IEEE 71th Electronic Components and Technology Conference | Virtual | June 1 – July 4, 2021 5

Key Technologies : Overview

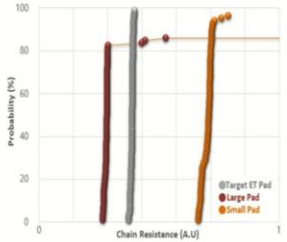
- **3 Key Technologies : CMP / TSV / Edge Engineering**
  - ① CMP : Flatness control of hybrid bonding surface for both wafer front and backside.
  - ② TSV : Stable etch process control for bottom metal landing
  - ③ Edge engineering : Edge treatment to prevent any wafer edge abnormality such as chipping.



[3 Stack Structure]

Connectivity Yield : Cu-Cu Bonding

- **Cu-Cu Bonding Chain**
  - The test pattern is composed of diverse bonding pad sizes. There is still difficulty to acquire good connectivity for all pad sizes.
  - Small and larg bonding pad shows the degradation due to different Cu protrusion behavior while good resistance trend for target ET pad.



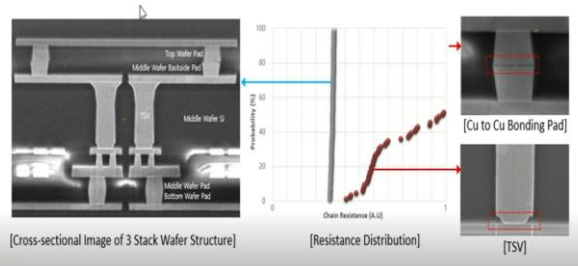
[Resistance Distribution]

SAMSUNG

2021 IEEE 71th Electronic Components and Technology Conference | Virtual | June 1 – July 4, 2021 11

Connectivity Yield : 3 Stack

- **3 Stack Wafer Bonding Structure**
  - Excellent distribution of 3 stacked wafer bonding chain including Cu-Cu Bonding and TSV.
  - 100% connectivity yield has been verified through control of hybrid bonding surface flatness, TSV process optimization and proper wafer edge treatment.



[Cross-sectional Image of 3 Stack Wafer Structure] [Resistance Distribution]

SAMSUNG

2021 IEEE 71th Electronic Components and Technology Conference | Virtual | June 1 – July 4, 2021 12

Conclusion

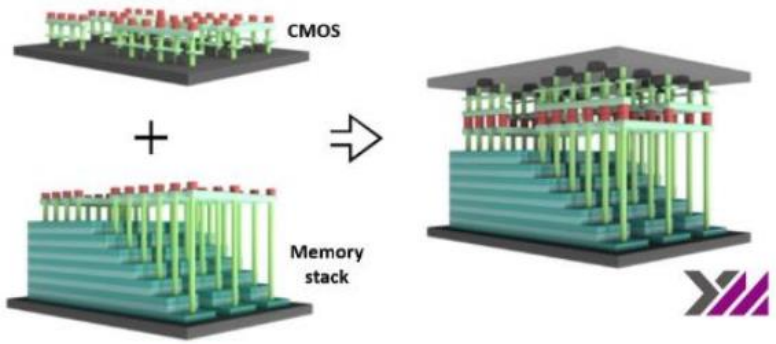
- **Achieved both wafer front and backside hybrid bonding friendly CMP planarization.**
- **Achieved stable TSV process for the back to face bonding.**
- **Defined the proper edge treatment method for multi wafer stacking.**
- **Above processes has been verified through 100% connectivity yield.**

SAMSUNG

2021 IEEE 71th Electronic Components and Technology Conference | Virtual | June 1 – July 4, 2021 13

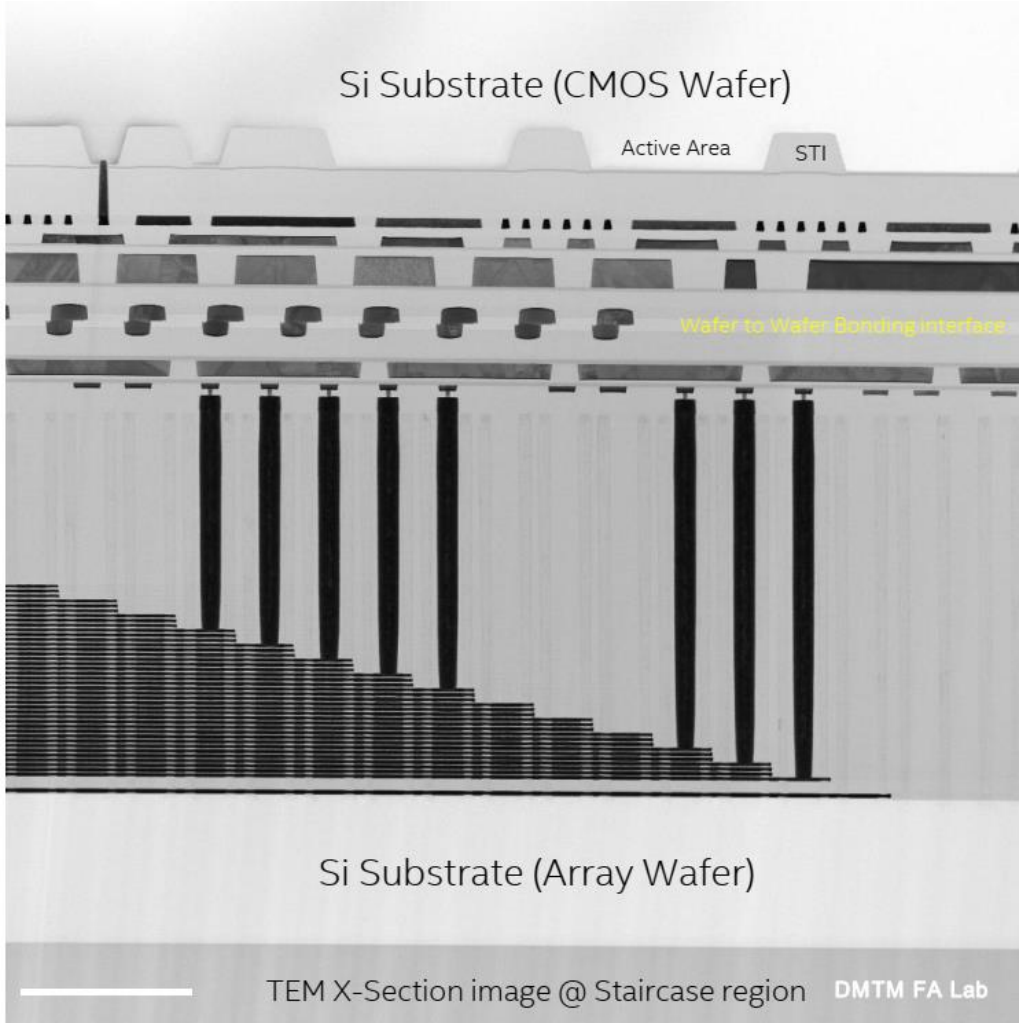
HB for 3DNAND Stacking

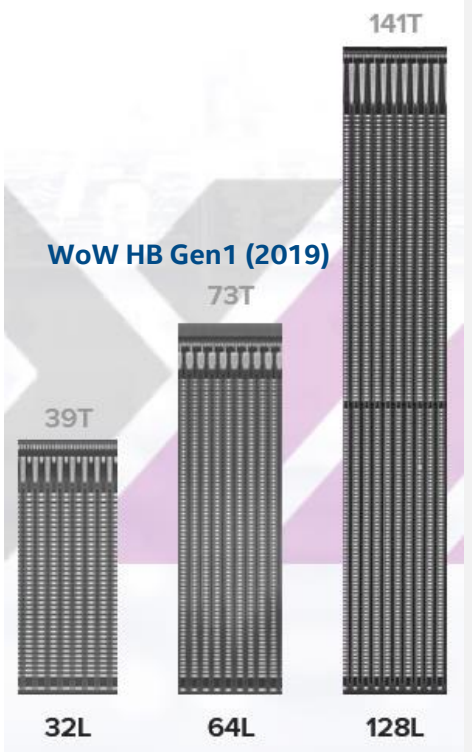
YMTC Xtacking Architecture



Source: YMTC

With Xtacking®, the periphery circuits which handle data I/O as well as memory cell operations are processed on a separate wafer using the logic technology node that enables the desired I/O speed and functions. Once the processing of the array wafer is completed, the two wafers are connected electrically through billions of metal VIAs (Vertical Interconnect Accesses) that are formed simultaneously across the whole wafer in one process step, using the innovative Xtacking® technology, with limited increase in total cost.





HB for NVM (NOR) on Logic:

An approach to embedding traditional non-volatile memories into a deep sub-micron CMOS

TSMC Co., Ltd., 8, Li-Hsin Rd. 6, Hsinchu Science Park, Hsinchu, Taiwan

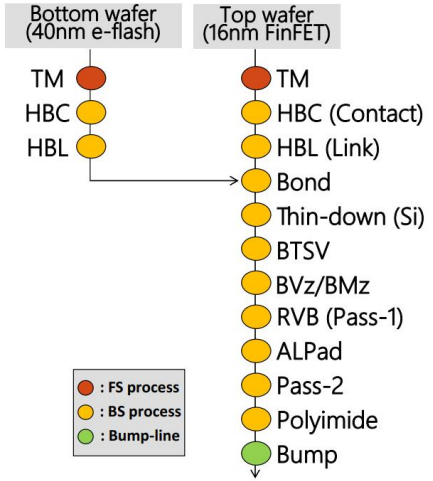
TSMC @ VLSI'20

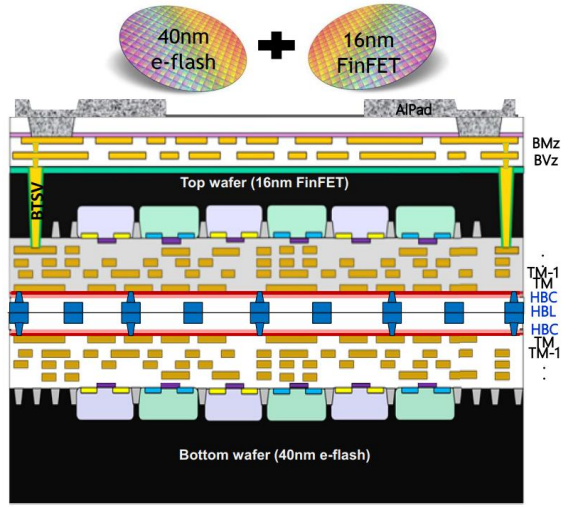
50mm^2 die, F2B

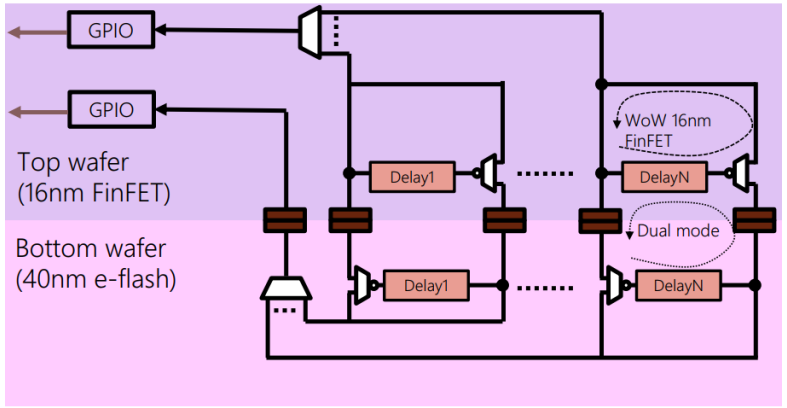
Increasing process complexity and long development cycle time is the critical bottlenecks for implementing e-flash in advanced logic nodes.

Comparison Items	This work	On developing	Available technology	
	WoW 16nm FinFET & 40nm e-flash	SoC 16nm e-flash	SiP External NOR flash	SoC 40nm e-flash
Current status	available	not available	available	available
Delivery complexity	medium	very high	low	medium
CMOS logic node (computing power)	16nm node	16nm node	16nm node	40nm node
Flash data access speed	> 200MB/sec	> 200MB/sec	< 70 MB/sec	> 200MB/sec
IO Bus width	> 32	> 32	< 10	> 32
Reliability	Grade-1 capable 10yr@125C data retention	-	10yr at 85C	Grade-1 capable 10yr@125C data retention
Package	simple	simple	complicate	simple

WoW Brief Process Flow





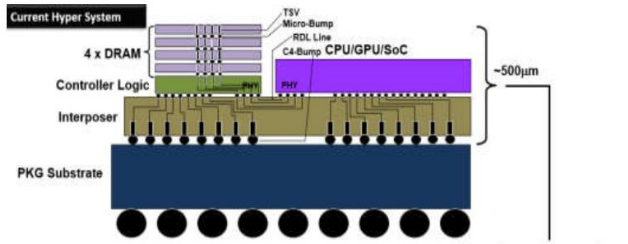


Technology automotive qualified w/ matched eflash/improved CMOS (SM, TC, ..)

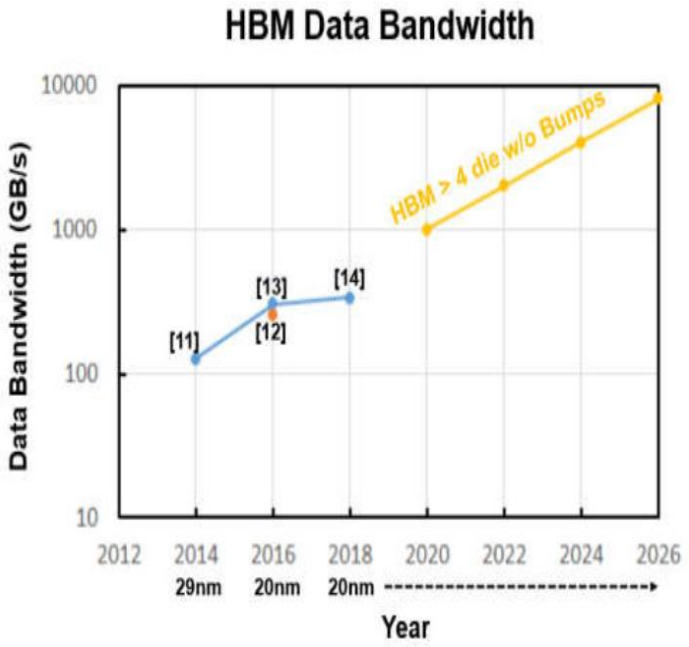
Tech	PROs	CONs
WoW Approach (16nm+40nm e-flash for example)	◆ **Time to market** (vs. SOC) ◆ **Better inter-connection performance** (vs. SIP) ◆ **Low cost** (vs. SIP ; for more die matching) ◆ **SPICE model consistency**	◆ Need customer's design efforts for stacking chips.

Features		
CMOS logic	Process	16nm FinFET
	Structure	ARM-A53
	Die THK	2.85um
	Power	0.8V/1.8V
e-Flash Memory	Process	40nm
	Structure	ESF3 split-gate
	Die THK	780um
	Power	0.81V ~1.21V 1.62V ~ 1.98V
	Access time	28ns (max)
	PGM/ERS time	16us/20ms (max)

HB for DRAM Stacking [Cube]..... (1/2)



- Many channel architecture
- Individual channel vertical interconnects
- Wide noise margin using decupling interposer usage
- High power integrity supply
- Bandwidth 1TB/s, 4TB/s, 8TB/s
- Full memory capacity 32GB, 64GB, 128GB
- Energy efficiency <10mJ/Tb



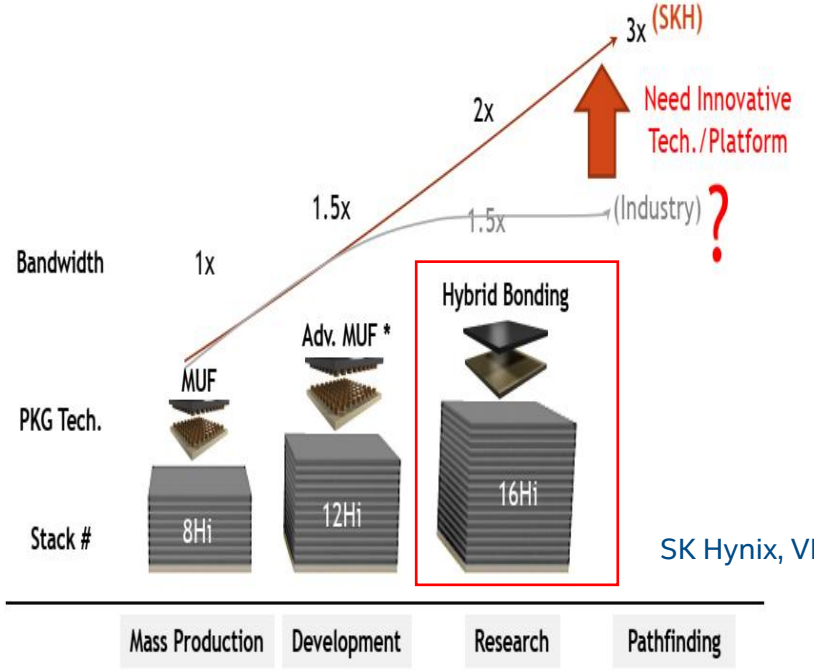
SK Hynix, Samsung

Technique	Requirements	Microbump/TSV pitch size	Die layer thickness	Processing thermal budget	Heat extraction capacity
Microbump (solder)	Microbump + TSV landing pad + Underfill	55µm	50µm (plus 30µm thick underfill)	~250^{0}C for a few (2-3) minutes	Poor due to underfill
Hybrid	Direct electrical connection	2.5µm for wafer-2-wafer	5µm-20µm	~400^{0}C for an hour	Very good
Direct Oxide	TSV after bonding (TSV last)	15-20µm	5-20µm	~150^{0}C for an hour	Very good

PKG Technology Roadmap for Future HBM

PKG technology is a key for the next-generation HBM product

Direction to higher stack w/ thinner chip, narrower gap height, fine pitch Int'n



SK Hynix, VLSI 2021

HBM roadmap includes bumpless stacking with Hybrid Bonding

HB for DRAM Stacking [on Logic]..... (2/2)

WoW Alliance + Micron

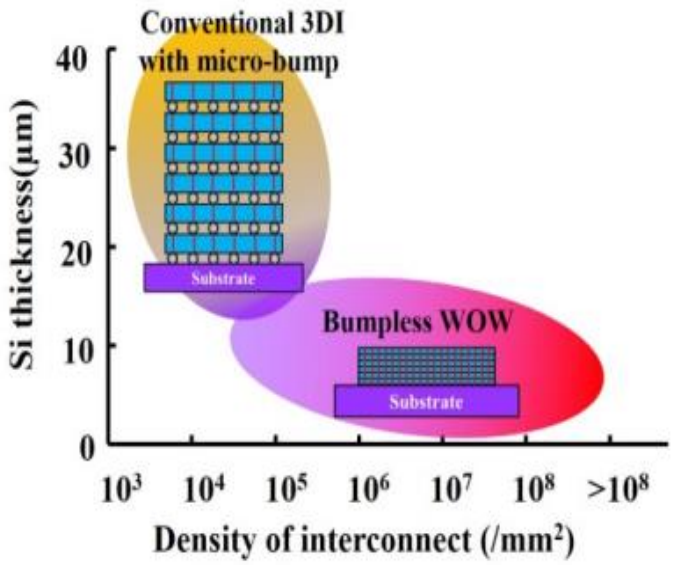
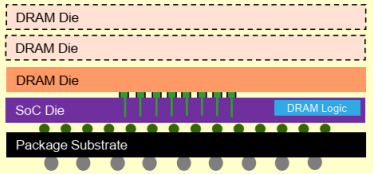


Figure 1. Difference of Si thickness and interconnect density between Bumpless WOW and conventional 3D integration with bumps.

Winbond 3DIC – Proposal



- General spec
  - ✓ 0.5~1GB per die x 1~4 layers
  - ✓ **256GB/s (or more) bandwidth with 4096 I/O**
  - ✓ **Die-to-wafer Hybrid bonding by UMC**
  - or
  - ✓ **1TB/s (or more) bandwidth with 16K I/O**
  - ✓ **Wafer-on-wafer 3DFabric by TSMC**
- Advantages
  - ✓ Cost: No need for interposer and Logic Die
  - ✓ Lower power consumption
  - ✓ Lower latency

Xian, Powerchip, IEDM 2020

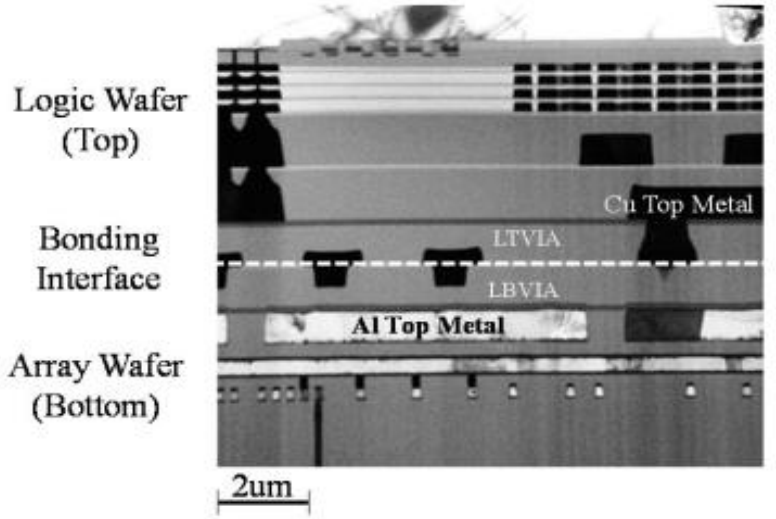
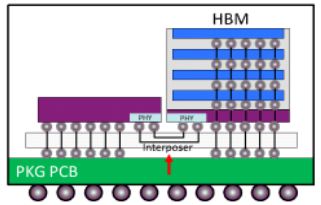
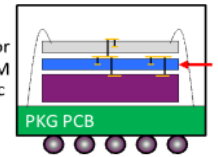
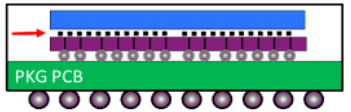


Fig.3. SEDRAM cross sectional TEM images.

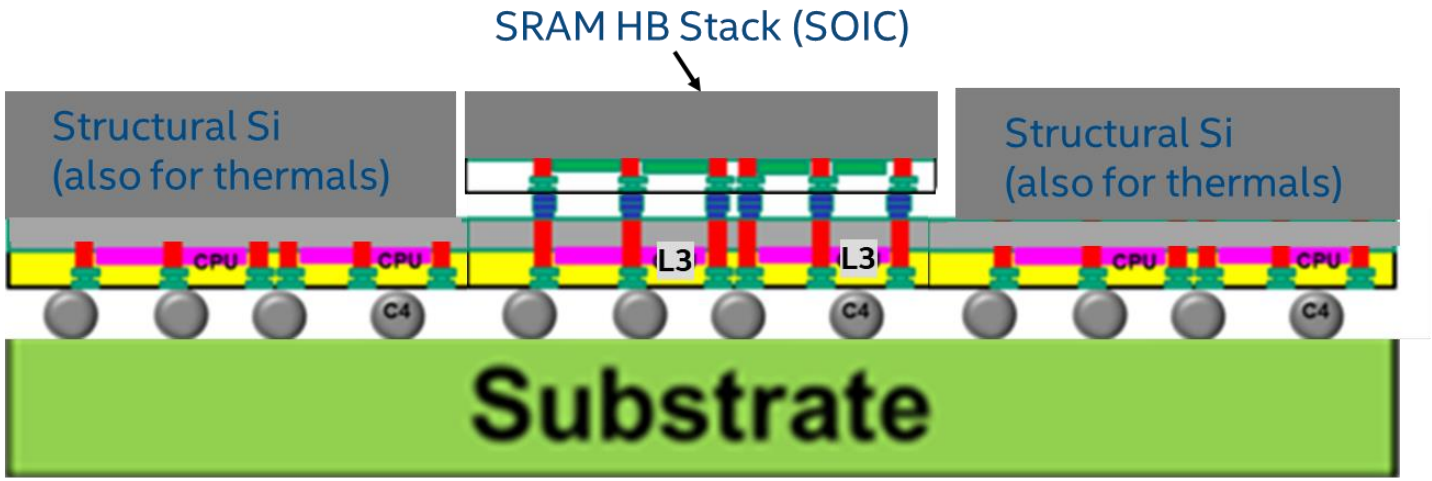
	ISSCC14[3]	ISSCC20[4]	ISSCC17[5], IEDM17[6]	this proposal
Structure diagram				
Connection,pitch(um)	**microbump, 48 x 55, interposer,**		**TSV,6.3 x 6.3, no interposer**	**Hybrid Bonding, 3 x 3, no interposer**
Connection length	~5mm, microbump+wiring		~10um, TSV+wiring	2um, via thickness
PHY needed	Yes		No	**No**
Energy efficiency(pJ/b)	~1.5[2]		N/A	**0.88**
Total Density	8Gb	128Gb	1Gb	4Gb
# of Stack dies	4	8	1	1
Density per die	2Gb	16Gb	1Gb	4Gb
# of Channel	8	8x2	4	**32**
Data bus width	1024(128/ch)	1024(64/ch)	512(128/ch)	**4096(128/ch)**
Data rate per pin(Mbps/pin)	1000	4000	200	**266**
Bandwidth(GBps)	128	512	12.8	136
Bandwidth per die(GBps)	32	64	12.8	**136**

- Scaling Si thickness and increasing 3DID
- HB of 3D stacked DRAM (customized with high I/O) on Logic for optimal bandwidth & power

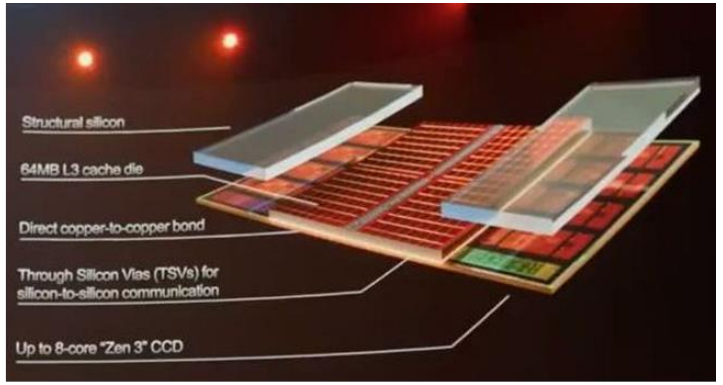
HB for SRAM/Cache on Logic: AMD/TSMC 1/3

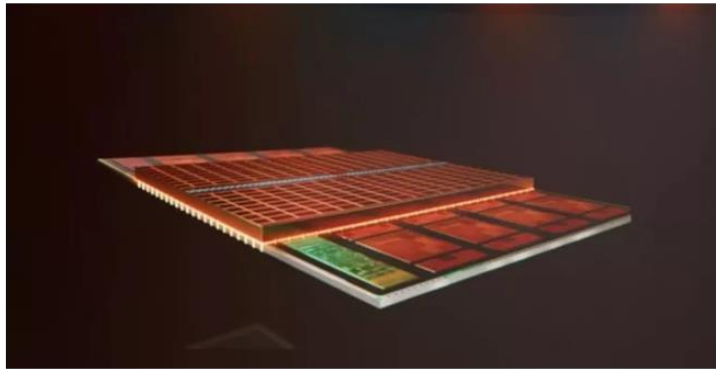
AMD: 3D Chiplet (V-Cache)

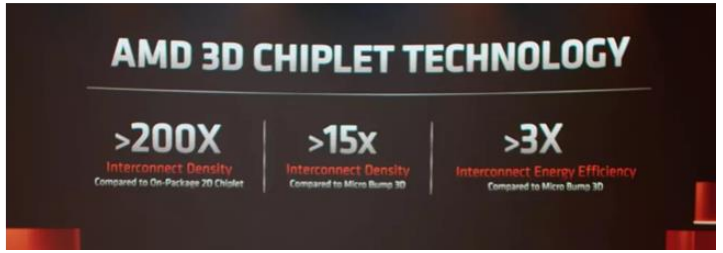
1st Implementation of X3D?



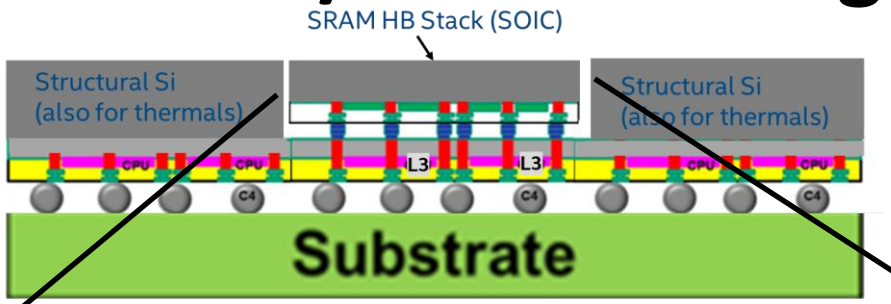
Stacked CCD + V-Cache



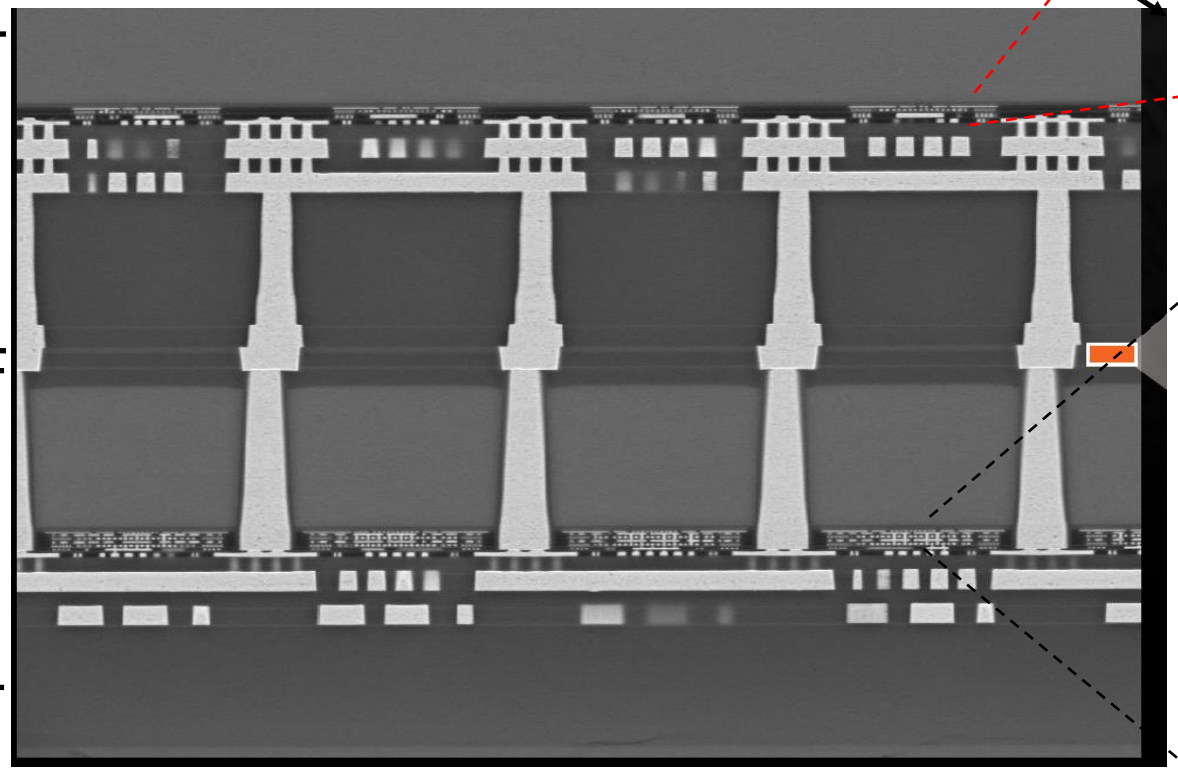




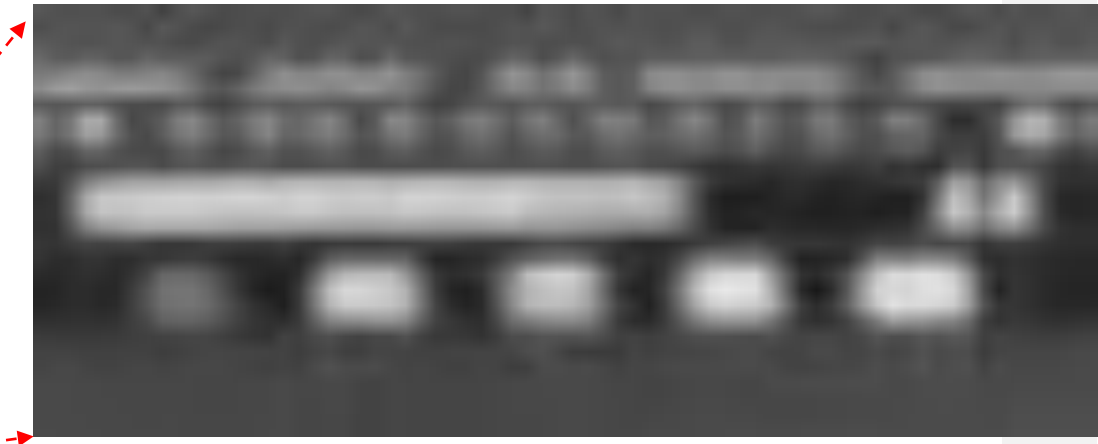
HB for SRAM/Cache on Logic: AMD/TSMC 2/3



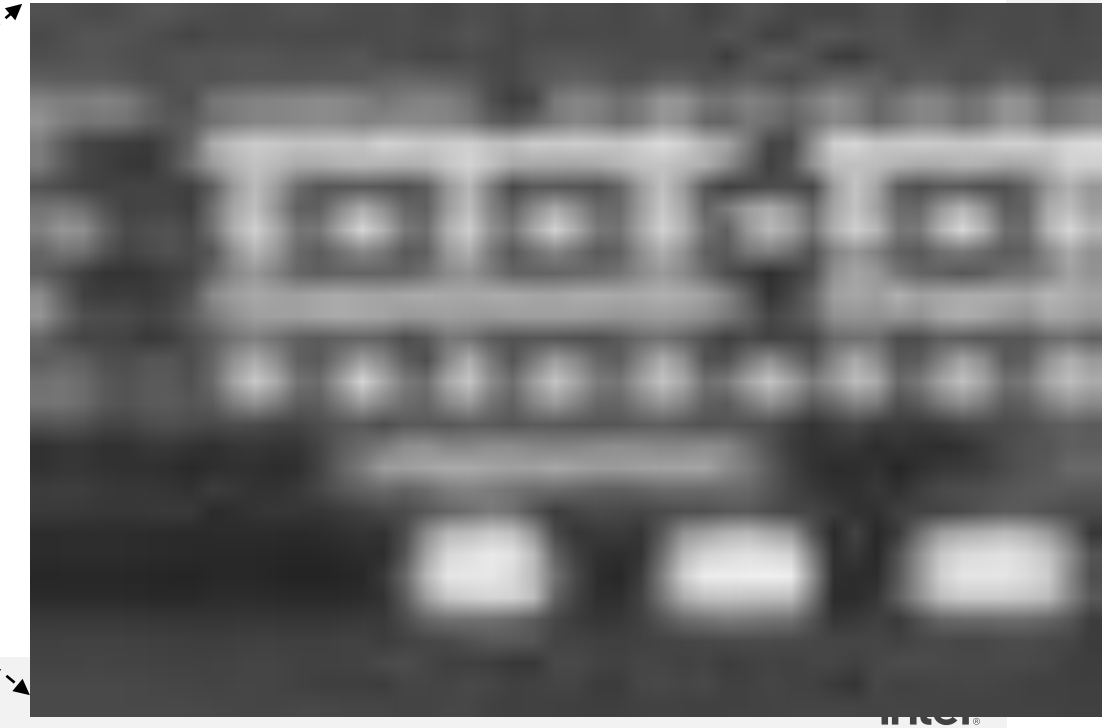
Stacked CCD + V-Cache



Top: SRAM



Bottom: CCD XPU



AMD @ HotChips '21

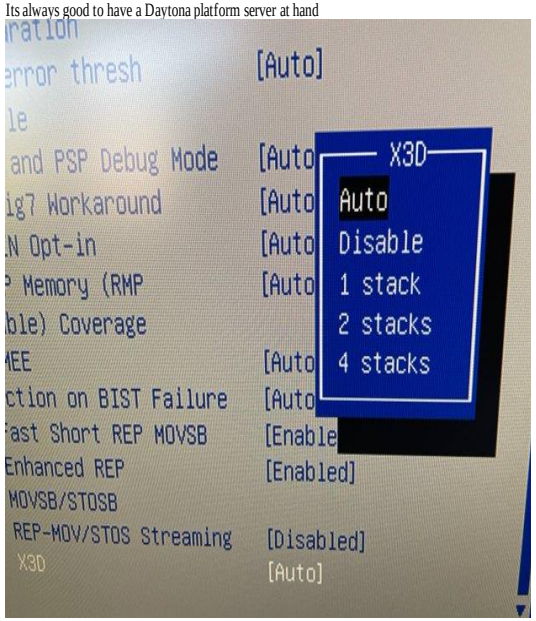
HB for SRAM/Cache on Logic: AMD/TSMC 2/3

AMD V-Cache: Future?

Epyc BIOS with #N Stack

https://twitter.com/aschilling/status/1399701274489151489

Its always good to have a Daytona platform server at hand

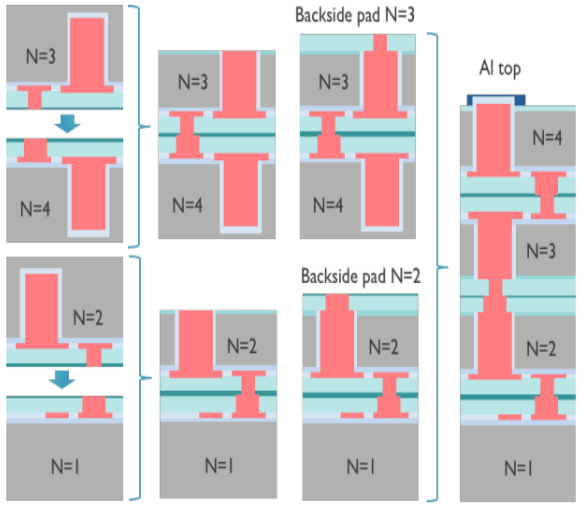


$N=2^k$ Multi-wafer stacking

"Binary tree" STRATEGY

Scalable $N>2^k$ approach :

- Face-to-face bonding N=2
- Back-to-back stack bonding N ≥ 4



N=2 F2F Bonding + N=4 B2B bonding

Pitch scaling: ⇒ **2 µm**

Repeat hybrid B2B bonding for N=8, N=16

imec partner technical week

55

Next-Generation Design and Technology Co-optimization (DTCO) of System on Integrated Chip (SoIC) for Mobile and HPC Applications

Y.-K. Cheng, F. Lee, M.-F. Chen, J. Yuan, T.-C. Huang, K.-J. Chen, C.-T. Wang, C.-L. Chen, C.-H. Tsai, and Douglas Yu
R&D, Taiwan Semiconductor Manufacturing Company, Ltd., HsinChu, Taiwan, email: yk_cheng@tsmc.com

***Abstract*—This paper demonstrates the next-generation design and technology co-optimization (DTCO) of system on integrated chip (SoIC) for mobile and HPC applications, where the SoIC technology was proposed to integrate multi-chips with different functionality and technology into a single SoC chip. The new DTCO includes overall die partitioning, die integration, and interconnect. These methodologies can be used for improving time-to-market and trade-off between performance and cost. In this paper, two prototypes of stacking CPU and memory dies are demonstrated with 15% performance gain and 30% average point-to-point distance reduction.**

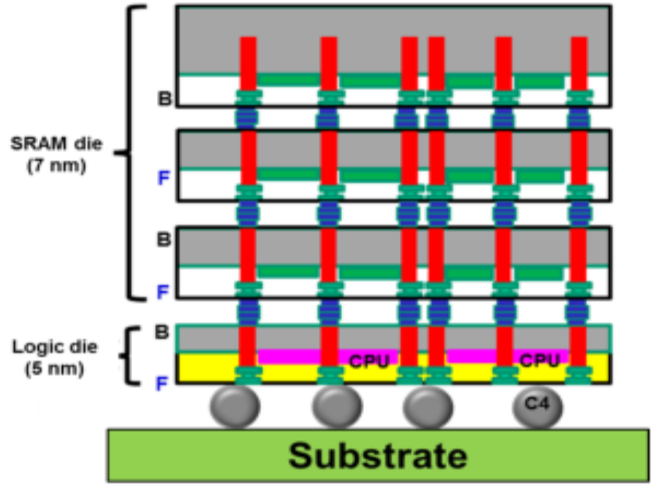


Fig. 9. SoIC-PTV2 stacking view

TSMC IEDM '20

3D-optimized SRAM Macro Design and Application to Memory-on-Logic 3D-IC at Advanced Nodes

R. Chen[1], P. Weckx[1], S. M. Salahuddin[1], S.-W. Kim[1], G. Sisto[1,2], G. Van der Plas[1], M. Stucchi[1], R. Baert[1], P. Debacker[1], M.H. Na[1], J. Ryckaert[1], D. Milojevic[1,3], E. Beyne[1]

[1]IMEC, Leuven, Belgium, email: Rongmei.Chen@imec.be, [2]Cadence, CA, USA, [3]Université libre de Bruxelles, Belgium

Abstract – We present local & global SRAM macro optimizations for 3nm FinFET and 2nm Nanosheet using Face-to-Face (F2F) and Wafer-to-Wafer (W2W) hybrid bonding at sub 1um pitch. Bonding pad parasitics are measured experimentally to calibrate RC models of the pad used to evaluate 3D-optimized memory macro delays. 3D-optimized macros are designed to reduce the macro external delay by ~50%. With customized SRAM BEOL, performance improvement of up to 70% for larger memories is observed compared with 2D macro. We also show that bit-cell tech-level optimizations have minor impact on the performance of large caches at advanced nodes due to high metal resistance in the macro global routing. Finally, at system-level we partition a L2 data memory (with 3D-optimized macro) from logic showing that the 3D implementation achieves a total of 33% performance gain with respect to a 2D implementation.

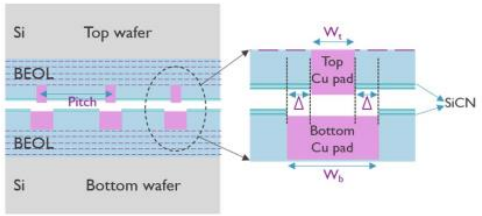


Fig. 1. Schematic of F2F&W2W hybrid Cu/SiCN-to-Cu/SiCN bonding with definitions of different dimensions: top pad width W_t, bottom pad width W_b, overlay tolerance Δ and pitch.

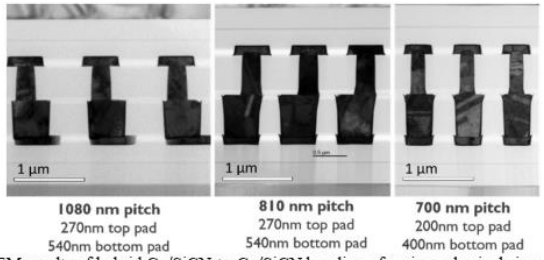


Fig. 2. TEM results of hybrid Cu/SiCN-to-Cu/SiCN bonding of various physical sizes.

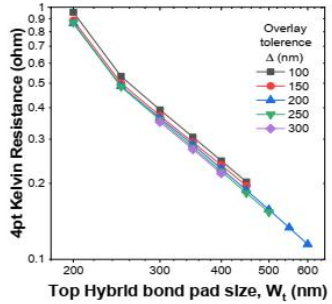


Fig. 3. 4pt Kelvin measured results of hybrid bonding resistance as a function of top pad size considering the impact of overlay tolerance. The standard deviation of the measurement is below 5%, except for the 200 nm point (10% instead).

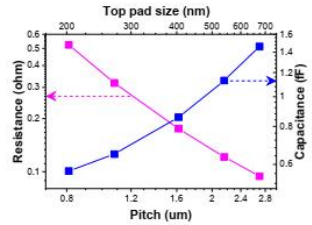


Fig. 4. Modelling of hybrid bonding RC as a function of bonding pitch/top pad size. Various coupling capacitances including the top & bottom bonding plane and neighboring bonding pads are considered.

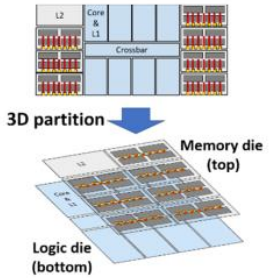


Fig. 5. Multi-core SoC and Memory-on-Logic partitioning of scheme b): partition SRAM macros from logic die.

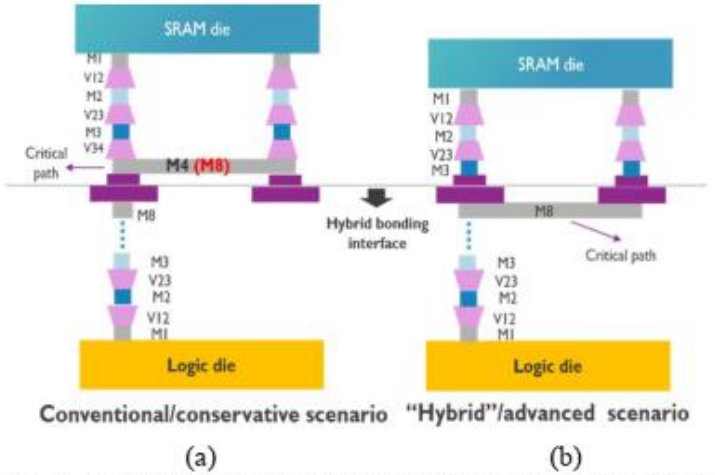


Fig. 8. Customization of BEOL for SRAM die. (a) configure SRAM die M4 layer (critical path routing) with M8 process, i.e. change in metal thickness, aspect ratio, etc. leading to lower resistance without increasing capacitance; (b) use logic die M8 for memory die M4 routing via the hybrid bonding pad without increasing BEOL routing conjestion in logic die.

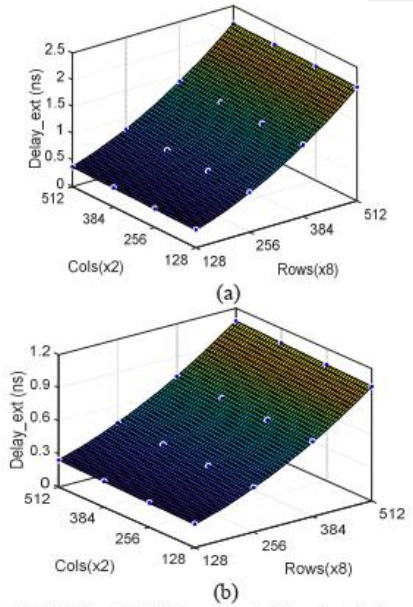


Fig. 14. 2D (a) & 3D (b) external delay for 2x8 array macro (3nm, non-BPR). The external delay is sensitive to the number of rows due to IO and address routing (M4) in the BL direction (refer to Fig. 7).

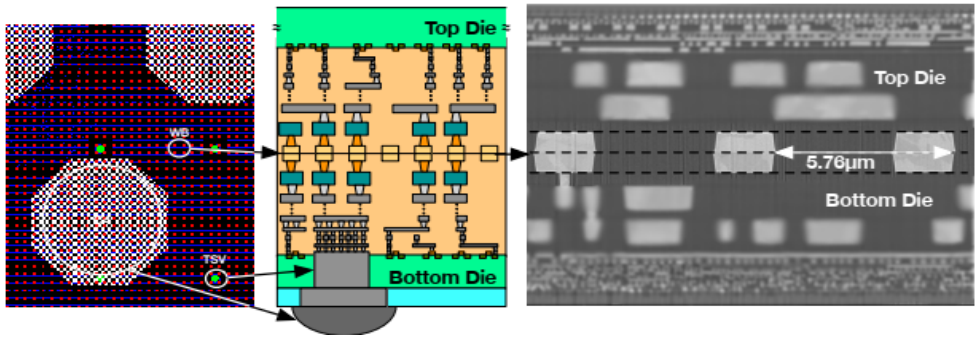
HB for Logic on Logic:

ARM, GF IEDM 2020

A high-density logic-on-logic 3DIC design using face-to-face hybrid wafer-bonding on 12nm FinFET process

S. Sinha, S. Hung, D. Fisher†, X. Xu, C. Chao, P. Chandupatla, F. Frederick, H. Perry, D. Smith†, A. Cestero‡, J. Safran‡, V. Ayyavu, M. Bhargava, R. Mathur, D. Prasad, R. Katz‡, A. Kinsbruner‡, J. Garant‡, J. Lubguban‡, S. Knickerbocker‡, V. Soler‡, B. Cline, R. Christy, T. McLaurin, N. Robson‡, D. Berger‡
Arm Inc., 5707 Southwest Parkway, Austin, TX, 78735
†GLOBALFOUNDRIES, Malta, NY 12020 USA. ‡GLOBALFOUNDRIES, Hopewell Junction, NY 12533, USA.
Email: *saurabh.sinha@arm.com / daniel.fisher@globalfoundries.com*

***Abstract*—A high-density-3D test-vehicle showcasing a synchronous cache coherent mesh interconnect design (Arm Neoverse® CMN-600) operational at frequencies up to 2.4 GHz and partitioned in 3D using 5.76µm pitch face-to-face wafer-bond 3D connections on a 12nm FinFET process is presented. The test-vehicle is designed using an industry tool compatible innovative physical implementation flow and serves as the first known industry demonstration of the IEEE 1838 3DIC Design-for-Test (DFT) standard. We demonstrate a 3D aggregate bandwidth of 307 GB/s, a record bandwidth density of 3.4 TB/s/mm^2, and an energy efficiency of 0.02 pJ/bit for the 3D-stacked dies. We present measurement and analysis data from 945 dies where a total of 13.5 million signal 3D wafer-bond nets and 20 million power-delivery 3D wafer-bond nets on multiple wafer-bonded pairs are tested showing robust functionality, paving the path for 3D-stacked high performance logic-on-logic applications.**



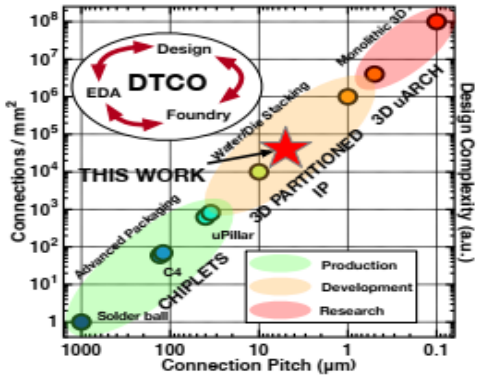


Fig. 1 The 3D integration roadmap. This work targets 5.76µm 3D face-to-face bonding pitch. Strong Design-Foundry-EDA collaboration is important for high-density 3D technologies.

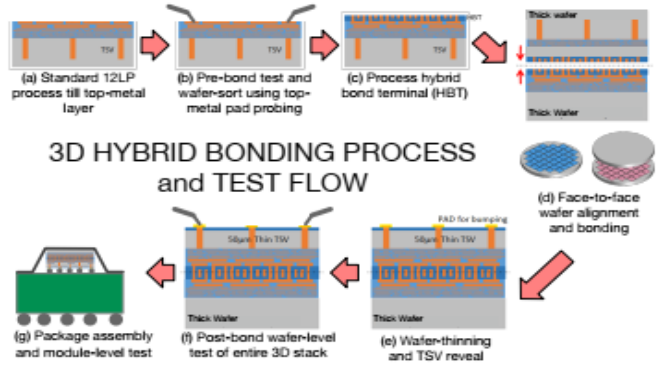


Fig. 2 3D process and test flow diagram showing (a-b) pre-bond tests using top metal test pads to enable wafer-sorting and matching, (c-d) hybrid wafer bonding at hybrid bonding terminal (HBT) layer and (e) TSV reveal with contact metal, (f-g) post-bond test through C4 bumps and TSV and packaging.

Metric	Value
Process technology	12nm FinFET
Metal layers per die	11
3D stacking	Face-to-face hybrid wafer bond
3D pitch	5.76µm
TSV diameter	5µm
C4 bump pitch	150µm
Active die area	1.18mm²
3D signals for CMN-600	1600 per XP
3D signals/die	13800
3D pads for power delivery/die	22158
Cumulative 3D signal-nets tested from 945 wafer-bonded dies	**13.5 million**

Table I Key metrics of the 3D stacked test-vehicle. The vehicle demonstrates the feasibility of hybrid-wafer bonding for logic-over-logic high-density 3D design.

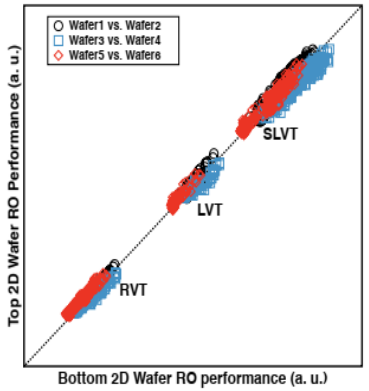


Fig. 8 Example pre-bond measurements from 6 wafers to match for bonding. Good correlation observed between selected wafers for bonding.

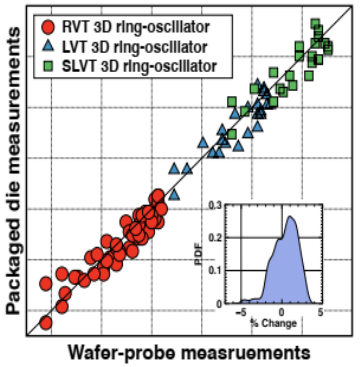


Fig. 9 Post-dicing and packaged device performance versus post-bond wafer-probe measurements.

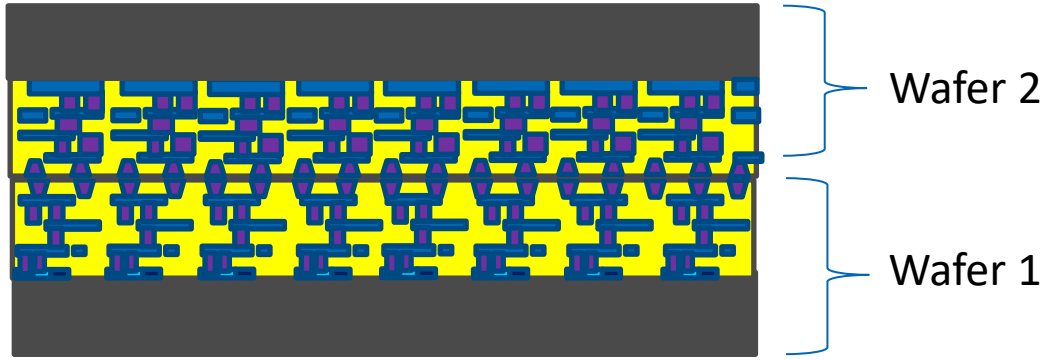
HB for Process Flow Split and Bond [Fab Cycle Time]:

Step 1: Manufacture wafer 1 and wafer 2 independently.

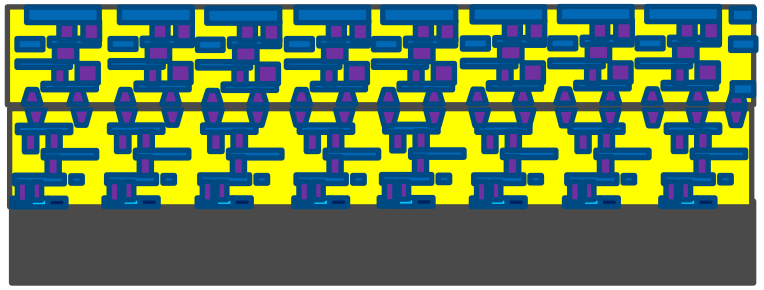
Wafer 1: FE + Upto Mx

Wafer 2: EOL – Mx+1

Step 2: Bond wafer 1 and wafer 2.



Step 3: Back grind/CMP/Dry etch wafer 2 to expose EOL pads/structures.



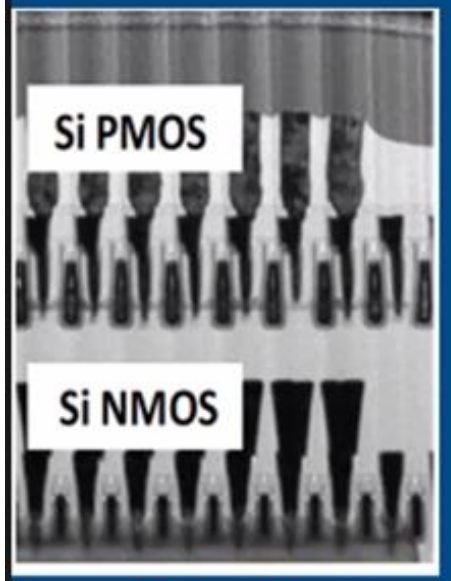
Advantages of Proposes Solution:

- Can use High temp materials for BE of the process, as wafer 2 processing does not impact wafer 1
- Manufacturability advantage
- Significant Cycle time advantage, Wafer 1 and 2 can be processed in parallel.
- Yield benefit
- Shorter development time
- Wafer 2 Does not need to be High Quality Si

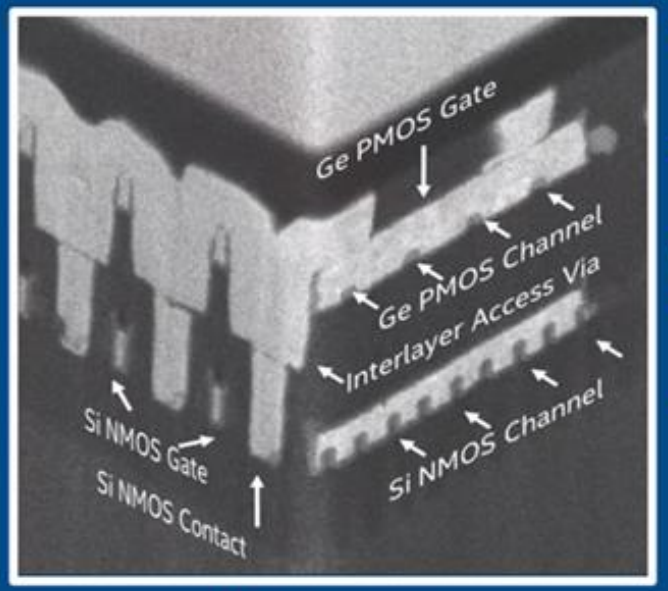
Anup Pancholi, Prashant Majhi, TMG/GSM, Intel Patent Filed 2018

Monolithic 3D Integration

Density Scaling, Performance & New Applications



Density
Cell Height Scaling



Performance
Ge PMOS Performance

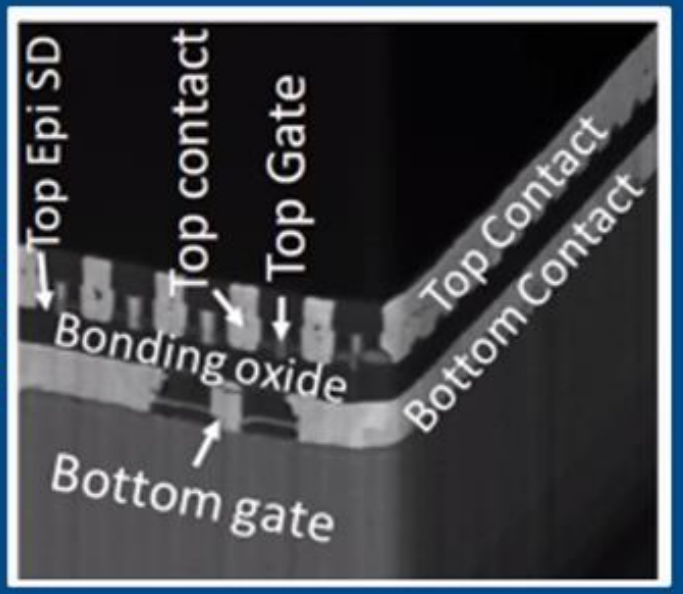
Top
Ge PMOS

Bonding Oxide

Bottom
Si NMOS

Top
Si PMOS

Bottom
GaN NMOS

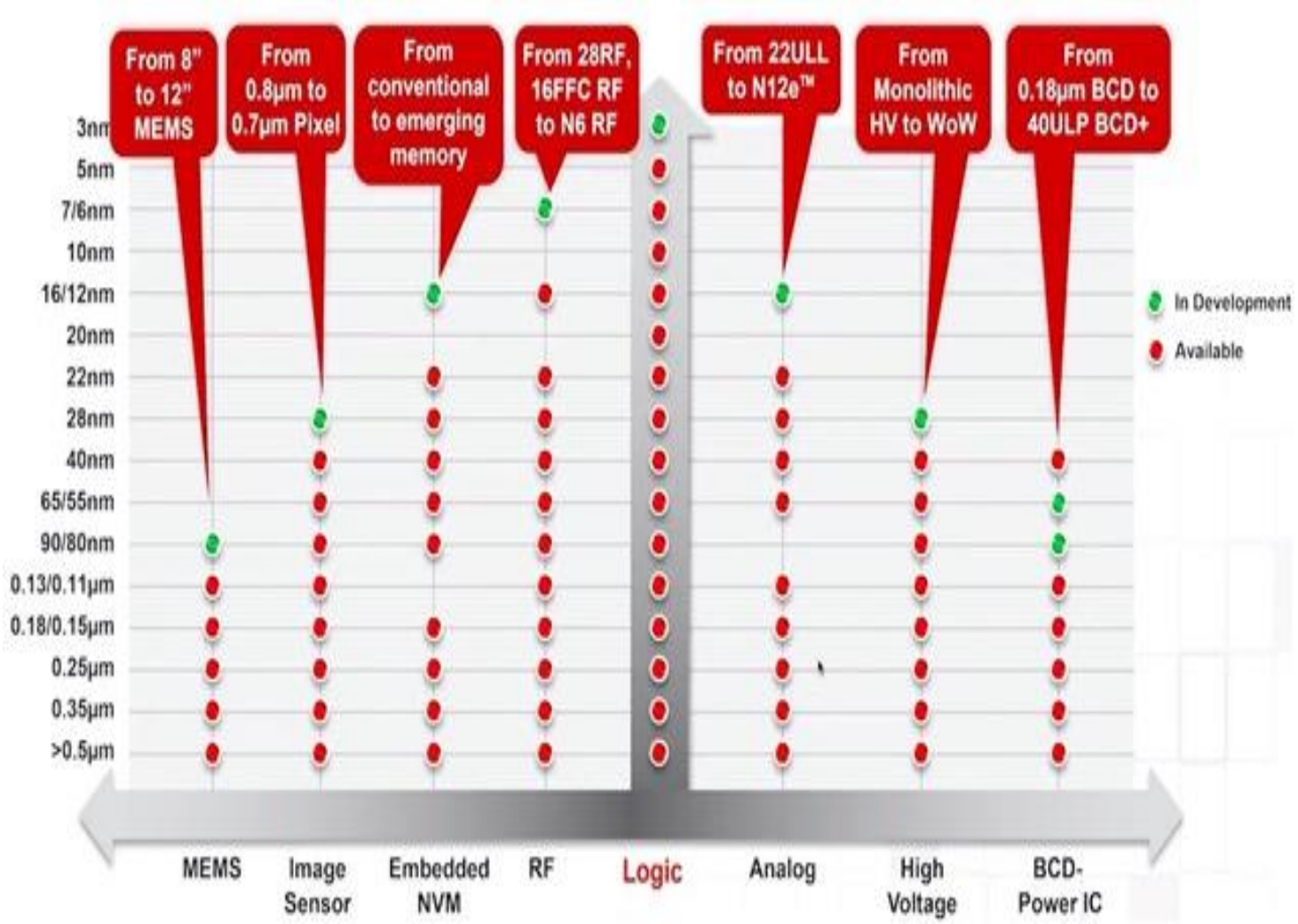


New Applications
Monolithic GaN NMOS + Si PMOS
Single Chip Fully Integrated
5G RF FE & Power Delivery

Other Bonding Applications

- TSMC: HV and LV Logic for Display Driver and other IOT ⇨
- TSMC/Samsung: DTC or ISC for tightly coupled high Density MIM Cap ⇨
- TSMC: Bonding of Corrugated Si for 3DIC immersion Cooling ⇨
- TSMC: COUPE for Tightly Coupled Si-Photonics [PE to XPU] ⇨
- GF/Samsung HB for (III-V) Laser on (Si SOI) PIC ⇨
- TSMC/Industry/Academia: Tightly coupled NVM to Logic [inc MRAM] ⇨
- TSMC Immersion In Memory Compute ⇨
- …

WoW for High Voltage



NVM	HV	Sensor	BCD
ESF3 MRAM RRAM	LCD Driver OLED Driver	CIS MEMS	Low Ron LDMOS Integrated Passives

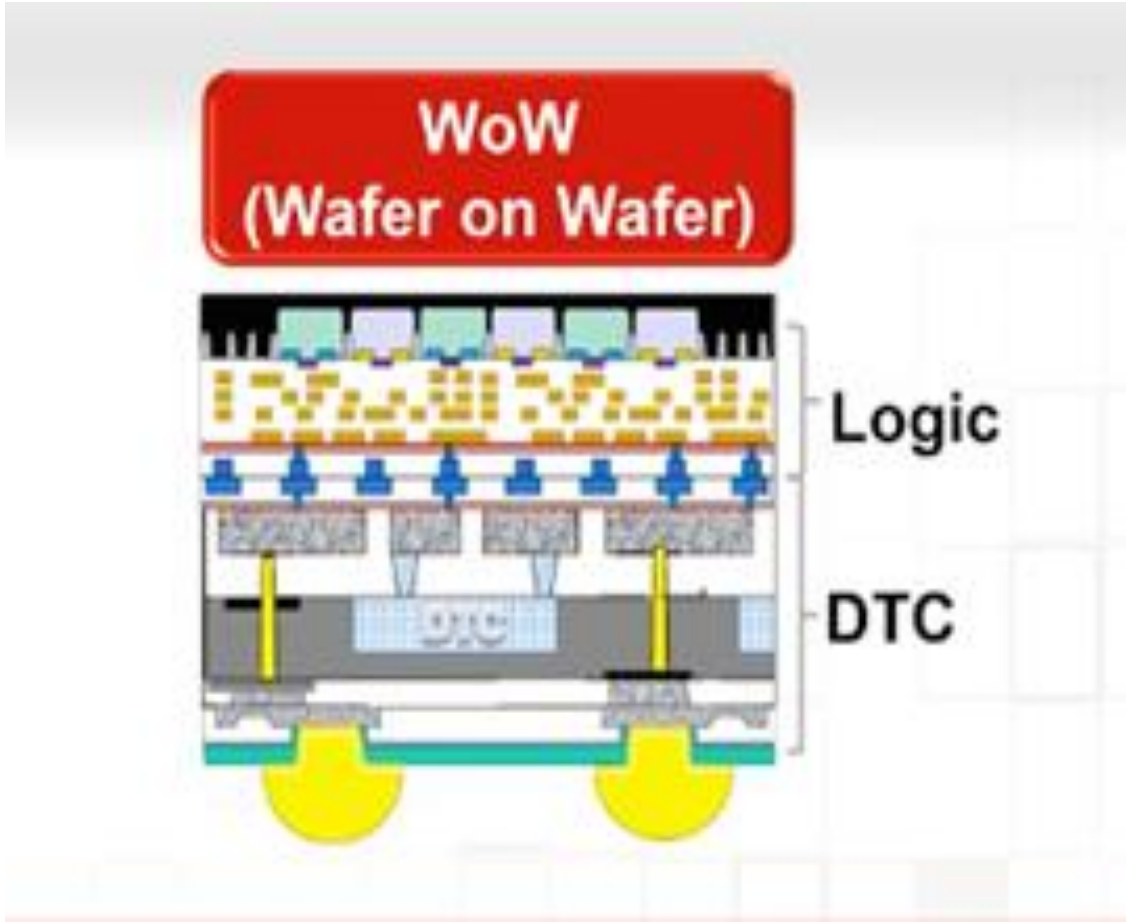
ULP/ULL	Analog	RF
eHVT eLVT ULL SRAM Low Vdd	6nm/16nm/22nm/40nm/55nm RF Active/Passive Devices RF Model/PDK Analog General Offers (LN, TaN, etc) Customization	

Logic Technology
5nm/7nm/16nm/12nm/22nm/28nm/40nm/55nm... Logic Technology Process Base Bank Model/PDK

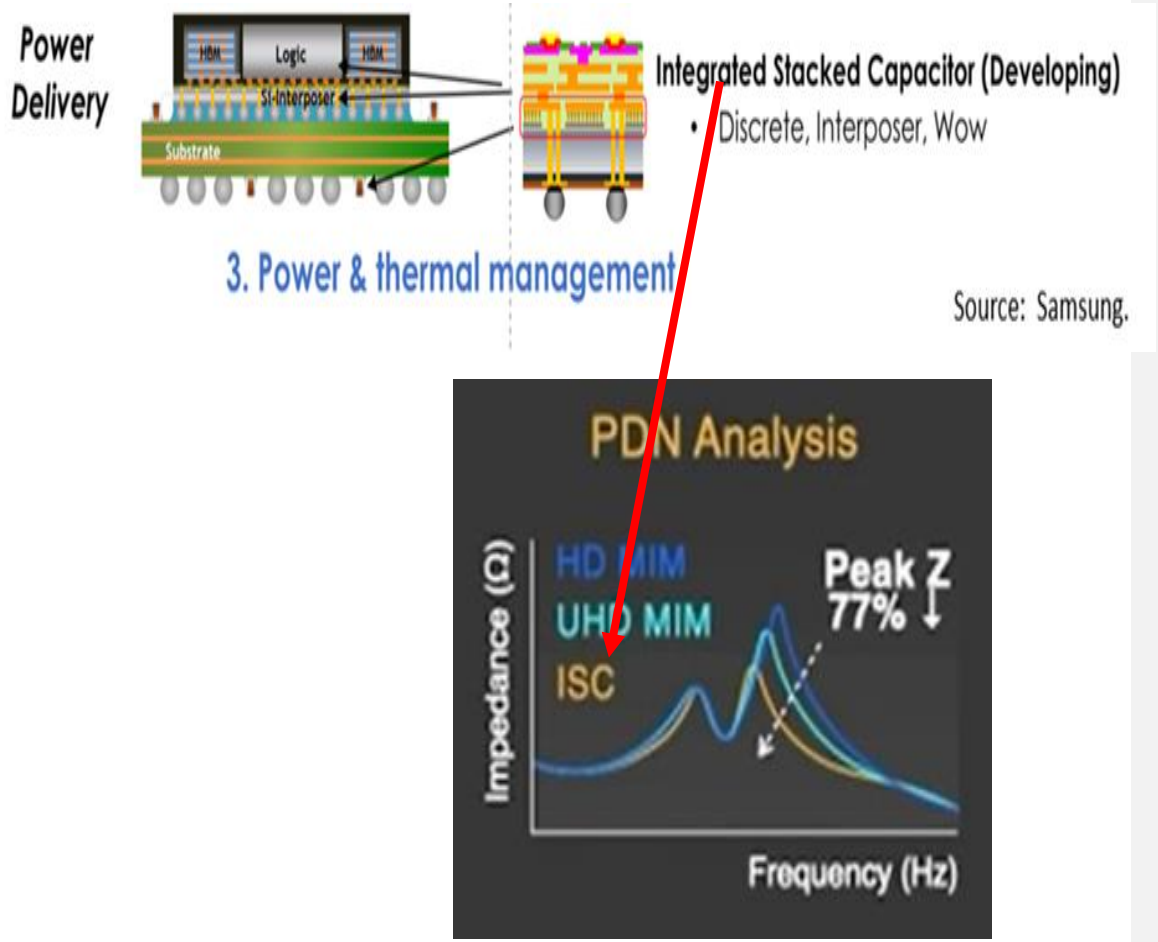
https://community.cadence.com/cadence_blogs_8/b/breakfast-bytes/posts/tsmc-2020-special

WoW Bonding for High Density MIM:

Deep Trench Cap (TSMC), Integrated Stacked Capacitor (Samsung)



https://www.eetimes.com/tsmcs-chip-scaling-efforts-reach-crossroads-at-2nm/#

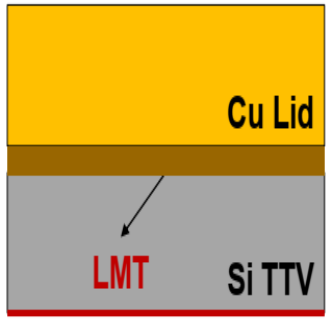


Bonding for 3DIC Cooling

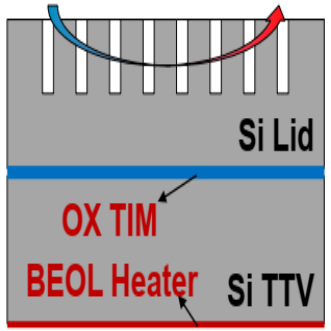
Integrated Si Micro-Cooler (ISMC) for Ultra-HPC

- Thin SiOx bonding interface (OX TIM) by fusion bonding Si lid and Si chips
- Low interface TR, even though K_{SiOx} at low single digit W/m·K

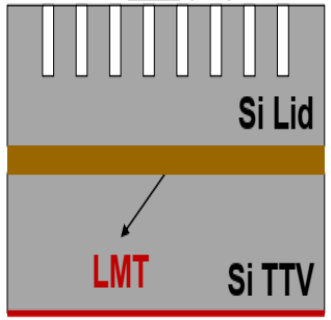
Cu Lid with LMT (Liquid Metal TIM)



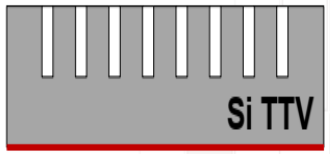
Si Lid with OX TIM



Si Lid with LMT (Liquid Metal TIM)

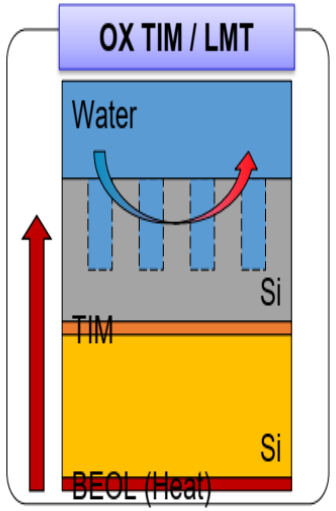


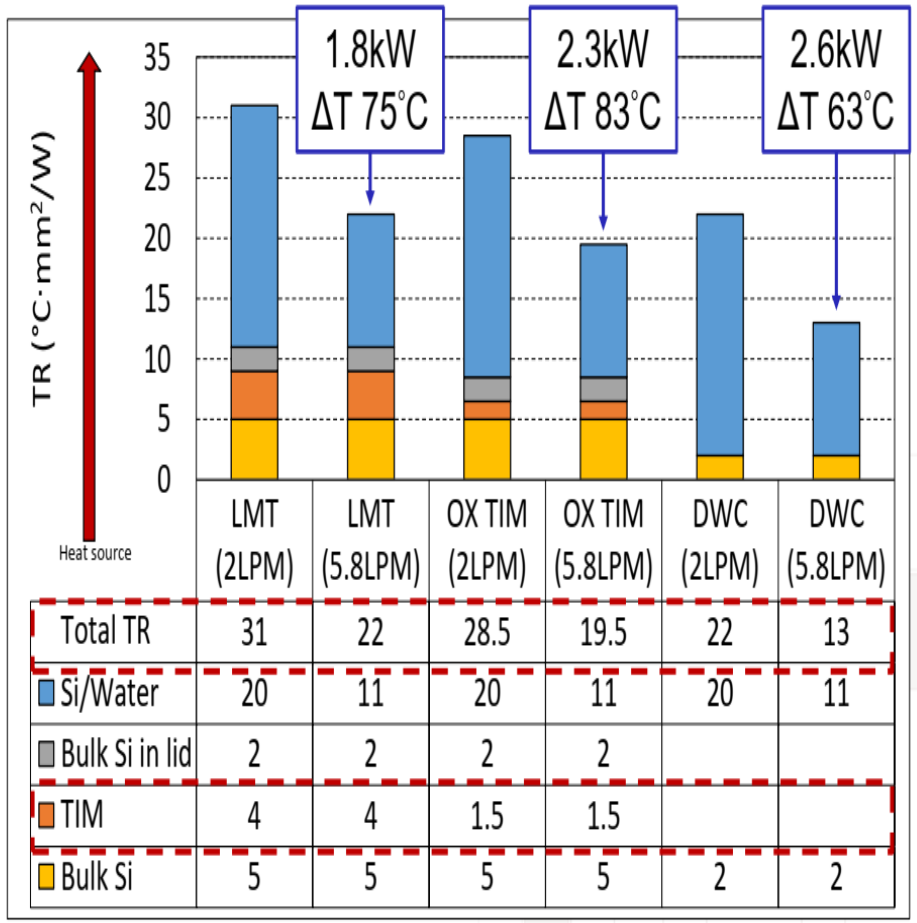
DWC (Direct Water Cooling)



TSMC, ECTC 2021

Cooling Performance Benchmark





	LMT (2LPM)	LMT (5.8LPM)	OX TIM (2LPM)	OX TIM (5.8LPM)	DWC (2LPM)	DWC (5.8LPM)
Total TR	31	22	28.5	19.5	22	13
Si/Water	20	11	20	11	20	11
Bulk Si in lid	2	2	2	2		
TIM	4	4	1.5	1.5		
Bulk Si	5	5	5	5	2	2

TSMC “COUPE” for Co-Packaged Si-Photonics CPO Leadership

COUPE: Compact Universal Photonic Engine

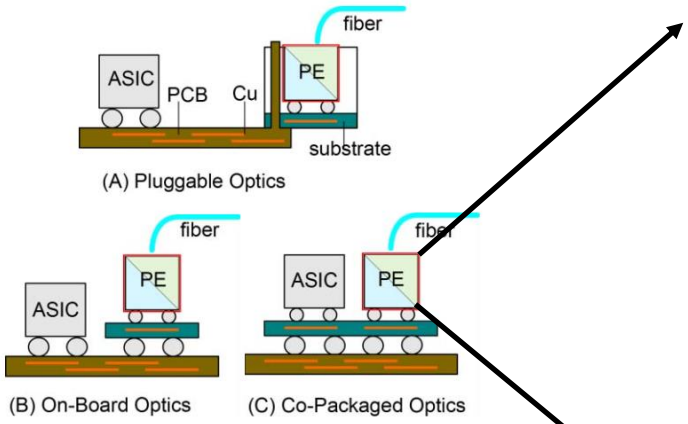
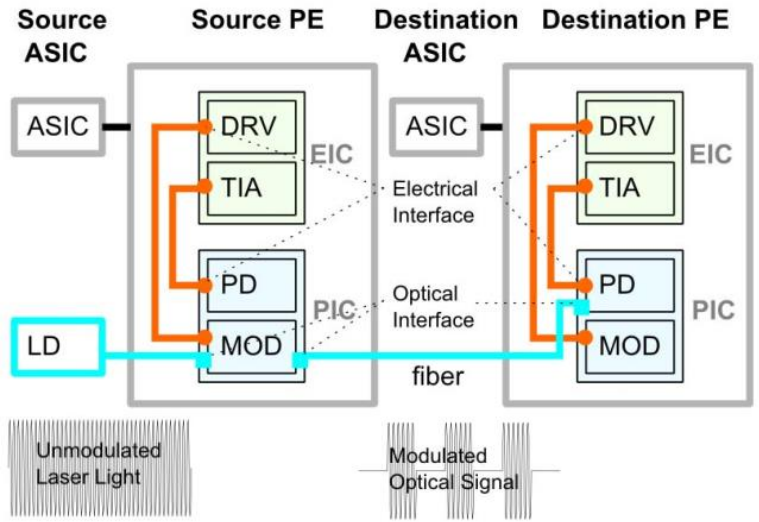
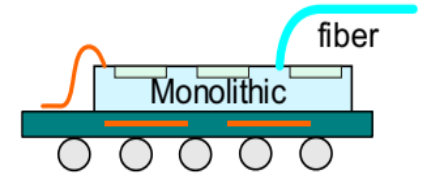


Fig.1. SiPh package configurations on motherboard: evolved from (A) Pluggable Optics, (B) On-Board Optics (OBO), to (C) Co-Packaged Optics (CPO).

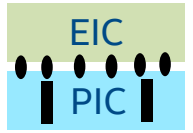
PE Logical Construct

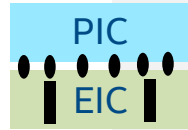


PE Physical Construct

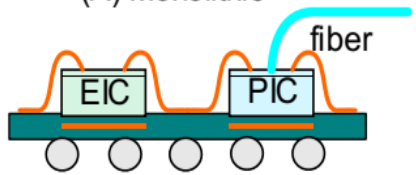


(A) Monolithic





(C) 3D [EIC/PIC w/ ubump, TSV]



(B) Heterogeneous Integration

(D) COUPE
(Not disclosed)

COUPE, being a heterogeneous integration technology by nature, is designed in to minimize electrical coupling loss as well as to avoid the reoccurring KOZ engineering and the TSV loss. In our link analysis that compares performances between conventional heterogeneously integrated PE and COUPE, up to roughly 40% savings on both driving current and energy consumption can be observed for COUPE when 112Gbps none-return-zero (NRZ) modulation is applied to micro-ring modulators (MRM) (See Fig.4 and Table I).

TSMC, ECTC 2021

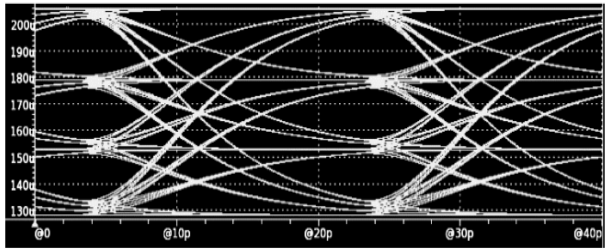


Fig.4. In the transmitter demonstrator using 7nm FinFET and 112Gbps/ PAM4 modulation format on MRM, COUPE can yield satisfying eye openings. For the transmitter using conventional heterogeneous integration technology, almost 1.7X increase in driving current is needed to reach the same eye openings.

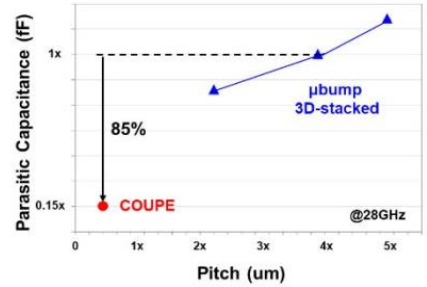


Fig.6. EIC-to-PIC interface parasitic comparison.

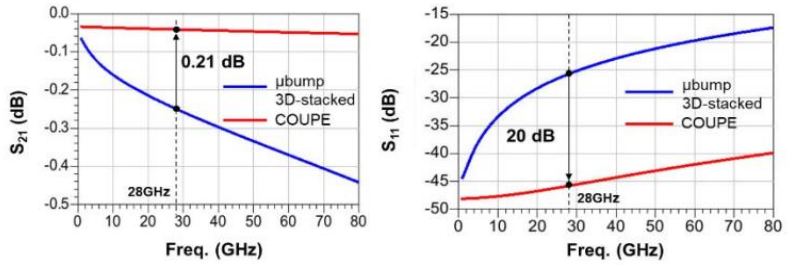
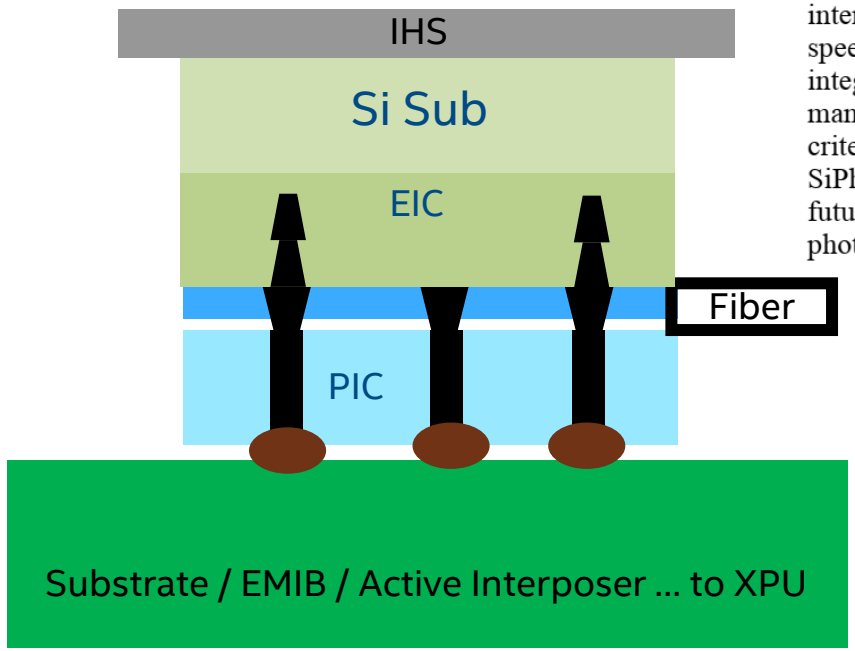


Fig.7. Interface insertion loss and reflection loss comparison.

COUPE: Speculation of Config ?

- TSV in PIC
- Hybrid Bonding Connections?
  - Pitch Scaled
  - Parasitics Scaled
- Thermals for PIC?



V. CONCLUSION

Through the structural survey, we reach the conclusion that a 3D stacking, with TSV in PIC is the best choice for OE, from both electrical and optical interface point of view. COUPE, the heterogeneous integration technology uniquely leverages the bandwidth, bandwidth density, and latency advantages of SiPh interconnect while accommodating both GC and EC to meet speed and power consumption requirements. Being able to be integrated at wafer scale and at FEOL makes it a low cost, manufacturable and uniquely meet the most demanding PPAC criterions. We believe that COUPE- the compact and universal SiPh integration solution can serve as the building block for future wafer level system integration (WLSI) based on silicon photonics.

VDD=0.8V, PAM4 Data Rate=112Gbps	PE by Conventional 3D Stacking	COUPE
Current Consumption (mA)	1X	0.6X
Energy Consumption (pJ/bit)	1X	0.6X

Table I. In our link analysis that compares performances between conventional heterogeneously integrated PE and COUPE, up to 40% savings on both driving current and energy consumption can be observed for COUPE when 112Gbps PAM4 modulation is used.

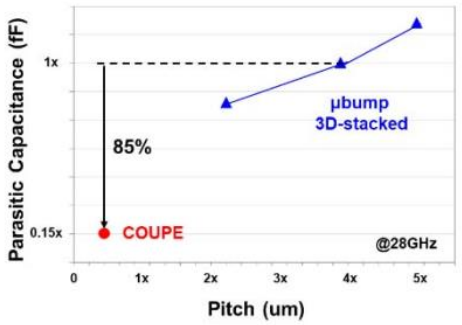


Fig.6. EIC-to-PIC interface parasitic comparison.

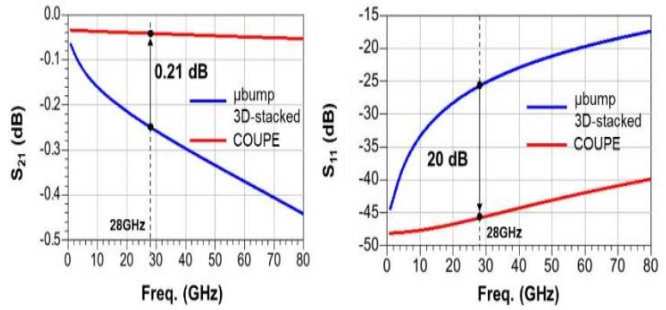


Fig.7. Interface insertion loss and reflection loss comparison.

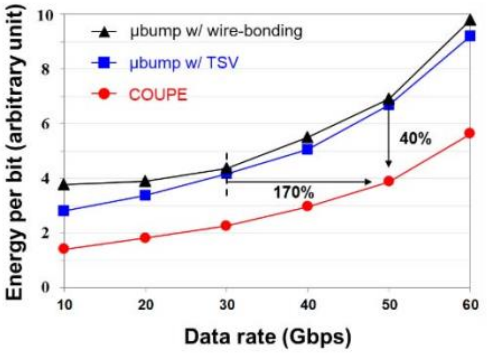
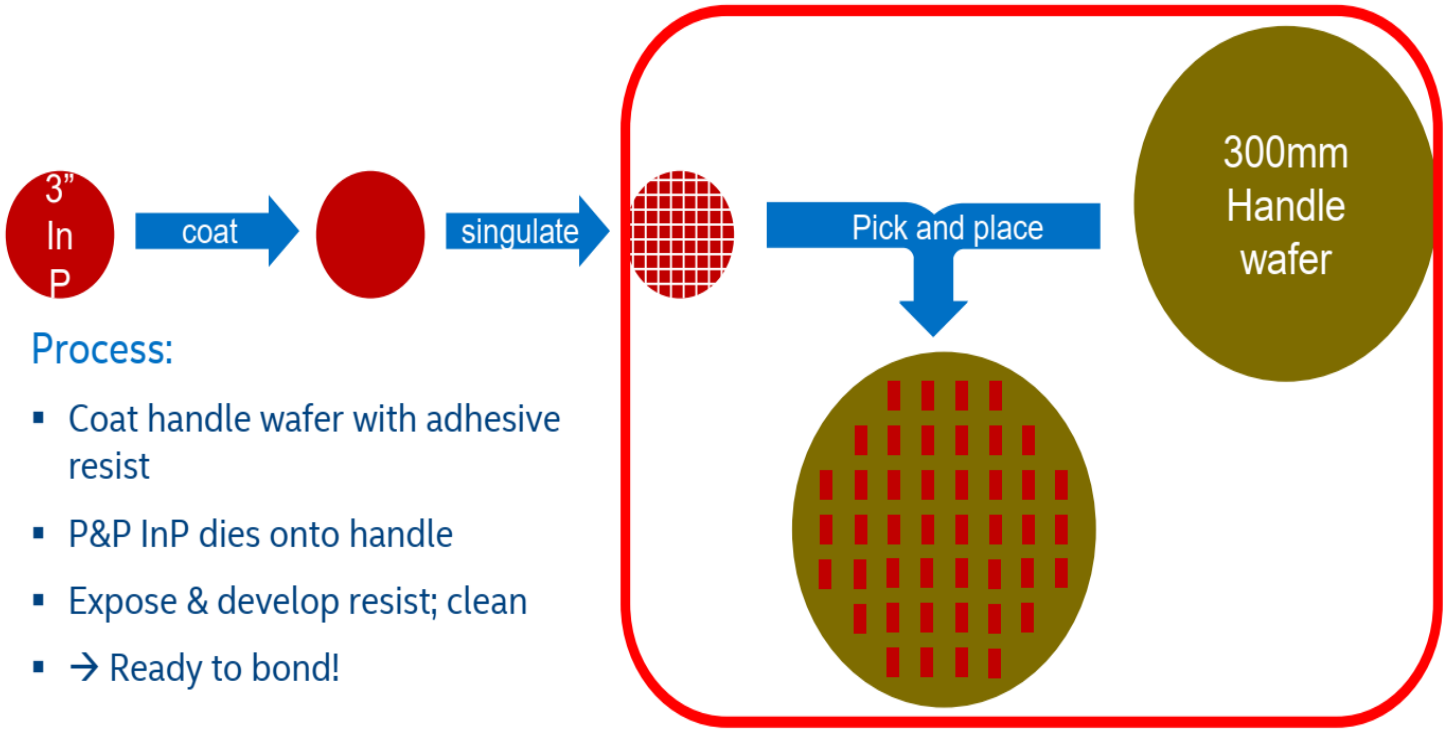


Fig.11. Photonics engine's transmitter power consumption.

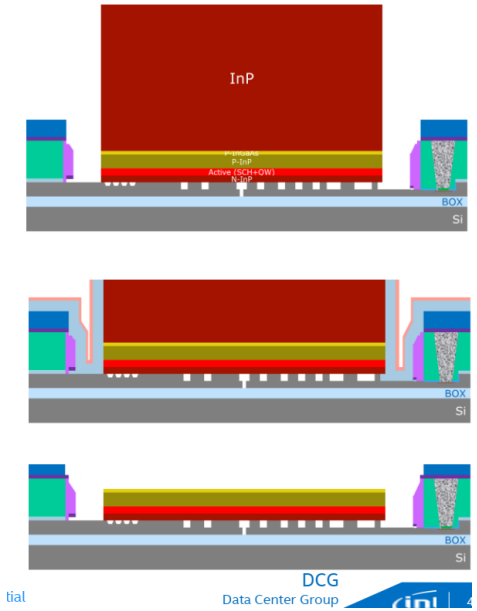
Intel's Differentiation With (III-V) Laser on Si

Pick-and-place to Handle wafer



Process:

- Coat handle wafer with adhesive resist
- P&P InP dies onto handle
- Expose & develop resist; clean
- → Ready to bond!



John Heck DCG-SPPD

Bonding of (III-V) Laser on Si

Global Foundries

3D Integrated Laser Attach Technology on 300-mm Monolithic Silicon Photonics Platform

Yusheng Bian[1,*], Koushik Ramachandran[2], Bo Peng[4], Brittany Hedrick[2], Keith Donegan[1], Jorge Lubguban[2], Benjamin Fasano[2], Armand Rundquist[5], Jim Pape[2], Asli Sahin[2], Thomas Houghton[2], Karen Nummy[2], Jay Steffes[2], Louis Medina[2], Subharup Gupta Roy[3], Harry Cox[2], Bart Green[3], Kevin Dezfulian[2], Won Suk Lee[1], Andy Stricker[1,2], Kate Mclean[3], Shuren Hu[4], Zoey Sowinski[2], Colleen Meagher[2], Abdelsalam Aboketaf[5], Michal Rakowski[1], Mai Randall[1], Ian Melville[2], Dave Riggs[2], Ajey Jacob[4], Rod Augur[1], Daniel Berger[2], Anthony Yu[2], Ken Giewont[2] and John Pellerin[1]

[1]*GLOBALFOUNDRIES, 400 Stone Break Rd Ext., Malta, NY 12020, USA*
[2]*GLOBALFOUNDRIES, 2070 Route 52, Hopewell Junction, NY 12533, USA*
[3]*GLOBALFOUNDRIES, 1000 River St., Essex Junction, VT 05452, USA*
[4]*Formerly with GLOBALFOUNDRIES, USA*
[5]*NeoPhotonics Corporation, 2911 Zanker Road, San Jose, CA 95134, USA*
[*]*yusheng.bian@globalfoundries.com*

***Abstract*—A hybrid laser attach technology was demonstrated on GLOBALFOUNDRIES 300-mm monolithic silicon photonics (SiPh) platform. High accuracy bonding of laser inside a cavity in the SiPh die was accomplished. Optical power up to 10dBm was demonstrated through direct butt-coupling of the laser to SiPh die.**

Keywords—Semiconductor lasers, monolithic silicon photonics, hybrid integration, photonic integrated circuits

I. Introduction

Silicon photonics (SiPh) has been identified as one of the key enabling technologies to overcome microelectronics bottlenecks and address the ever-increasing demands in global data communication [1]. SiPh-based photonic integrated circuits (PICs) offer the promise of low-cost and high-volume solutions for next-generation, high speed energy-efficient optical interconnects [2]. While remarkable advances have been achieved at both the component and system level in SiPh, the on-chip integration of low cost and power efficient laser sources onto a SiPh PIC remains a significant challenge. Among a variety of demonstrated approaches (including monolithic integration (e.g. heteroepitaxy) [3] and heterogeneous integration using direct wafer bonding techniques [4]), hybrid integration technology represents a viable solution towards attachment of high-performance on-chip lasers by leveraging flip-chip bonding processes and butt coupling of the laser to SiPh PICs [5]. However, the large divergence of the III-V laser and significant mode mismatch between the laser core and SiPh waveguide (WG) poses stringent requirements on the chip alignment in all dimensions, as well as the intermediate coupling elements. By leveraging the advanced manufacturing and packaging techniques, we report here for the first time a 3D hybrid integrated laser attach technology based on GLOBALFOUNDRIES (GF) 300 mm monolithic SiPh platform [6].

II. Laser Integration on Monolithic SiPh Platform

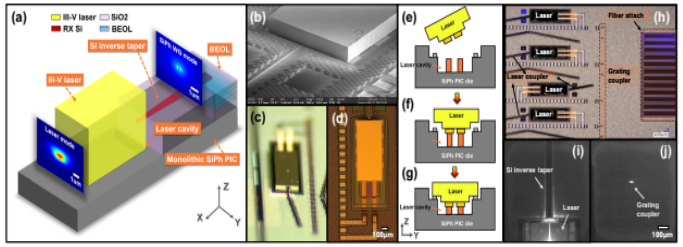


Fig.1. Laser attach on GF monolithic SiPh platform. (a) 3D schematic of the laser-integrated SiPh PIC; (b) SEM image of the laser diode in a cavity; (c) 3D pespective microscope image of the laser attach with an angled inverse taper; (d) Top-down microscope image of the laser with a straight inverse taper; (e) –(g) Process flow showing flip-chip attach of the laser diode inside the PIC cavity. (e) before alignment; (f) during alignment; (g) after reflow. (h) Microscope top view of a testing chip that comprises laser attach, grating coupler-based testing macro for wafer-level testing and a cavity to V-groove for chip level fiber attach testing; (i)-(j) Laser emission and output beam measured from the grating coupler.

ECTC 2021

Single-Chip Beam Scanner with Integrated Light Source for Real-Time Light Detection and Ranging

Jisan Lee[1*], Dongjae Shin[1], Bongyong Jang[1], Hyunil Byun[1], Changbum Lee[1], Changgyun Shin[1], Inoh Hwang[1], Dongshik Shim[1], Eunkyung Lee[1], Jinmyung Kim[1], Kyunghyun Son[1], Tatsuhiro Otsuka[1], Kyoungho Ha[1], and Hyuck Choo[1]
[1]Imaging Device Lab, Samsung Advanced Institute of Technology, Samsung Electronics, email: jisan2.lee@samsung.com

***Abstract*— For the first time to our knowledge, we present a single-chip solution for a solid-state 2D beam scanner achieving 10-m light detection and ranging (LIDAR) operation at 20 frames per second (fps). The beam scanner is integrated with a fully functional 32-channel optical phased array (OPA), 36 optical amplifiers, and a tunable laser, all on a 7.5×3-mm^2 single chip fabricated using III-V-on-silicon processes. In addition, we created and applied an ultrafast self-evolving OPA-calibration algorithm and digital signal processing to demonstrate real-time LIDAR operation. This work presents the first demonstration of a chip-scale LIDAR solution without using an external optical source or amplifier, making an ultra-low cost and compact LIDAR technology a reality.**

ECTC 2021

Samsung

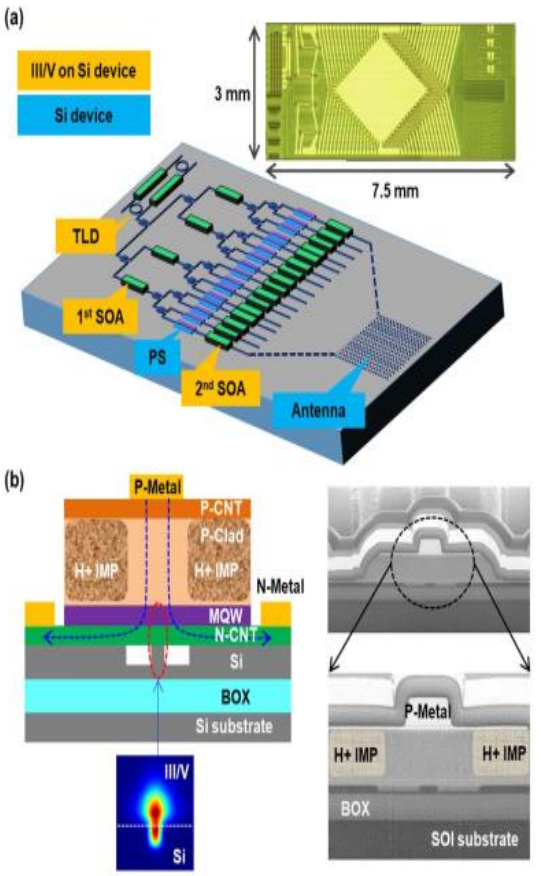


Fig. 1. Lateral and vertical structure of solid-state beam scanner. (a) Illustration and microscope image of fully-integrated 32-channel scanner. The chip size is 7.5 mm × 3.0 mm. (b) Illustration and vertical SEM image of III/V on Si device (TLD and SOA).

HB for NVM (STTMRAM) on Logic:

System exploration and technology demonstration of 3D Wafer-to-Wafer integrated STT-MRAM based caches for advanced Mobile SoCs

M. Perumkunnil, F. Yasin, S. Rao, S. M. Salahuddin, D. Milojevic, G. Van der Plas, J. Ryckaert, Eric Beyne, A. Furnémont, G.S. Kar
imec, Leuven, Belgium, Email: Manu.Perumkunnil@imec.be

***Abstract*—This paper analyzes the most feasible 3D integration and partitioning scheme for STT-MRAM based caches in an advanced Mobile SoC based on the process demonstration of the first ever functional 3D integrated STT devices. We present 3D partitioning schemes from a design - architecture perspective and Power Performance and Area (PPA) analysis is carried out for the 2D and 3D SoC designs with both SRAM and STT-MRAM caches. Our work shows that the PPA benefits from 3D Memory on Logic partitioning are magnified when it can be exploited to accommodate larger caches in general. We also show that STT-MRAM based 3D partitioned caches can exploit this potential increase in capacity to improve performance even more than SRAM. These 3D Wafer-to-Wafer (W2W) integrated STT-MRAM caches can result in up-to 30% performance improvement at 17% power and 15% footprint reduction for our target SoC.**

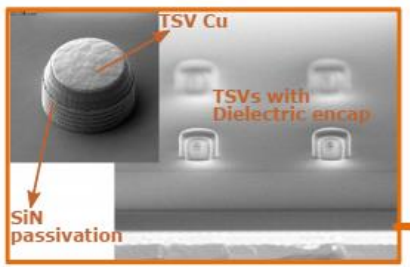
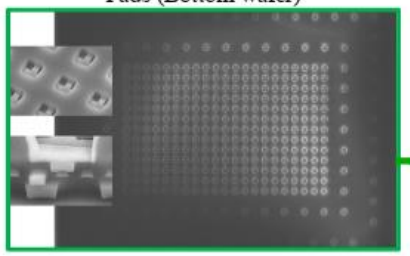


Fig. 5. TSV reveal on the TOP wafer

Fig. 6. Top view of the etched Bottom Pads (Bottom wafer)



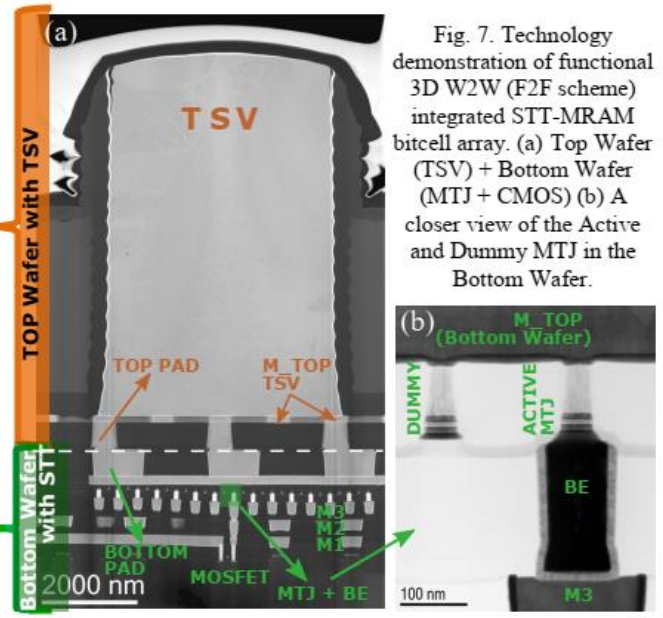


Fig. 7. Technology demonstration of functional 3D W2W (F2F scheme) integrated STT-MRAM bitcell array. (a) Top Wafer (TSV) + Bottom Wafer (MTJ + CMOS) (b) A closer view of the Active and Dummy MTJ in the Bottom Wafer.

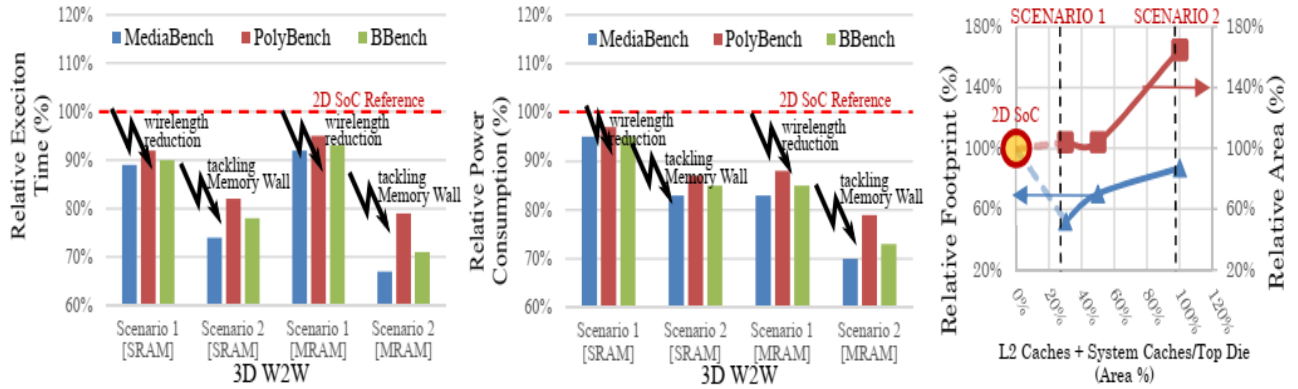


Fig. 11. Comparative a) Performance, b) Power and c) Area analysis for Scenario 1 and 2 with SRAM and MRAM (L2 caches and SLC) as the cache memories.

IMEC, VLSI 2021

TSMC HB enabled Extreme Disagg for HPC
[HB enabled D2D parasitic scaling]

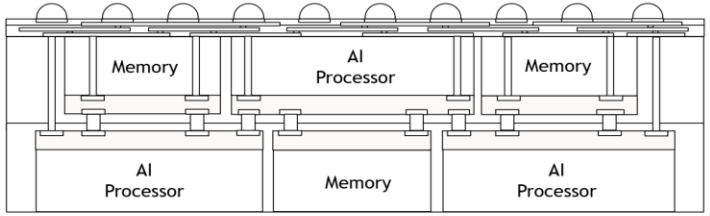
Motivation & Challenges

- Develop a technology to solve the challenges of compute wall, memory wall, and connectivity wall on computing systems.
- HPC for AI application faces compute wall, memory wall, and connectivity wall challenges.
  - Compute wall:
    - Limit in transistor numbers from reticle size, yield and cost constraints.
  - Memory wall:
    - Limits in on-chip memory capacity and bandwidth between off-chip memory and compute chip.
  - Connectivity wall:
    - Limit in physical connectivity numbers between heterogeneous technologies.
    - Limit in connection density between artificial neuron chips in brain-inspired computing system.

[1] M. D. Bishop et al., IEEE Micro, 2019, pp. 16-27.

Immersion in Memory Compute (ImMC) Technology

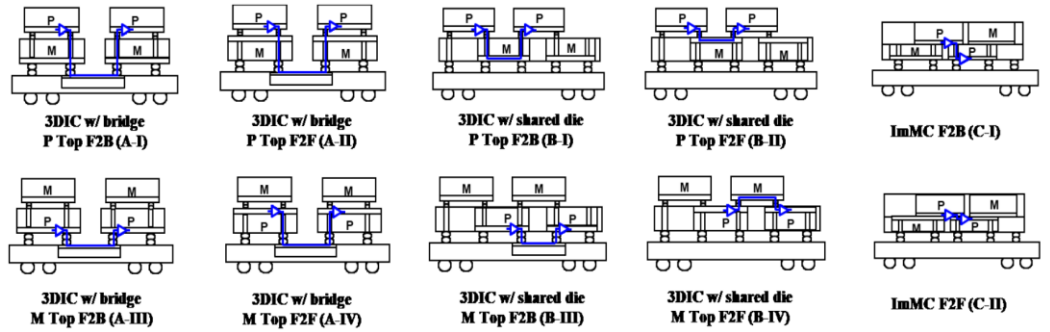
- ImMC technology structure:



✓ Multiple AI processors F2F bond at die edge.
✓ Partitioned cache memory F2F bond to AI processor with high BW density.
✓ SoIC™ bonding technology for the F2F bonds.

- **The ImMC technology is the solution to solve all challenges altogether.**

Processor-to-Processor (P2P) Interconnect Performance



Interconnect Benchmark — 3DIC: μbump pitch 36 μm, TSV height 50 μm; ImMC: Bond pitch 9 μm.

Structures	A-I	A-II	A-III	A-IV	B-I	B-II	B-III	B-IV	C-I	C-II
Bump Density	1x	1x	1x	1x	2x	2x	1x	2x	16x	16x
Speed†	1x	1x	1.2x	1x	1.4x	1.7x	1.2x	1.7x	200x	300x
Bandwidth Density††	1x	1x	1.2x	1x	2.8x	3.6x	1.2x	3.6x	3150x	6200x

†Speed: 1/total wire delay ††Bandwidth Density: Bump Density*Speed

- C-II F2F ImMC better than A-I F2B 3DIC:16x in bump density, 300x in speed and 6200x in bandwidth density

Conclusions

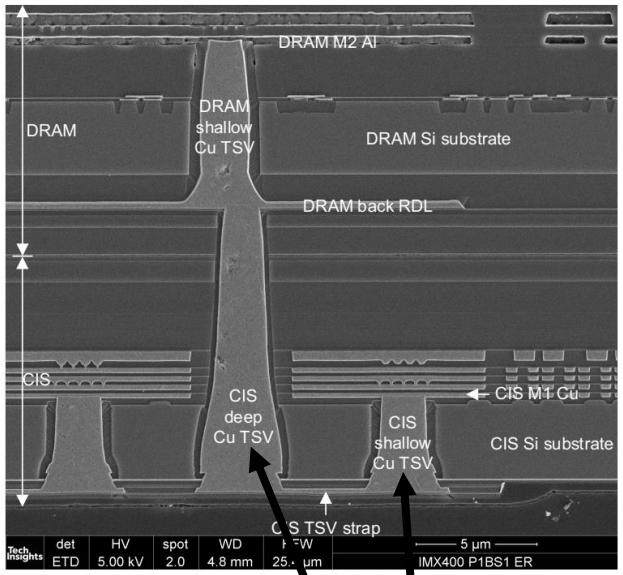
- **ImMC technology can offer multiple chips in multi-layer structure with ultra-short and ultra-high-density interconnects for system integration to solve compute, memory and connectivity wall challenges.**
- **ImMC technology has better electrical performance than conventional 3DIC structures.**

➢ **PPA improvement for P2P from bridge structure to ImMC**
  - **Bandwidth density: 224x improvement.**
  - **Energy/bit: 98% total power reduction.**
  - **Driver area: 99% area reduction.**

➢ **PPA improvement for P2M from mBump structure to ImMC**
  - **Bandwidth density: 20x improvement.**
  - **Energy/bit: 94% total power reduction.**
  - **Driver area: 75% area reduction.**

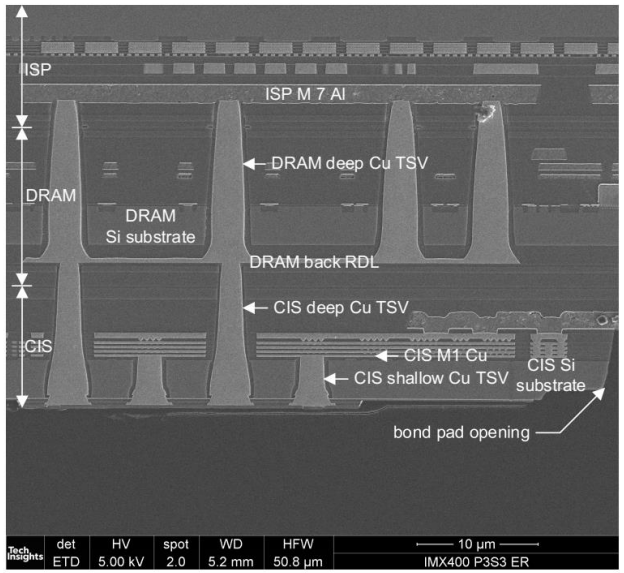
Sony 3 Wafer CIS

IMX400 TSVs Overview – Cross Section



\6 SEM_Cross-section_Images\...BS1\325_CIS+DRAM_TSVs_5K_224636

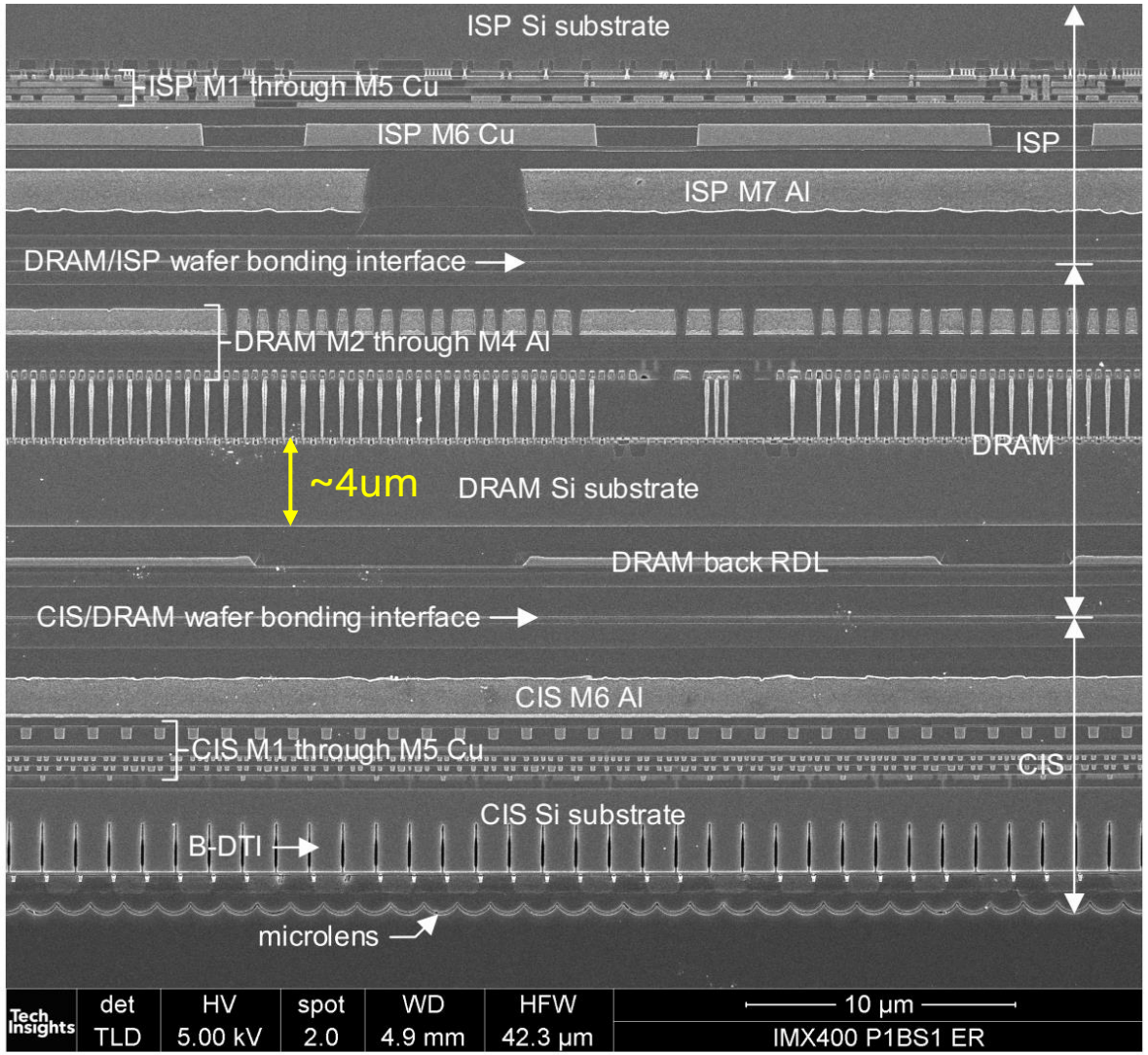
IMX400 CIS/DRAM Column TSVs



\6 SEM_Cross-section_Images\P3S3\803_CIS+MEM+ISP_BondPad_Edge_TSVs_2p5K_240795

IMX400 CIS/DRAM Bond Pad TSVs

Multi-Depth TSV (+RDL) for D2D connections in 3D Stack



\6 SEM_Cross-section_Images\P1BS1\518_CIS_PixelArray_General_Structure_3K_224636