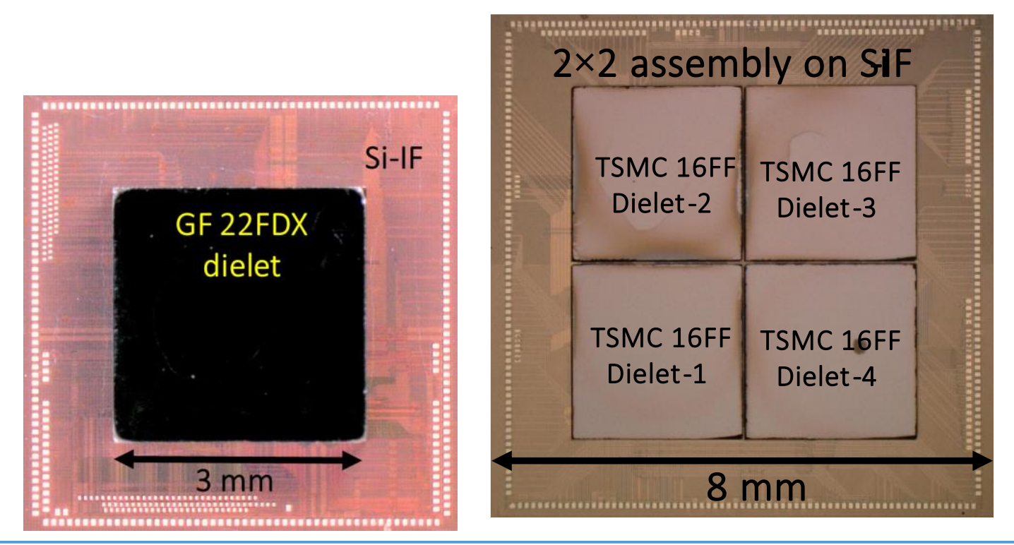
A High Throughput Two-Stage Die-to-Wafer Thermal Compression Bonding Scheme for Heterogeneous Integration

Krutikesh Sahoo, Haoxiang Ren, Zoe Chen & Subramanian S. Iyer | UCLA CHIPS

Introduction

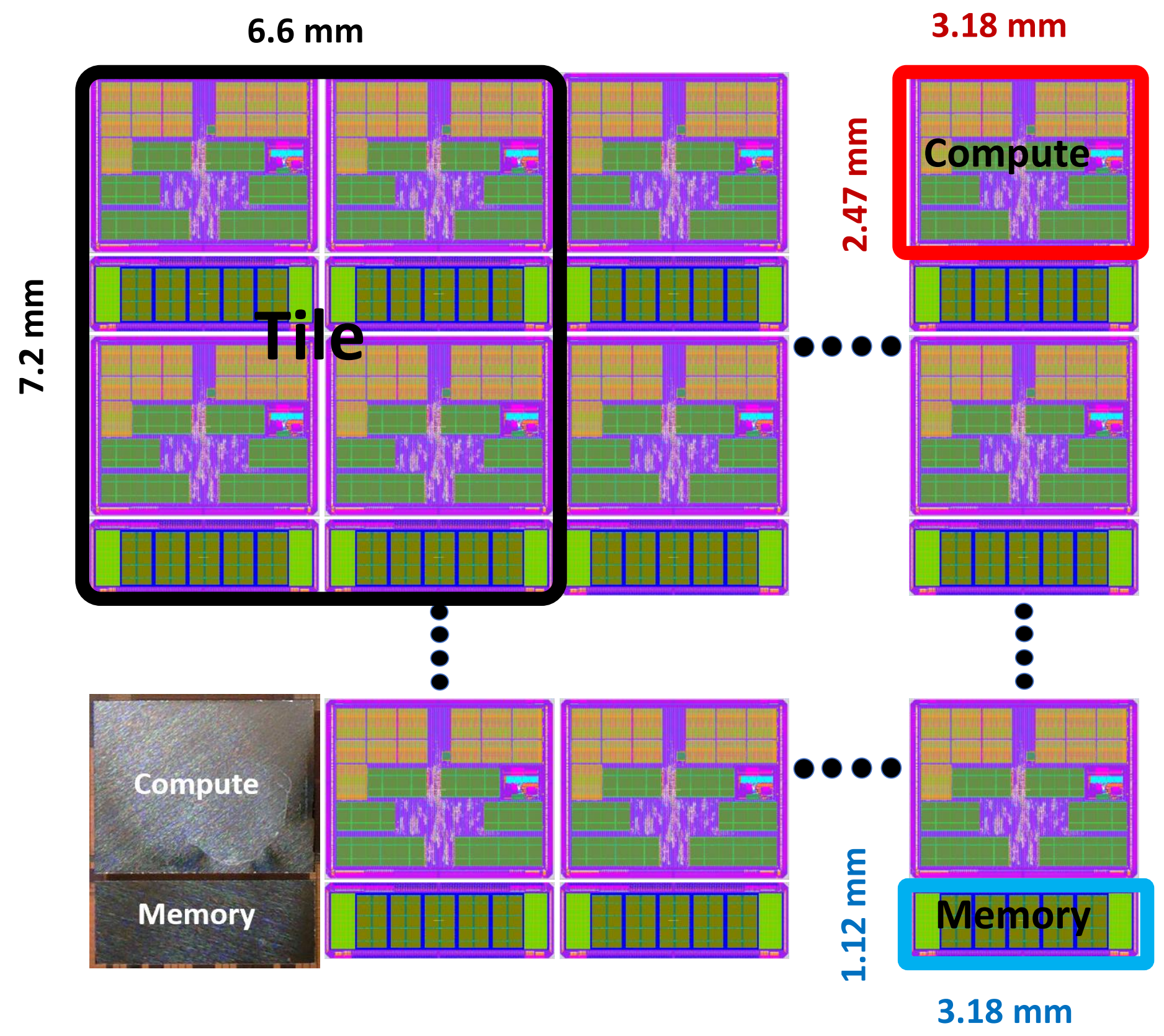
Cu-Cu thermal compression bonding (TCB) for die to wafer (D2W) assemblies

- First ever demonstration of a functional multi-chiplet assembly at sub-10 µm Cu-Cu bonding pitch using foundry supplied dielets[1-4].
- Key enabler for Cu-Cu TCB is the in-situ formic acid vapor cleaning process[1] which takes place within the bonding chamber.



Several functional dielet assemblies on Silicon Interconnect Fabric [1-3].

Building a graph processor on a 300 mm wafer by scaling down and scaling out TCB

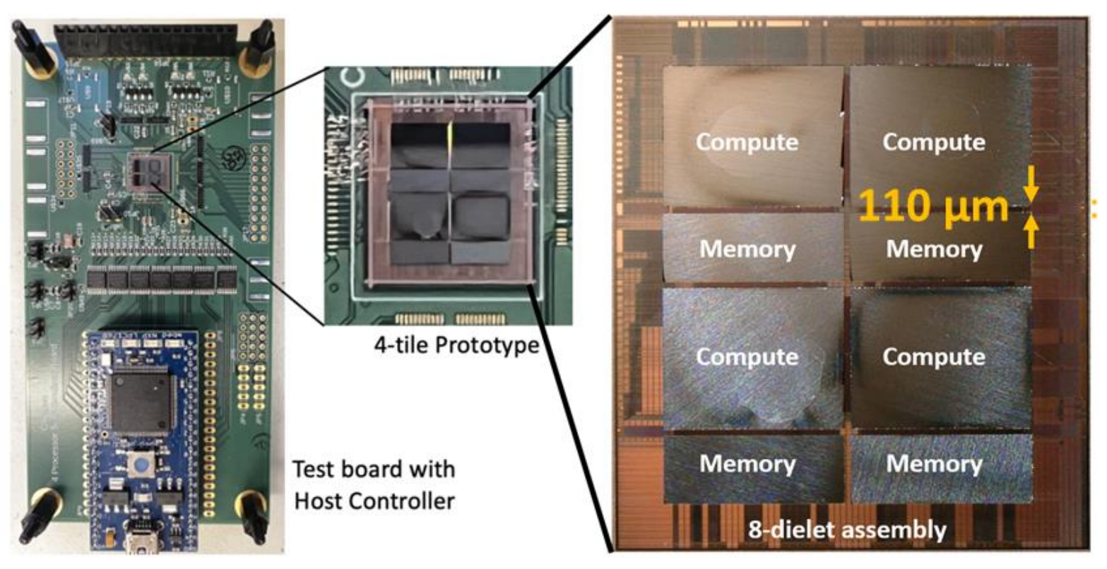


- **Scaling down:**
  - Chiplet size (< 100 mm^2)
  - Inter-chiplet spacing
  - Bonding pitch (< 10µm)
- **Scaling Out:**
  - Processor Architecture to wafer-scale
  - Memory architecture to Shared & Unified memory

***dielet designed jointly by UCLA and UIUC and fabricated at TSMC, and the Si-IF substrate fabricated at UCLA**

Lessons from prototyping

Small-scale prototyping



Small-scale prototype of graph processor

- Eight dielets per tile built in TSMC 40LP process:

  4 Compute dielets – **14** ARM CORTEX-M3 per dielet

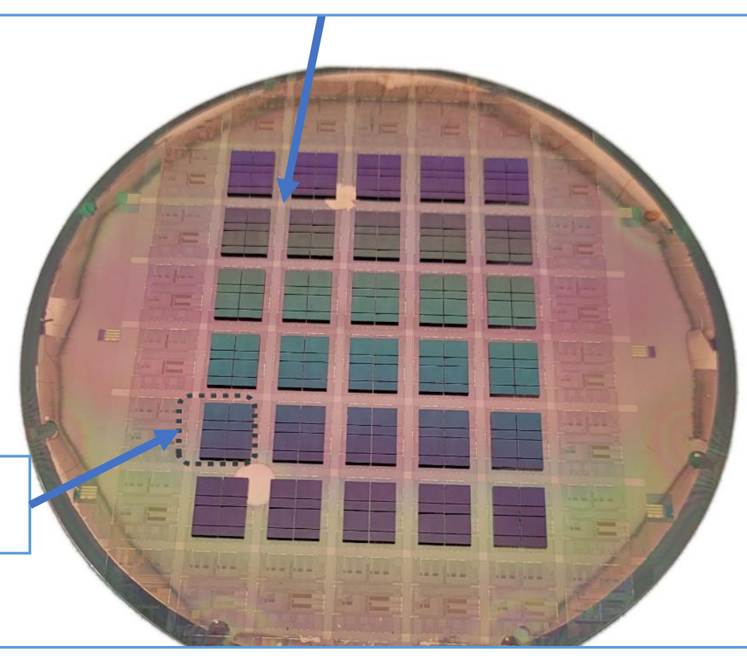
  4 Memory dielets -- **640**KB SRAM per dielet

- 240 passive dielet prototype on 100 mm wafer to identify challenges with Scale out of thermal compression bonding:

  120 dielets: 3.5x2.5 mm^2

  120 dielets: 3.5x1.5 mm^2

Gap between tiles for testing and front-side TSV-less power delivery [1]



1 tile = 8 dielets

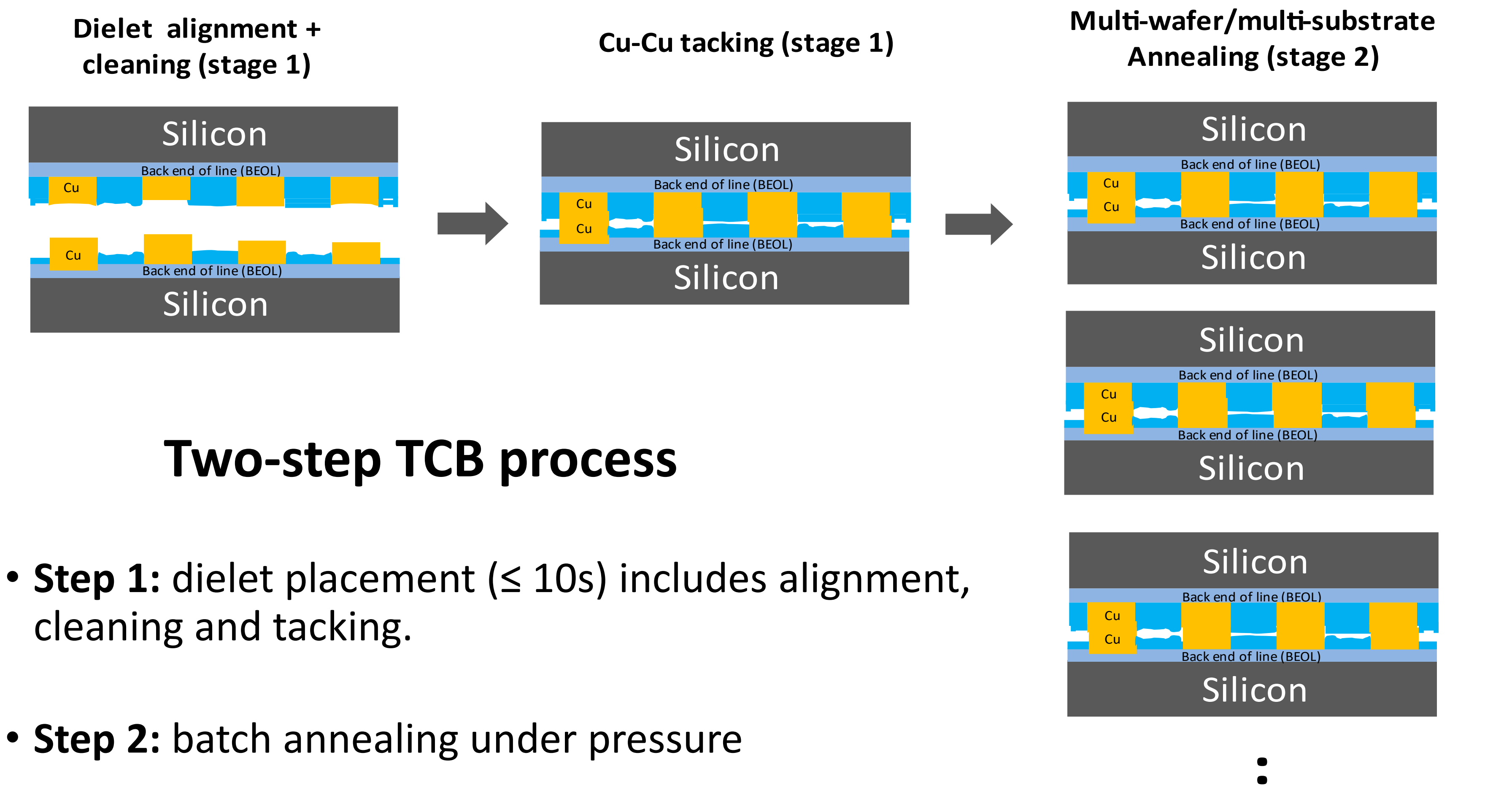
Prototype 100 mm Si-IF populated with compute and memory dielets.

- All Interconnected on Silicon interconnect Fabric

Learnings from Small-scale prototyping

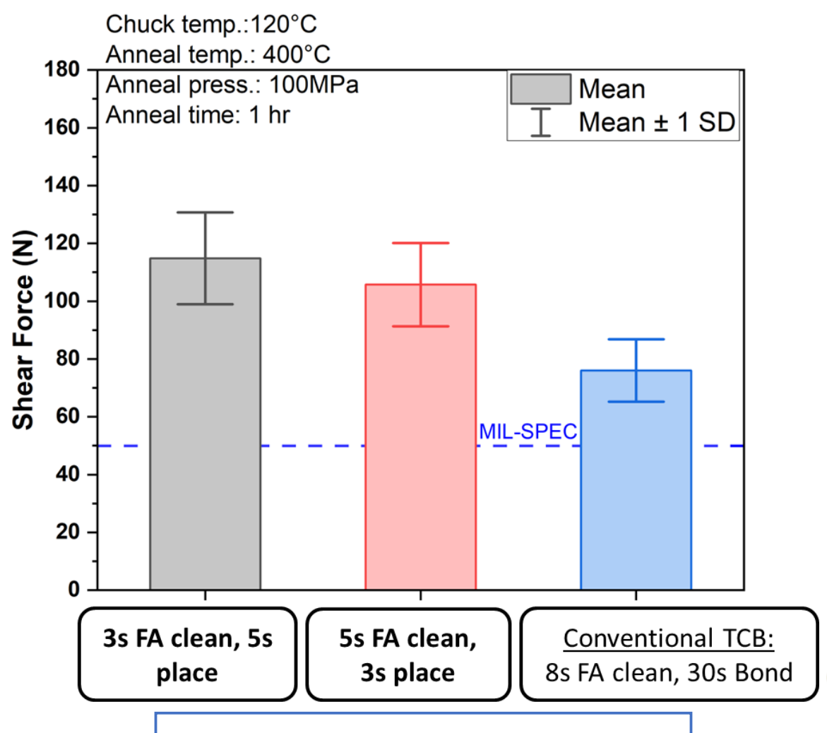
- Sequential single step TCB of dielets results in low throughput (30 seconds/dielet) and combines both placement and temperature/pressure induced grain growth.
- Bonding time of 30 – 40 seconds/die => 2 – 2.5 hours to populate a wafer-scale sample.
- Scaling this process to 300 mm wafers requires long assembly times leading to re-contamination of the Cu pillars on the substrate and requires periodic recleaning of the surface.
- **A higher throughput process is therefore desirable.**

Two step TCB process for higher throughput

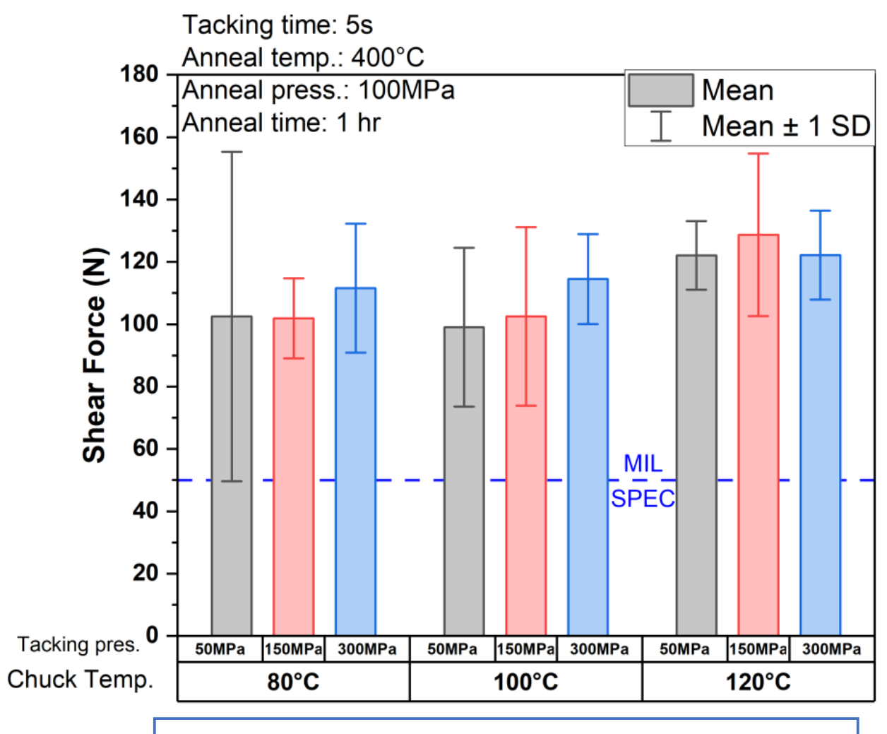


Two-step TCB process

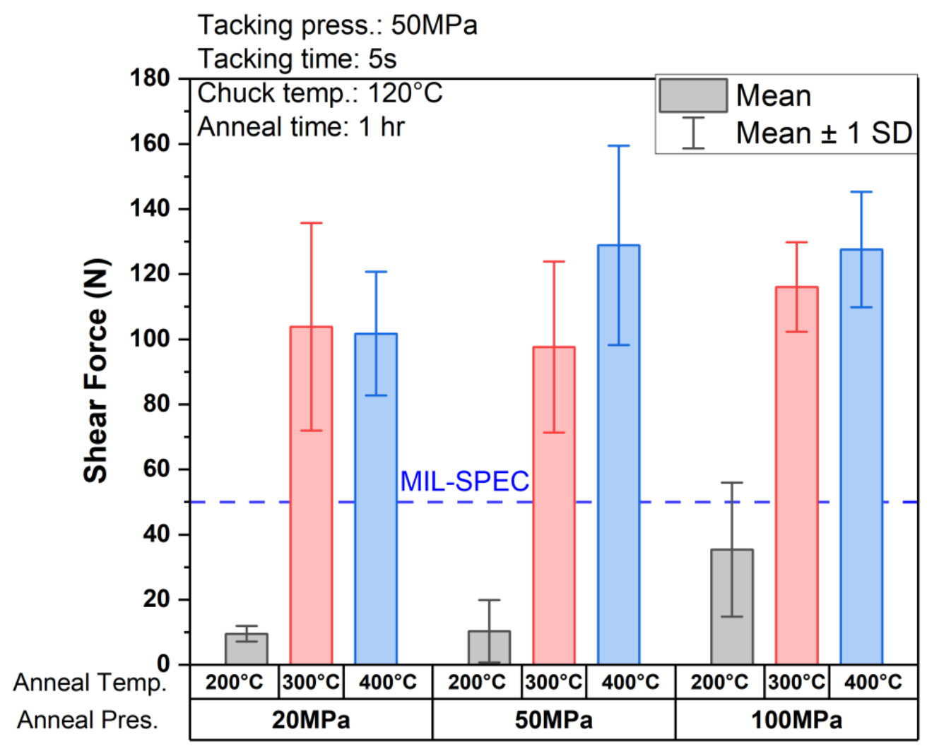
- **Step 1:** dielet placement (≤ 10s) includes alignment, cleaning and tacking.
- **Step 2:** batch annealing under pressure



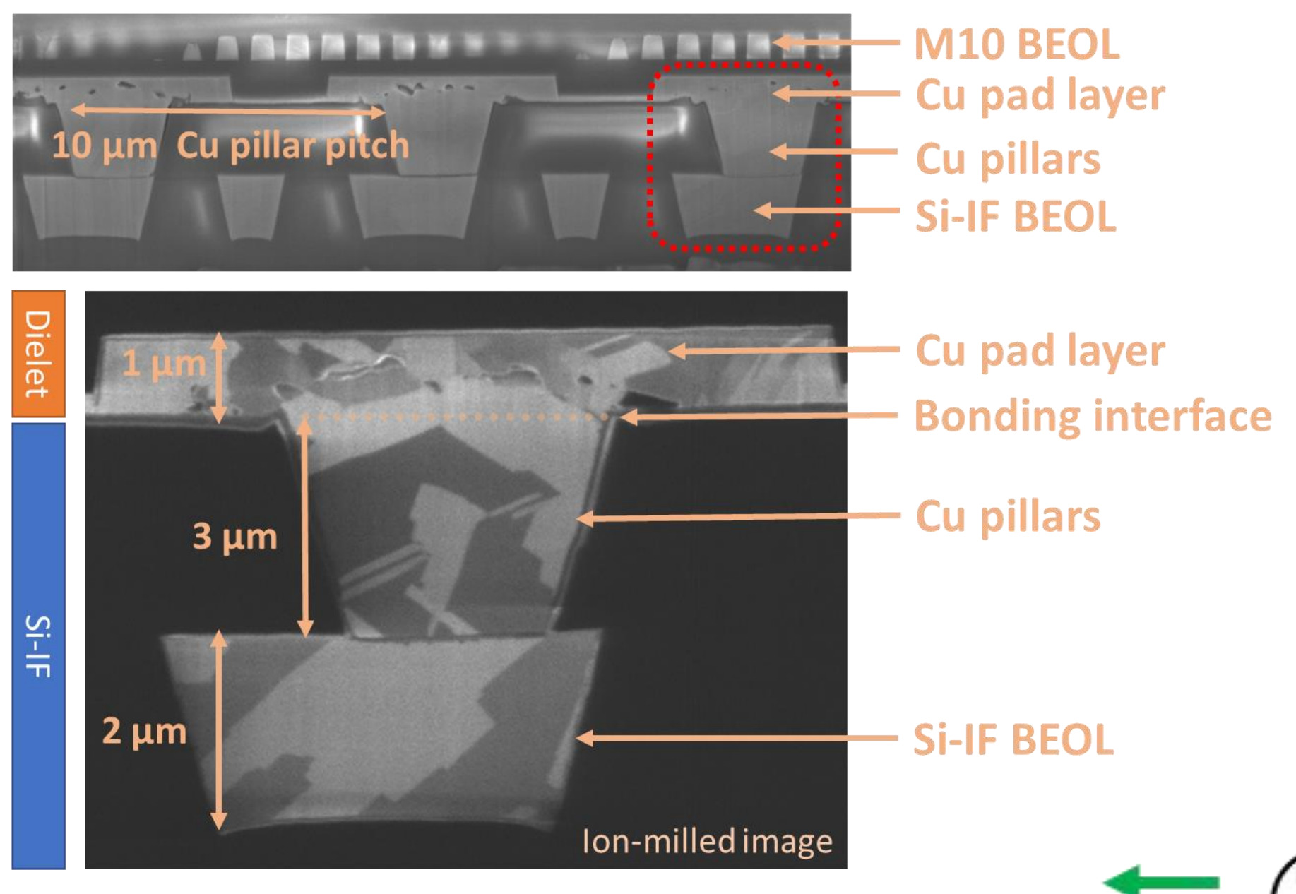
P1: tacking time opt.



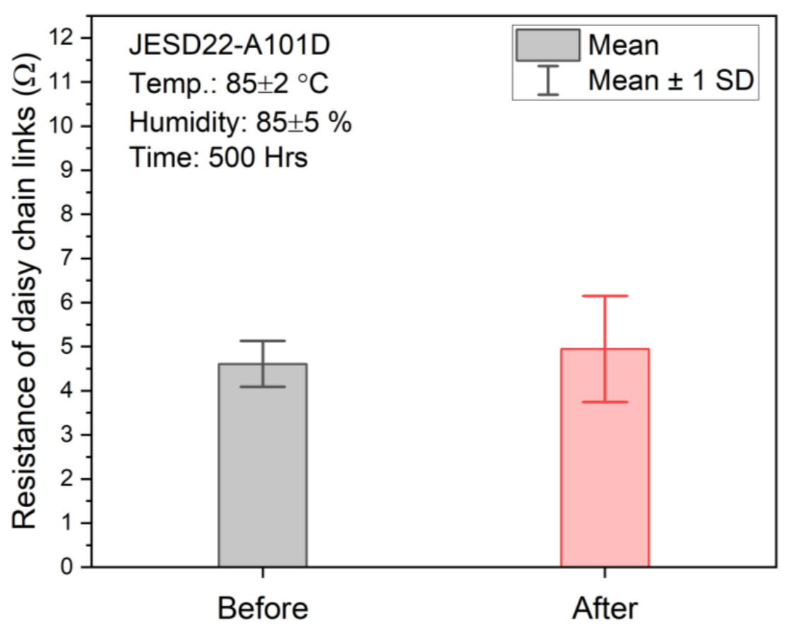
P2: tacking parameters opt.



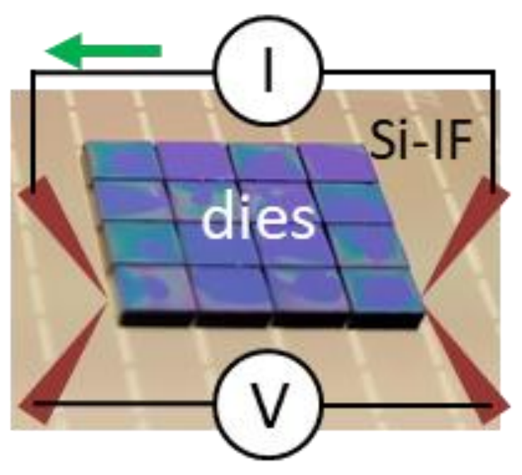
P3: annealing opt.



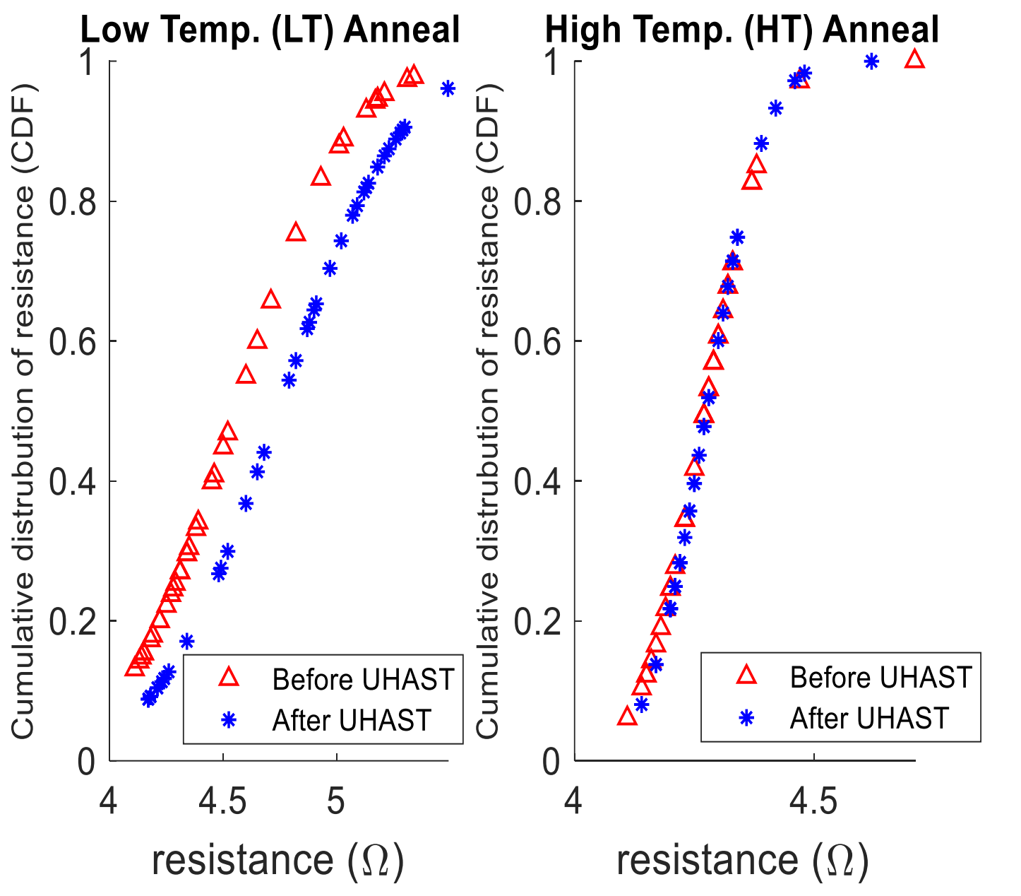
Parameter name	Low Temp. **(LT)** anneal	High Temp. **(HT)** anneal
Die cleaning time (s)	3	3
Die placement time (s)	5	5
Chuck temperature (°C)	120	120
Tacking pressure (MPa)	50	50
Annealing temp. (°C)	300	400
Annealing press. (MPa)	100	50



JEDEC test of 72000 Cu-Cu connections using two-step TCB for 500 hours[5].



Multiple Daisy Chain Samples



Cumulative resistance distribution of 40 daisy chains[5] before and after Unbiased Highly Accelerated Stress Tests (UHAST)

A comparison with Hybrid bonding

	Hybrid Bonding	Direct thermal compression bonding
CMP process Development	Strict dielectric flatness (6σ roughness ≤ 1 nm) and metal recess requirements	Relaxed metal roughness requirements (6σ roughness ~10nm acceptable)
Dicing process	Mandatory particle-free dicing.	Cu pads/pillars are recessed so, blade dicing with standard wet cleaning is feasible.
Bonding environment & activation	ISO-4 or below	ISO-8 and above, even outside cleanroom.
Dielet size	Almost any size due to less tacking pressure requirements.	Limited by max. bond-head pressure during tacking.
Throughput	1000+ units-per-hour (UPH) due to fast dielectric tacking	1000+ UPH possible with **optimized tack and anneal process**.
Conclusion	• **TCB has low process development cost & operation cost compared to HB.** • **TCB is less sensitive to particles during dicing and bonding.** • **Therefore, we believe TCB should be used for bonding pitches up-to 7 µm.**	

References: [1] S. Jangam and S. S. Iyer, TCPMT 2021, [2] K. Sahoo et al., ECTC 2022, [3] S. Nagi et al., JSSC 2022
[4] U. Rathore et al., ISSCC 2022, [5] Sahoo et al., ECTC 2023

Acknowledgements: The Authors thank SRC JUMP 2.0 CHIMES and UCLA CHIPS Consortium for supporting this work.

2023 UCLA CHIPS Workshop
Poster Session