White paper · 12 min read

Inline AFM for hybrid bonding: closing the metrology gap in HBM manufacturing

Why the move to Cu-to-Cu hybrid bonding turns HBM yield into a nanometre-scale metrology problem, why neither optical tools nor single-tip AFM solve it alone, and how parallel, miniaturized AFM at the tool can.

Akshay Harlalka · November 3, 2026

Summary

High-bandwidth memory (HBM) is now the largest component cost in AI accelerators, and the industry is preparing to move future HBM stacks from microbumps to direct Cu-to-Cu hybrid bonding. Hybrid bonding only works when every copper pad sits inside a recess window of a few nanometres and the dielectric is smooth to a fraction of a nanometre. Atomic force microscopy (AFM) is the standard way to measure those surfaces, but a single tip images an 80 × 80 µm area in about five minutes, under 0.1 mm² per hour, so fabs can only sample a handful of sites. This paper explains the gap and describes Pivector's approach: miniaturized 20 mm AFM heads operating in parallel, and a multi-head module in development that works inside a FOUP at the process tool.

1. Memory is the bottleneck of AI hardware

Every leading AI accelerator is surrounded by stacks of HBM. NVIDIA's Rubin platform, for example, carries 288 GB of HBM4 in eight 12-high stacks. By Epoch AI's estimate, HBM accounted for about 63% of AI-chip component

spend in late 2025. Supply has struggled to keep up: SK hynix expects memory demand to exceed supply through 2030, and Micron has agreements covering the vast majority of its 2027 HBM bit supply.

Each HBM stack is a tower of thinned DRAM dies. If any die in a stack is defective, the entire stack is discarded, including every good die bonded to it. Yield losses therefore compound with stack height, and they are paid at the most expensive point in the flow.

2. Hybrid bonding moves the problem to the surface

Today’s HBM joins dies with microbumps, using mass-reflow molded underfill (MR-MUF) or thermocompression bonding with non-conductive film (TC-NCF). As stacks grow taller inside the JEDEC package-height limit, hybrid bonding is expected in the HBM4E and HBM5 generations: dielectric bonds to dielectric at room temperature, then an anneal expands the copper so pads fuse directly. No solder and no bump gap, so finer pitch, thinner stacks and better heat flow.

The price is extreme surface control. After CMP, the copper pad must sit slightly below the dielectric so the dielectric bonds first, but only by a few nanometres, so the copper can still close the gap during anneal. Published targets are on the order of:

Surface parameter	Published figure
Cu recess	A few nanometres
Dishing	≤ 5 nm
Planarity	About 0.5 nm RMS
Dielectric roughness (Rq)	0.1–0.2 nm

Copper atoms are about 0.26 nm apart, so a recess of 1–3 nm is only about 4 to 12 atoms deep. Pads that are too deep never connect; pads that protrude hold the dielectric apart and leave voids. Both failures show up only after bonding, when the stack is already built.

3. Why current metrology leaves a gap

Optical tools such as white-light interferometry and scatterometry are fast and essential. Interferometers see copper and dielectric with different phase offsets, and scatterometry fits a model to periodic targets; both are calibrated against AFM. As interconnect pitch shrinks from around 10 µm toward 2 µm and below, each pad spans only a few optical resolution elements, and pad-level 3D numbers become harder to trust without an AFM reference. Research has already shown sub-micron pitch.

AFM measures height directly, with sub-nanometre vertical resolution and no optical model, which is why it is the standard method for Cu recess. Its problem is speed. Imaging an 80 × 80 µm area takes about five minutes. Scaled to a full 300 mm wafer (about 70,700 mm²), one tip would need on the order of a century. In practice, fabs measure a few sites on a few wafers per lot, and drift between those points surfaces later as bond failures.

> Fast optical tools need a reference, and the reference, AFM, is not fast. Production hybrid bonding needs a reference that keeps up.

4. Pivector's approach: miniaturize, parallelize, integrate

Pivector keeps the physics of AFM and changes the architecture.

- Miniaturize. A conventional AFM head is about 100 mm across. Pivector's head is 20 × 20 mm, small enough to tile many over one wafer.
- Parallelize. Heads scan different sites at the same time. The engineering challenge, and our focus, is running them together: synchronization, calibration, stability, electronics and control.
- Integrate. The production module is designed in a FOUP form factor for a standard 300 mm load port, so wafers never leave the automated flow, and it reports to the fab host over SECS/GEM for run-to-run control.

With 40+ heads, the target is 50–100 mm^2 of AFM data per hour, against under 0.1 mm^2 for one tip at the published rate. That needs both more heads and faster scanning per head. Even at the target, imaging a whole wafer would take one to two months, so the aim is not a full-wafer map but enough sites on each wafer to see within-wafer and edge variation, instead of a few sites per lot.

5. What has been demonstrated

Milestone	Status
Single-head AFM, manual benchtop	Done
Single-head AFM, automated; < 2 nm vertical resolution without vibration isolation	Done
20 × 20 mm miniaturized head	Done
Dual-head prototype scanning a 300 mm wafer	Done
Dual-head synchronization, calibration and control	In progress
Multi-Head AFM pilots with partners	Next
First AFM FOUP systems	Planned for 2027

< 5 Å (0.5 nm) vertical resolution, < 1.5 nm lateral resolution, 40+ heads and 50–100 mm² per hour are engineering targets for the production system. At < 2 nm, today's prototype cannot yet measure dielectric roughness at the 0.1–0.2 nm level. We expect resolution to improve with electrical noise reduction and acoustic and vibration isolation, which the benchtop prototype does not yet have.

6. What this enables

- Bond-readiness decisions before stacking, based on many sites per wafer rather than a few per lot.
- Recess-based run-to-run CMP control: correct the polish on the next wafer, not the next lot.

- Faster recipe development for each new bonding layer, from dense maps on every learning wafer.
- A dense AFM reference to calibrate fast optical metrology on the same wafers.

7. Beyond HBM

The same need for fast, direct 3D surface data appears in SiC and silicon wafer polishing, EUV mask blanks and high-NA resist (where an electron beam can shrink the resist it measures), and process steps in gate-all-around and CFET transistors.

Work with us

We measure customer samples on our prototype, work with the NYCU 3DIC Research Center as our validation partner, and are planning partner pilots of the Pivector Multi-Head AFM. Contact contact@pivectorlabs.com or book a call at pivectorlabs.com/contact.

Sources

1. Epoch AI, AI chip component cost shares (May 21, 2026)
2. SK hynix, 3Q26 financial results (Oct 28, 2026)
3. Micron FQ4 2026 earnings call (Sep 30, 2026), via TrendForce and Reuters
4. NVIDIA, Vera Rubin NVL72 platform specifications
5. TrendForce on HBM stack yield (Mar 6, 2024) and hybrid bonding adoption (Jul 7, 2026)
6. imec (2024) and Besi via SemiEngineering (2026) on Cu recess, dishing and planarity targets

7. Bruker application note AN5001 on dielectric roughness for direct bonding
8. Adeia via 3D InCites (May 21, 2024): AFM as the standard method for Cu recess; 80 × 80 µm in about 5 minutes
9. Throughput per wafer and atoms-per-nanometre figures are derived estimates (Cu lattice constant 361.49 pm).