Scaling Up

The MonolithIC 3D Record of Invention, 2010-2026

"Dimensional scaling is ending. Scaling up is the way forward. Here is the practical, cost-effective way to do it."

A documented sixteen-year public record, and the industry that followed it.

MonolithIC 3D Inc. • Rev. 12, October 2026

I. The Case in One Paragraph

Beginning in 2010, MonolithIC 3D Inc. told the semiconductor industry, in keynotes, peer-reviewed papers, book chapters, free e-books, open simulators, and more than a decade of continuous public blogging, that dimensional (2D) scaling was approaching its economic end, and that the industry's future lay in *scaling up*: building integrated circuits vertically, using layer transfer, precision wafer bonding, and ultimately hybrid bonding, on existing fab equipment at existing nodes. The company did not merely predict the destination; it published the practical process flows, the cost models proving economic viability, the thermal solutions answering the standard objections, and the system architectures (memory-on-logic, two-wafer NAND, 3D DRAM) that define today's memory industry. At the heart of the thesis is the heterogeneous case: logic, memory, analog and photonics are made in different fabs on different processes, and today's systems join them on a circuit board, or in the industry's first small steps a 2.5D interposer (now itself moving to bumpless hybrid bonding, as in Apple's M5 Pro; TechInsights, June 2026), at a large cost in power, performance and area; scaling up bonds these strata with millions of vertical connections microns long, so that the system is integrated in the fab rather than on the board (2011 keynote; IntSim, 2011; "The Future is the Interconnect," 2012; "Is the Cost Reduction Associated with Scaling Over?", 2012; the 1,000x paper, 2017). In parallel, it filed and prosecuted the patent estate covering that path: 402 issued US patents as of May 2026, with many more applications pending, the earliest grants issuing in 2011 from applications filed in 2009-2010. Every major element of the thesis has since been adopted by the industry: YMTC's Xtacking (2018), AMD's 3D V-Cache (2022), Kioxia/SanDisk's CMOS-directly-Bonded-to-Array NAND (2023-2026), hybrid bonding on every HBM roadmap, $4F^2$/vertical 3D DRAM as the industry's declared next architecture, and, in mid-2026, memory-on-logic near-memory computing announced as the datacenter roadmap of Qualcomm (High Bandwidth Compute) and placed by TSMC on its 2029 roadmap as DRAM-on-logic, while Huawei codified equivalent scaling through 3D architecture as its "Tau Scaling Law" and, in September 2026, shipped it in a phone. **The publications came first. The patents came first. The industry followed.**

II. The Thesis Stated Early (2010-2011)

The founding claim was made publicly, by name, in the industry's own venues, before any of today's commercial 3D memory products existed:

- **January 2010**: Invited presentation at the Applied Materials 3D Interconnect Symposium: "3D-FPGA: The Path to ASIC-like Density, Power and Performance."
- **December 2010**: Invited presentation at the 3D-ASIP Conference: "Monolithic 3D Technology: A Semiconductor Industry Game Changer."
- **February 2011**: "Redundancy & Repair with Monolithic 3D" presentation (February 4, 2011), published on the company site: 3D stacked redundant logic replaceable at logic-cone granularity and a single stacked repair layer, stated to enable "ultra-large scale integration, up to wafer scale integration, without yield issues" and "high yield wafer-scale integration with current defect density," opening "new frontiers for supercomputing." The wafer-scale 3D teaching, fifteen years before Cerebras's CS-6.
- **March 2011**: The MonolithIC 3D technical blog launches; it has published continuously since, creating a dated, public, searchable record of the company's teachings for fifteen years. A parallel channel of dated technical articles has run on the founder's LinkedIn profile, and further teaching appeared as EE Times and Semiconductor Digest columns throughout.
- **April 2011**: "Monolithic 3D Integrated Circuits," Future-Fab International, Issue #37.
- **May-July 2011**: Invited talks at the International Memory Workshop ("Monolithic 3D FPGAs"), the Technical University of Munich, the American Vacuum Society Workshop on 3D Packaging ("Monolithic 3D DRAM Technology", the 3D DRAM teaching fifteen years before the industry's current 3D DRAM race), the CMOS Emerging Technologies Workshop, and Intel Research (Photonics Technology Lab).
- **September 2011**: Invited talk at UC Berkeley: "Resistive RAM: Technology and Market Applications."
- **October 7, 2011**: Invited keynote, International Electronics Forum, Seville: "Monolithic 3D-ICs: The Effective Alternative to Dimensional Scaling." The thesis, stated in the title, on a keynote stage, in 2011.
- **November 2011**: Invited lecture at the Technion: "Monolithic 3D - The Effective Alternative to Dimensional Scaling"; invited talk at the 9th International SoC Conference.
- **December 15, 2011**: Blog post "Monolithic 3D IC Could Increase Circuit Integration by 1,000x": the motivation (Nvidia data showing off-chip data movement takes >>100× the energy, with energy named the limiting factor of computing), yield identified as the limit to integration, and stacked logic-cone repair as the enabler, concluding that yield is removed as a constraint and "we could even integrate 1,000 such devices." The first reader comment, posted the same day, reads: "Extraordinary claims require extraordinary evidence." The skepticism stage, timestamped.
- **2011-2012**: Industry recognition of the breakthrough: finalist, "Best of Semicon West 2011"; finalist, EE Times Innovator of the Year, 2011 and 2012; named Top Embedded Innovator (Silicon) by Embedded Computing Design.

The patents were already issuing. The first US patents issued in June and July 2011, from applications with 2009-2010 priority, with grants issuing every year since, now totaling 402 issued US patents.

III. Making It Practical and Cost-Effective (2012-2015)

The industry's objections to 3D were always the same: too hot, too expensive, no manufacturable process. Between 2012 and 2015 MonolithIC 3D published the answers to each. This is the substance behind "practical and cost effective."

The process. The core enabler taught was layer transfer of ultra-thin (<100 nm) single-crystal silicon at under 400 °C (compatible with copper/low-k interconnect beneath) via Ion-Cut and Modified ELTRAN, delivering roughly 10,000× better vertical connectivity than TSV-based stacking:

- "Monolithic 3D-ICs with Single Crystal Silicon Layers," invited paper, IEEE 3DIC Conference, 2012.
- "Practical Process Flows for Monolithic 3D," Plenary Talk, IEEE S3S Conference, October 2013. Note the word "practical" in the plenary title.
- "Pulsed Laser Annealing: A Scalable and Practical Technology for Monolithic 3D IC," IEEE 3DIC 2013.
- "Scaling Makes Monolithic 3D IC Practical," Solid State Technology, October 2013. This is the elegant inversion: as dimensional scaling gets harder, its own thinner layers and lower thermal budgets make monolithic 3D easier. The two curves cross, and the industry changes direction.
- "Precision Bonders - A Game Changer for Monolithic 3D," IEEE S3S, October 2014, identifying precision wafer bonding, the enabling equipment class of today's hybrid bonding, as the practical vehicle, years before commercial hybrid-bonded memory existed.
- "Modified ELTRAN® - A Game Changer for Monolithic 3D," October 2015.

Power from below, and the cut-layer teaching. Two of the practical flow's teachings later became the foundations of backside power delivery. The first is power delivery from below: routing power to the transistors from beneath the device layer, taught in a US application filed October 2010 in a chain with priority to October 2009 (issued March 2013), including Figures 83L to 83L4. The second is the built-in cut layer: an epitaxial SiGe layer formed beneath the silicon device layer so that, after bonding, the wafer can be ground and then etched with the etch stopping selectively on the SiGe (a soft landing), the SiGe then removed to leave a precisely controlled, ultra-thin silicon layer. This SiGe etch-stop approach to precision extreme wafer thinning was taught in a PCT application filed September 21, 2016 and published March 30, 2017 (Figure 44, layer 4404), and was republished in book-chapter form in NANO-CHIPS 2030, Chapter 8, Section 8.3 (2020), including the flow: cuttable substrate, standard fab, precise bond, grind and etch to the cut layer.

The thermal objection, answered at IEDM. "Cooling Three-Dimensional Integrated Circuits Using Power Delivery Networks," IEDM, December 2012: the industry's premier device conference, on the field's most-cited objection to 3D.

The economics. Using the company's IntSim 3D CAD simulator (published free on the site, alongside its other open simulators), MonolithIC 3D quantified the claim: a two-layer monolithic 3D IC at the same node delivers ~2× power reduction, ~2× silicon-area reduction, and 4× footprint reduction (the benefits of a full node shrink) while using predominantly existing capital equipment and R&D, against leading-edge fabs then costing $5B+ per generation (and far more today). Cost-of-Ownership estimates built on the Sematech framework were published openly. This is the "cost effective" pillar, stated with numbers, in public, from 2011 onward.

The teaching infrastructure. Free e-books ("Monolithic 3D Advantage"; "Monolithic 3D Memory - NAND Flash, DRAM and R-RAM," March 2013), IEEE tutorials, university guest lectures, webcasts and webinars: an open campaign to teach the industry the path.

The industry began to listen. IEEE Spectrum, February 2014: "The Rise of the Monolithic 3D Chip." Qualcomm's VP keynoted 2015 ISPD on 3D VLSI; EE Times reported "3D Qualcomm SoCs by 2016"; CEA-Leti built CoolCube and extended it with Qualcomm; Stanford's N3XT program (December 2015) announced "skyscraper-style" chips boosting performance "by a factor of a thousand". The community was converging on the direction MonolithIC 3D had been mapping since 2010.

IV. The Economic Call: "Moore's Law Stopped at 28nm" (2014 → validated 2023)

In 2014, while Intel was publishing roadmaps promising continued cost-per-transistor reduction for the foreseeable future, Zvi Or-Bach published the contrarian, falsifiable claim that **transistor cost scaling had stopped at the 28 nm node** (EE Times, "Intel vs. Intel," 2014, with a deeper technical analysis in 2016 covering lithography cost escalation and transistor complexity). The claim was widely covered (Design & Reuse, EE Times Asia, SmartBrief, and others) and widely doubted.

Nine years later it was confirmed from the industry's own podium. At IEDM 2023, Short Course SC1.6, Milind Shah of Google stated: *"Transistor cost scaling (0.7X) stalled at 28 nm and remains flat gen over gen."*

This matters beyond vindication: the end of cost scaling is the **economic premise** of the entire scaling-up thesis. If cost-per-transistor no longer falls with node shrinks, the only remaining path to more capability per dollar is vertical integration on mature nodes, exactly what the memory industry then did.

V. The Memory Wall and the 1,000× Roadmap (2017-2020)

With the economic premise established, MonolithIC 3D published the destination architecture:

- **2017, IEEE S3S**: "A 1,000× Improvement in Computer Systems by Bridging the Processor-Memory Gap": hybrid bonding of memory directly onto logic to defeat the Memory Wall, quantified at three orders of magnitude.
- **September 2017, the US Government adopts the direction.** DARPA's Electronics Resurgence Initiative (announced September 13, 2017; $500-800M) included the 3DSoC program calling for monolithic 3D technology enabling greater than 50× improvement in SoC digital performance-at-power versus 7nm 2D CMOS, documented in real time in the company's September 19, 2017 article "DARPA Calls for Monolithic 3D - 3DSoC." At the same IEEE S3S 2017, Intel Senior Fellow Al Fazio's plenary presented 3D NAND and 3D XPoint as the trailblazing monolithic 3D technologies already matured to volume production.
- **2020, Springer, NANO-CHIPS 2030**: four monolithic-3D chapters, including Chapter 15, "A 1000× Improvement of the Processor-Memory Gap," and Chapter 11, "3D for Efficient FPGA": the thesis in book-chapter permanence, including the wafer-scale, redundancy-and-repair, and system-integration teachings that extend it.
- **Earlier foundations**: the redundancy-and-repair teachings answering Gene Amdahl's famous objection that wafer-scale integration "will only work with 99.99% yield, which won't happen for 100 years", published as part of the 3D-IC Edge series and patented in the "Automation for Monolithic 3D Devices" and repair families.

VI. The Validation Cascade: The Industry Executes the Thesis (2017-2026)

Each row below sets a dated MonolithIC 3D teaching against a later industry development that adopts, demonstrates or announces the same element, together with a few of the company's own forecasts and research milestones:

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
Sept 2017	**DARPA ERI / 3DSoC program** ($500-800M initiative): calls for monolithic 3D enabling >50× SoC performance-at-power vs 7nm 2D CMOS;	US Government funding of the scaling-up direction; Intel confirming memory as	Thesis publications 2010-2017; documented in real

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
	Intel's Al Fazio plenary at IEEE S3S 2017 presents 3D NAND / 3D XPoint as trailblazing monolithic 3D in volume production	the early adopter	time, Sept 19, 2017
Aug 2018	**YMTC Xtacking** launches at Flash Memory Summit ("Best of Show"): periphery CMOS wafer hybrid-bonded onto 3D NAND array wafer	Two-wafer NAND: logic wafer + memory wafer, bonded	Two-wafer / layer-transfer memory architectures published 2011-2015; patents from 2009-2010 priorities
Feb 2022	Alibaba, ISSCC 2022: a 3D logic-to-DRAM hybrid-bonded prototype, 55 nm logic on 25 nm DRAM with matched 602 mm^2 dies, reporting roughly 10× the performance and more than 300× the energy efficiency of the tested CPU-plus-DRAM system (14 nm Xeon) on its recommendation-system workload	Large application-specific performance and energy-efficiency gains from a hybrid-bonded near-memory architecture on mature process nodes	2017 S3S: "A 1,000× Improvement…"; flagged the following month in "China May Win in AI Computing," EE Times, Mar 2022
2022-2027	**AMD 3D V-Cache** ships; Dr. Lisa Su, ISSCC 2023: "Tighter Integration of Compute and Memory." **Intel Diamond Rapids** (Hot Chips 2026, shipping 2027): 16 core chiplets on 18A-P hybrid-bonded with Foveros Direct onto four base tiles that hold the entire 1.28 GB last-level cache, following Xeon 6+ Clearwater Forest, the first hybrid-bonded Xeon (Tom's Hardware; ServeTheHome)	Hybrid-bonded memory-on-logic in volume production at both x86 vendors, with the cache moved beneath the cores	Memory-on-logic teachings 2011-2020
2022	imec publishes the BSPDN flow (ECTC 2022, Jourdain et al.): a 50nm epitaxial SiGe25% layer as the etch stop; grind plus selective soft-landing wet etch on the SiGe; sub-40nm total thickness variation at final silicon thickness below 500nm, enabling nano-TSVs to buried power rails	The SiGe cut / etch-stop layer as the enabler of backside power delivery	SiGe cut layer: PCT application filed Sept 2016, published Mar 30, 2017 (Fig. 44). Power from below: application filed Oct 2010 with priority to Oct 2009, Figs. 83L-83L4. Republished: NANO-CHIPS 2030 Ch. 8 §8.3 (2020)
2020-2026	The Chinese ecosystem, from paper to national strategy: IC League **HITOC** paper, "Breaking the Memory Wall for AI Chip with a New Dimension," logic wafer + DRAM wafer, hybrid bonded (Sept 2020); **Xinmeng** 3D $4F^2$ DRAM, **Jasminer X4**, the industry's first DBI hybrid bonding with DRAM (TechInsights), and **Seehi SH9000** tiled DRAM-on-logic (2023); **Dongfang Suanxin** exits stealth (July 2026), led by Wei Shaojun, vice-president of the China Semiconductor Industry Association, with 3D stacked near-memory computing on a fully domestic supply chain (SCMP, July 5, 2026)	Memory-on-logic via hybrid bonding and $4F^2$ 3D DRAM on mature nodes, progressing from research papers to national-champion strategy	2017 S3S paper; NANO-CHIPS 2030 Ch. 15 (2020); "China May Win in AI Computing," EE Times, Mar 2022; "Update," Semiconductor Digest, May 2023; 3D DRAM publications from 2011

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
Dec 2023	MonolithIC 3D forecast in print: 2024 = hybrid-bonding inflection across logic, DRAM, NAND; **Kioxia/WD BiCS8 218L CBA** to volume production; **HBM4 with hybrid bonding**; **Samsung 4F² + hybrid-bonding DRAM at IEDM**; Intel PowerVia extreme-thinning. Same month, **Google, IEDM 2023 SC1.6**: transistor cost scaling stalled at 28 nm	Named, dated predictions across all three segments; the 2014 economic call confirmed from the industry's own podium	"2024 Forecast: Hybrid Bonding Steps Up," Dec 2023; "Intel vs. Intel," EE Times 2014; technical analysis 2016
2023-2026	**Kioxia/SanDisk CBA generation ships and scales**: BiCS8 (218L) to volume; BiCS10 (332L) sample-shipping July 2026, built as separate logic and NAND-cell wafers, bonded; SK hynix 321L; Samsung Cell-on-Periphery	CMOS-directly-Bonded-to-Array = the two-wafer NAND architecture, now the industry's mainstream direction	Two-wafer NAND teachings and patents, 2009-2015; predicted by name Dec 2023
2023-2026	Backside power delivery goes mainstream: Intel commercializes PowerVia (announced 2023, first products on 18A-class nodes); TSMC announces A16 with Super Power Rail; Samsung announces SF2Z with BSPDN. SiGe layers are native to gate-all-around nanosheet nodes, making the cut-layer approach the industry's natural path. An independent teardown of Intel's Panther Lake (SemiAnalysis, Sept 26, 2026) confirms the flow in shipping silicon: the device wafer is bonded to a carrier and flipped, the original substrate is removed until the nano-TSV tips are exposed, and the backside power network is built on the revealed vias.	Extreme-thinning backside architecture as the leading-edge logic standard	The power-from-below (2009) and SiGe cut-layer (2016/2017) teachings; named in the December 2023 forecast
2024-2026	Hybrid bonding on every **HBM** roadmap (HBM4/HBM4E era); all DRAM vendors pursuing **3D DRAM** via hybrid bonding and 4F²/vertical architectures. A January 2026 JEDEC decision raising the HBM package height ceiling from 720 µm to 775 µm let 16-high HBM4 ship on microbumps, deferring hybrid bonding's memory debut to HBM4E and HBM5 (Tom's Hardware, Sept 2, 2026)	Vertical DRAM and bonded stacking as DRAM's future	3D DRAM talks from June 2011; 3D DRAM patent families throughout
Nov 2025	**IEEE TENCON 2025 Best Paper Award**: "Design of Highly Stackable Charge Trap-Based 3D DRAM" (1T poly-Si channel, Schottky-barrier metal-silicide S/D, CXL-class applications)	Ongoing, award-recognized invention in exactly the industry's next battleground	MonolithIC 3D's own current R&D
Mid-2025	The world's largest memory maker takes a **patent license** to the portfolio	Commercial validation by the industry leader	The portfolio itself
June 2026	**Qualcomm announces High Bandwidth Compute (HBC)** at its Investor Day: LPDDR stacked in 3D directly on a custom compute	Memory-on-logic near-memory computing as a flagship compute-vendor	2017 S3S: "A 1,000× Improvement…"; memory-on-logic teachings 2011-

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
	base die with near-memory processing, anchoring its multi-generation datacenter roadmap (AI250 sampling mid-2027) as an alternative to the "HBM tax"	roadmap	2020
July-Sept 2026	**TSMC / Winbond wafer-on-wafer DRAM stacking** (industry reports, unconfirmed): DRAM wafers hybrid-bonded directly with logic wafers to attack the AI memory bottleneck. TSMC's own OIP 2026 material presents DRAM-on-Logic, built with SoIC hybrid bonding, as one of three standard memory options for its advanced nodes, alongside 3D-stacked SRAM and HBM on CoWoS (TSMC, EE Times, Sept 24, 2026). TSMC's OIP projections put 3D DRAM on logic on its 2029 roadmap at less than half the energy per bit of HBM and roughly three times its bandwidth (SemiEngineering; Counterpoint Research, Sept 2026).	Two-wafer logic-plus-memory bonding entering the foundry layer	Two-wafer teachings and patents, 2009-2015; 2017 S3S paper; PCT application with priority to Oct 2016, Figs. 21A-25J
Aug 2026	Samsung introduces **V10 BV-NAND** (Bonding V-NAND) at FMS 2026: an industry-first NAND architecture enabled by a new wafer bonding technology, with more than 400 layers and about 58% higher density than the previous generation	Bonded two-wafer construction as the NAND architecture itself, now named in the product	"Monolithic 3D Memory: NAND Flash, DRAM and R-RAM," 2013; layer transfer and bonded memory teachings, 2010 to 2020
Aug 2026	**Samsung previews zHBM** at FMS 2026: HBM stacked directly above the AI accelerator rather than beside it, with wafer bonding named as the density enabler and customer IP contemplated in the interlayer between memory and accelerator. At Hot Chips 2026 (Aug 23) Samsung set zHBM as the end point of a three-phase HBM roadmap (logic base die, then compute in the base die, then true 3D integration with the interposer removed), built on wafer-on-wafer and hybrid copper bonding with a unified SoC-and-DRAM design flow; Samsung projects about 70% lower I/O power than HBM5 and roughly 2.3 times the DRAM bandwidth with about 100 W less memory power than a four-stack HBM4E system, at about four-high stacks because of heat (Tom's Hardware; ServeTheHome)	Compute and memory in a single vertical stack as the destination architecture, with logic in the bonding interlayer	"A 1,000x Improvement in Computer Systems by Bridging the Processor-Memory Gap," IEEE S3S 2017
Aug-Sept 2026	Hot Chips 2026 (Aug 23-25): Micron calls the memory wall a worsening problem (accelerator compute up about 3× every two years against under 2× for HBM bandwidth; memory silicon about 90% of a GPU package) and names	The processor-memory gap named by the memory makers as AI's binding constraint, with hybrid-bonded memory-on-logic	"A 1,000x Improvement in Computer Systems by Bridging the Processor-Memory Gap," IEEE S3S 2017 (blog, May 2018), Figs.

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
	hybrid and fusion bonding among the remedies; at SEMICON Taiwan (Sept 2) Micron adds that tightly coupled DRAM could deliver more than ten times the bandwidth of HBM at substantially less energy per bit (DigiTimes); SK hynix charts HBM packaging from hybrid bonding to full 3D integration; d-Matrix presents the 3D DRAM stacking behind its Raptor XPU (4 nm compute die face to face on a custom DRAM die; Pavehawk test-chip silicon; targets above 100 TB/s per card). Sept 8: SK hynix's Future Forum sets its 3D direction as wide data buses, ultra-short-reach transmission and DRAM periphery moving to logic-foundry processing. Sept 10: d-Matrix and Nvidia announce Raptor's entry into Nvidia's MGX rack architecture (tape-out expected 2026; availability Q4 2027) Sept 16: at the AI Infra Summit in Santa Clara, SK hynix says the capacity and cost limits of processing-in-memory are solved by hybrid bonding, manufacturing memory dies and processing dies separately and connecting them vertically, with PIM offering about 300 times the capacity of SRAM in the same area, and calls for design across the whole memory hierarchy beyond GPUs plus HBM (Business Korea).	and wide-bus, short-reach 3D as the remedy; memory-and-compute stacking demonstrated in test silicon	9-14: memory and control-logic strata made on separate lines, periphery over the array, wide short buses; PCT application filed Sept 2017 with priority to Oct 2016, Figs. 21A-25J
Aug 2026	Cerebras, Hot Chips 2026: CS-6 roadmap: DRAM 3D-stacked at wafer scale on its wafer-scale engine, with a yield-resilient architecture, vertical power delivery and integrated cooling	Wafer-scale memory-on-logic, with its three enablers: redundancy and repair, power from below, integrated cooling	"Redundancy & Repair," Feb 4, 2011; blog, Dec 15, 2011 ("we could even integrate 1,000 such devices"); 3D-IC Edge, "3D WSI"; IEEE S3S 2017 (Figs. 11, 13, 14); power from below (2009 filing); cooling, IEDM 2012
Aug 2026	Xiaomi unveils the **XRING O100** AI chip (Aug 24, 2026): a 6 nm logic layer, an NPU layer, and two layers of vertical DRAM in one hybrid-bonded 3D stack at 1.4 µm pitch, forming 28,672 vertical data connections at 1.22 TB/s; slated for devices by 2027 (Xiaomi AI Cube shown)	Hybrid-bonded memory-on-logic near-memory compute reaching consumer and edge silicon, designed by a phone maker	"A 1,000x Improvement in Computer Systems by Bridging the Processor-Memory Gap," IEEE S3S 2017; memory-on-logic teachings 2011-2020
Aug 2026	**Fujitsu Monaka**, Hot Chips 2026 (Aug 24, 2026): 144-core Arm server CPU with 2 nm compute dies hybrid-bonded face-to-face onto 5 nm dies holding the entire last-level cache and the per-core voltage regulators; 2 nm	Memory-on-logic with mixed nodes as the answer to stalled cost scaling, and power regulation delivered from the die beneath the	"Intel vs. Intel," EE Times, 2014 (cost scaling); memory-on-logic teachings 2011-2020; power from below (2009 filing)

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
	silicon held under 30% of die area to cut cost; volume production 2027; NEDO-subsidized	transistors	
Aug 2026	The equipment layer commits: **Applied Materials** (Seoul briefing, Aug 31, 2026): data movement, not compute, is the AI bottleneck and 3D scaling is the industry's answer; HBM must move from microbumps to hybrid bonding; advanced-packaging revenue to grow more than 70% in 2026. **Besi** (Q2 2026 results and August 2026 investor presentation): hybrid-bonding orders roughly doubled year on year; hybrid-bonding customers grown from 2 in 2021 to 21 by mid-2026 across logic, memory and photonics; HBM4E named as the insertion point for hybrid bonding in HBM, with the first mover preparing for high-volume manufacturing in 2027 and all leading HBM makers evaluating; hybrid bonding expected to be the largest assembly-equipment segment by 2030; bonder roadmap from 100 nm accuracy today to 25 nm and sub-micron pitch	The equipment industry organizing around hybrid bonding as the mainstream path; the December 2023 forecast of HBM4-generation hybrid bonding now carried on the equipment supplier's own timeline	"Precision Bonders - A Game Changer for Monolithic 3D," IEEE S3S, Oct 2014; "Scaling Makes Monolithic 3D IC Practical," 2013; "2024 Forecast: Hybrid Bonding Steps Up," Dec 2023
Sept 2026	Samsung, SEMICON Taiwan (Sept 1, 2026): **"CUBE"** presented as the governing strategy of its memory business: build memory upward rather than outward and place logic and memory within the stack ("3D is not simply about stacking"); zHBM targeted at 8× HBM4E performance, zNAND-O sampling in 2028	Vertical integration of memory and logic as the organizing principle of the memory industry's leader	"Monolithic 3D-ICs: The Effective Alternative to Dimensional Scaling" (keynote, 2011); "Monolithic 3D DRAM Technology," June 2011; "Monolithic 3D Memory," 2013; "A 1,000× Improvement…," IEEE S3S 2017
May-Oct 2026	**Huawei's "Tau Scaling Law" and LogicFolding**: proposed at IEEE ISCAS (May 25); silicon measurements reported in a preprint (Tingbo He, arXiv 2609.04287, Sept 3); the **Kirin 9050 Pro launched Sept 7** in the Mate XT 2 as the first commercial LogicFolding chip. Huawei reports two logic wafers hybrid-bonded at about 1.5 µm pitch with roughly 50 million vertical connections, on the same node as its predecessor: 238 million transistors per mm^2 against roughly 155 million for the planar predecessor (about 55% higher), power at matched performance lower by 66% for the NPU, 58% for the GPU and 41% for the CPU performance core, and, in a third paper, a chip that runs cooler than its predecessor, the gain attributed to shorter interconnect ("it is the	Logic folding, the single-process case of the thesis: a design partitioned across two bonded strata on an existing node, gaining density, power and performance from fine vertical connectivity; the thermal objection answered on a shipping product; monolithic 3D distinguished from die stacking in the industry's own words	2011 keynote; "Monolithic 3D-ICs with Single Crystal Silicon Layers," IEEE 3DIC 2012 (monolithic 3D vs TSV stacking); IntSim v2.0 (2011); "The Future is the Interconnect," May 2012; "Is the Cost Reduction Associated with Scaling Over?", June 2012 (folding into two strata, IntSim example); "Is TSV for real?", Apr 2012; IEDM 2012; "Precision Bonders," IEEE S3S 2014; "China May Win in AI Computing," EE Times, Mar 4, 2022 (hybrid

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
	travel that burns the energy, not the work"). The figures are Huawei's own; independent benchmarks have since confirmed performance and efficiency behavior, but no independent teardown of the structure, node or foundry has been published. Analysts now draw the distinction the thesis rests on: at coarse pitch a bond joins two finished chips, at fine pitch the two layers are designed as one netlist, "a coarse bond is a stack, a fine bond is a fold" (Convequity, Sept 10; TechTimes, Sept 7). In October Huawei extended LogicFolding to its flagship Mate 90 series, and its consumer chief said more of its phones will use LogicFolding chips (Tech Wire Asia, Oct 2, 2026).		bonding replacing node scaling in China under EUV restriction)
2025-2026	**The RF front end folds: Qorvo QM18566** antenna tuner in the iPhone 17 Pro splits its RF switch network across two active RF-SOI dies bonded wafer-to-wafer with hybrid bonding on **UMC**'s 55 nm 3DIC RF-SOI platform, replacing the planar single-die predecessor with comparable function in a substantially smaller footprint (TechInsights reverse engineering, 2026)	Vertical partitioning of one circuit across two bonded active layers, shipping in consumer volume and paying off at a mature node, independent of lithography	Two-wafer teachings, 2009-2015; design-partitioning filings, 2013, including allocation of RF circuits across strata; 2011 keynote
Sept 2026	Kepler Computing comes out of stealth (Sept 9, 2026) with an AI memory that, in the company's words, leapfrogs leading nodes without EUV in fabs the world already has: memory devices stacked on top of the transistors for density in the z-dimension, 3D between chips to put compute next to memory, the thermal and power problems of 3D addressed, at about a tenth of the capex of a new node; built on ferroelectric materials (as reported). Working silicon in a US fab's production flow, sampling by end of 2026; backed by Intel Capital, GlobalFoundries, AMD Ventures and Applied Ventures, with up to $245M in proposed CHIPS support	The thesis as a company's founding premise: scaling up on existing fabs at existing nodes instead of EUV; memory stacked over logic; ferroelectric memory in a 3D stack	2011 keynote; "Scaling Makes Monolithic 3D IC Practical," 2013; "Intel vs. Intel," 2014; blog, Dec 15, 2011; IEDM 2012; power from below (2009 filing); 3D ferroelectric memory, PCT application with priority to Feb 2017, Figs. 24A-26H; memory over logic by layer transfer with the cut layer, PCT published Mar 2017
Sept 2026	Silicon photonics enters the package: TSMC's COUPE, an electronic die hybrid-bonded (SoIC-X) directly onto a photonic die, enters volume production in 2026 (SEMICON Taiwan, Aug 31 to Sept 2); at the Silicon Photonics Global Summit, fourteen companies from Cisco and TSMC to ASE, imec, Soitec and Lam describe one roadmap, from pluggable optics through CPO and optical I/O to an optical interposer,	The electro-optic stratum of the heterogeneous thesis: photonics made on its own mature-node process and bonded into the stack, with an optical fabric over the package as the declared endgame	2011 keynote; application filed Oct 2011, Figs. 31A-35C: chip-to-chip and on-chip optical interconnect built by layer transfer; PCT application with priority to Sept 2017, Figs. 21A-27B: a generic optical fabric made on its own line and layer-

Year	Adopter / Event	The element validated	Taught by MonolithIC 3D
	with ASE calling the optical engine another chiplet inside the package (as reported, Sept 11)		transferred over the 3D logic-and-memory system
Sept 2026	**PieceMakers**, a Nanya-backed DRAM designer (Nanya holds about 34%), lists on Taiwan's Emerging Stock Board (Sept 16, 2026) at a valuation of roughly $1.4 billion on a business built around DRAM stacked wafer-on-wafer directly on the processor with hybrid bonding: more than 2 TB/s per layer at under 20 ns latency, sold as custom design, IP licensing and royalties, with the first volume program due in 2027 at the earliest. Its president names yield as the biggest hurdle, with the repair architecture designed in and testing before bonding, after bonding and after logic integration; at 80% per-layer yield, four layers would yield about 41% without repair (Tom's Hardware, Sept 16, 2026)	Memory-on-logic 3D DRAM as a standalone, publicly traded IP-and-royalty business; designed-in redundancy and repair named as the enabler of wafer-on-wafer stacks; a fourth DRAM maker choosing custom stacked memory over HBM	"Redundancy & Repair" presentation, Feb 2011; "Monolithic 3D DRAM Technology," AVS, June 2011; IEEE S3S 2017 (memory strata as repairable unit arrays, Figs. 11-14); NANO-CHIPS 2030 Ch. 15 (2020); redundancy, repair and built-in test patent families
Sept 2026	**Samsung** reportedly makes wafer-to-wafer hybrid bonding the baseline of its first sub-10 nm DRAM: B1b, previously an R&D project, moves onto the official roadmap as **D0a**, with memory cells and peripheral circuits made on separate wafers, each on the process that suits it, and joined by copper hybrid bonding; the vertical-channel-transistor design becomes a derivative (D0a-V). Commercialization is expected in roughly two years, with the production investment decision by mid-2027, and Samsung is evaluating the approach for HBM5 (ZDNet Korea, Sept 23; DigiTimes, Sept 24). Samsung's VCT flow itself relies on wafer bonding, burying the bitline by flipping the cell wafer onto a support wafer (SemiAnalysis)	The two-wafer architecture, cell array and periphery made separately and bonded, becoming DRAM's mainstream path as CBA became NAND's; bonding at the base of both sub-10 nm DRAM options	"Monolithic 3D DRAM Technology," AVS, June 2011; two-wafer memory teachings and patent families, 2009-2015; PCT application with priority to Oct 2016, Figs. 21A-25J: multiple DRAM strata sharing one control stratum, connected at the word-line and bit-line level by hybrid bonding; "Precision Bonders - A Game Changer for Monolithic 3D," IEEE S3S, Oct 2014; forecast of hybrid-bonded $4F^2$ DRAM, Dec 2023

The pattern is uniform: **publication → skepticism → third-party demonstration → industry adoption → commercial mainstream.** Sixteen years, one thesis, no reversals. Between May and September 2026, the convergence extended beyond the memory makers to the compute layer (Qualcomm), the foundry layer (TSMC), Chinese national doctrine (Huawei, Dongfang Suanxin), consumer silicon (Xiaomi), and the equipment suppliers (Applied Materials, Besi): every layer of the industry is now converging on memory-on-logic 3D integration. On September 1, Samsung presented vertical integration as the governing strategy of its memory business. At Hot Chips 2026 the memory makers themselves named the processor-memory gap as AI's binding constraint and hybrid-bonded 3D integration as the way through it, and Cerebras announced DRAM stacked at wafer scale on its wafer-scale engine. Independent

analysis of the conference reached the same conclusion under the title “The Memory Wall Is Making Hybrid Bonding a Necessity.”

In the first ten days of September 2026, Samsung announced CUBE, Huawei published measured silicon for logic folding, SK hynix described its future as wide-bus, short-reach 3D memory with the DRAM periphery turned into a logic layer, d-Matrix announced that its memory-on-compute accelerator will enter Nvidia’s rack-scale architecture, and Kepler Computing came out of stealth with an AI memory built on the premise that 3D and new materials, not EUV, are the way to leadership chips: companies spanning memory makers, mobile logic and AI compute, each describing scaling up as the path forward. On September 16 the capital markets joined: PieceMakers, a Nanya-backed designer of DRAM stacked directly on the processor, listed in Taipei at a valuation of roughly $1.4 billion as an IP-and-royalty business, with its president naming designed-in repair as the key to wafer-on-wafer yield. And on September 23 Samsung was reported to have made wafer-to-wafer bonding of a cell-array wafer onto a separately made peripheral wafer the baseline of its first sub-10 nm DRAM, taking DRAM down the two-wafer path that 3D NAND took; and folding has reached the radio of a mass-market phone.

Why the Convergence Came in 2026

For four decades the semiconductor industry was organized around cost: each node made the transistor cheaper, and consumer products absorbed the volume. Two things changed that. Transistor cost scaling stalled at 28 nm (Section IV), and, after the arrival of large language models in late 2022, AI became the industry’s largest market and its source of profit, buying performance and power efficiency rather than cost. With the memory makers, foundries and equipment suppliers now racing to supply that demand, the limit is no longer the transistor but the wires between transistors, dies and packages: data movement, not arithmetic, sets the power and the performance. Dimensional scaling shortens nothing at the system level. Vertical integration does, which is why logic folding, memory-on-logic, bonded NAND and DRAM, and now optical strata are all arriving at once, from every layer of the industry. Independent analysis published the same week reached the same conclusion: the next leap is memory stacked on logic, not smaller transistors (Asia Times, Sept 11, 2026). In 2022 a single research lab demonstrated DRAM hybrid-bonded to logic, and the industry largely passed over it; in 2026 the world’s largest foundry presented DRAM-on-logic as one of three standard ways to pair its advanced logic with memory, and placed it on its 2029 roadmap.

A Note on Engagement

This record is not a complaint that the industry was told and did not listen. Throughout these years MonolithIC 3D did what an inventor-founded company can do to put its work in the hands of those best placed to use it. It published the process flows, cost models and architectures openly rather than holding them back. It released IntSim as free open-source software so that anyone could test the scaling-up case on their own designs. And it went to the industry directly, year after year: presenting at Semicon West, meeting with semiconductor manufacturers, memory makers and equipment vendors, visiting research institutes such as imec and CEA-Leti, and pitching the 3D advantages and the path to get there, with the technology and the patent portfolio offered for license.

Large companies move on their own roadmaps and for their own reasons, and the industry reached the scaling-up path in its own time and by its own routes. That is how the semiconductor industry works. The record above simply documents the sequence: what was published, when it was published, and what followed. The inventions it describes remain available for license.

VII. The Patent Estate Beneath the Publications

The public teaching and the patent record are the same body of work, running in parallel:

- **402 issued US patents (the most recent issued May 12, 2026), with many more applications pending: an unbroken grant cadence from June 2011 to the present, averaging more than two dozen grants per year across the last five years.**
- **Priority dates of 2009-2010** on the foundational families, predating YMTC Xtacking by ~9 years, Kioxia CBA by ~14 years, and hybrid-bonded HBM by ~15 years.
- **Coverage tracking the published thesis element by element**: layer transfer and bonded multilayer structures; oxide and hybrid bonding; two-wafer memory-plus-logic architectures; 3D NAND structures and methods; 3D DRAM devices and control circuits; memory-on-logic and processor-plus-memory systems; power-delivery grids and connective paths through bonded stacks; single-crystal thin layers; redundancy, repair, and built-in test for ultra-large 3D integration; design automation for 3D ICs; and 3D image sensors and micro-displays.
- **The pending pipeline** of applications means the estate is still growing at the same cadence at which the industry's 3D adoption is accelerating.
- **The founder's personal record**: Zvi Or-Bach is a named inventor on more than 550 issued US patents (as of July 2026, with new patents issuing weekly), spanning monolithic 3D integration and three decades of semi-custom architecture invention: Chip Express (1989), eASIC (1999, acquired by Intel in 2018), and MonolithIC 3D (2009-present).

VIII. What This Record Is Not

It is not an acquired portfolio: every patent originates from the company's own inventors. It is not hindsight claiming: every prediction cited above was published, dated, and in most cases doubted before the industry's adoption. And it is not a passive archive: the 2025 IEEE TENCON Best Paper and the 2026 grant cadence demonstrate an invention program that is active in exactly the field, highly stackable 3D DRAM, where the industry's next decade will be decided.

IX. Uses of This Record

- **Website / Innovation Heritage page.** This document is the master narrative; each timeline entry links to the underlying paper, slide deck, blog post, or patent already hosted on monolithic3d.com.
- **Press briefings (Korea, Japan, US).** The one-line story: the companies now practicing scaling-up are following a map this company published, dated, and patented, starting in 2010.
- **Licensing discussions.** The record frames the license as the industry-standard mechanism for practicing a documented, foundational body of invention, as the market leader has already acknowledged.

X. Source Register (abridged)

All items publicly hosted at monolithic3d.com or the cited venues; full URLs on the site's Papers, Presentations and Patents page and blog archive (March 2011 - present).

1. "3D-FPGA: The Path to ASIC-like Density, Power and Performance," Applied Materials 3D Interconnect Symposium, Jan 2010.
2. "Monolithic 3D Technology: A Semiconductor Industry Game Changer," 3D-ASIP, Dec 2010.

3. “Monolithic 3D Integrated Circuits,” Future-Fab International #37, Apr 2011.

4. “Monolithic 3D DRAM Technology,” AVS 3D Packaging Workshop, Jun 2011.

5. Keynote: “Monolithic 3D-ICs: The Effective Alternative to Dimensional Scaling,” Intl. Electronics Forum, Seville, Oct 2011.

6. “Monolithic 3D-ICs with Single Crystal Silicon Layers,” IEEE 3DIC (invited), 2012.

7. “Cooling Three-Dimensional Integrated Circuits Using Power Delivery Networks,” IEDM, Dec 2012.

8. “Practical Process Flows for Monolithic 3D,” IEEE S3S Plenary, Oct 2013.

9. “Scaling Makes Monolithic 3D IC Practical,” Solid State Technology, Oct 2013.

10. “Intel vs. Intel” (Moore’s Law stopped at 28nm), EE Times, 2014; technical analysis, 2016.

11. “The Rise of the Monolithic 3D Chip,” IEEE Spectrum, Feb 2014.

12. “Precision Bonders - A Game Changer for Monolithic 3D,” IEEE S3S, Oct 2014.

13. “Monolithic 3D Integration,” CHIPS 2020 Vol. 2 (Springer), pp. 51-91, Oct 2015.

14. “Modified ELTRAN® - A Game Changer for Monolithic 3D,” Oct 2015.

15. “A 1,000× Improvement in Computer Systems by Bridging the Processor-Memory Gap,” IEEE S3S, 2017 (blog, May 23, 2018).

16. “DARPA Calls for Monolithic 3D - 3DSoC,” Z. Or-Bach, Sept 19, 2017 (re: DARPA ERI BAA, Sept 13, 2017; Al Fazio plenary, IEEE S3S 2017).

17. US patent application filed Oct 2010, priority chain to Oct 2009, issued Mar 2013, Figs. 83L-83L4: power delivery from below the device layer.

18. PCT application filed Sept 21, 2016; published Mar 30, 2017 (Fig. 44, layer 4404): SiGe cut / etch-stop layer for precision extreme wafer thinning.

19. A. Jourdain et al. (imec), “Buried Power Rails and Nano-Scale TSV: Technology Boosters for Backside Power Delivery Network and 3D Heterogeneous Integration,” IEEE ECTC 2022, pp. 1531-1537.

20. NANO-CHIPS 2030 (Springer, 2020): Ch. 15 “A 1000× Improvement of the Processor-Memory Gap”; Ch. 11 “3D for Efficient FPGA.”

21. “China May Win in AI Computing: Leveraging Hybrid Bonding as an Alternative to Dimensional Scaling,” EE Times, Mar 4, 2022 (Hebrew versions: Hayadan, Chiportal, Mar 2022).

22. Alibaba, “184QPS/W 64Mb/mm² 3D Logic-to-DRAM Hybrid Bonding…,” ISSCC 2022 (digest copy; 9.78× speedup and 317× energy efficiency vs. an Intel Xeon Gold 5220 CPU-DRAM system on a recommendation-system workload).

23. “Update: China May Win in AI Computing,” Semiconductor Digest, May 19, 2023.

24. “2024 Forecast: Hybrid Bonding Steps Up,” Dec 4, 2023.

25. Milind Shah (Google), IEDM 2023 Short Course SC1.6 (28nm cost-scaling confirmation); blog, “Moore’s Law Indeed Stopped at 28nm”.

26. “Design of Highly Stackable Charge Trap-Based 3D DRAM,” IEEE TENCON 2025 (Best Paper Award).

27. Qualcomm Investor Day, June 2026: High Bandwidth Compute (HBC) announcement; Dragonfly AI250/AI300 roadmap.

28. “Chinese AI chip start-up bets on 3D stacking to bypass US controls,” South China Morning Post, July 5, 2026 (Dongfang Suanxin; Huawei “Tau Scaling Law,” May 2026).

29. Industry reports, July 2026 (unconfirmed): TSMC/Winbond wafer-on-wafer DRAM-on-logic hybrid bonding (SemiWiki and financial press).

30. Samsung Semiconductor, “Samsung Unveils Next-Gen 3D Memory Vision at FMS 2026,” August 2026: V10 BV-NAND (Bonding V-NAND) and zHBM (archived copy retained).

31. Hot Chips 2026 (Stanford, Aug 23-25, 2026), as reported: Tom’s Hardware, “Micron warns HBM wafer penalty is widening with every generation” (Micron: compute ~3× per two years vs. HBM bandwidth <2×; memory wall “getting worse”); Tom’s Hardware, “SK hynix pushes hybrid bonding to HBM5 as AI memory hits 775-micron ceiling” and DigiTimes, “SK Hynix charts advanced HBM packaging path: hybrid bonding, Intel EMIB and 3D integration”;

Tom's Hardware and ServeTheHome, "d-Matrix stacks AI accelerator directly on custom DRAM for 100 TB/s per card" (Raptor: TSMC 4 nm logic face-to-face on custom DRAM, 32 GB, >100 TB/s, 0.37 pJ/bit).

32. "Redundancy & Repair with Monolithic 3D," presentation, Feb 4, 2011 (wafer-scale integration without yield issues via stacked redundancy and repair layers).
33. "Monolithic 3D IC Could Increase Circuit Integration by 1,000x," MonolithIC 3D blog, Dec 15, 2011.
34. "The Monolithic 3D Advantage," 3D-IC Edge chapter (c. 2013-2014): Sec. 6, logic redundancy allowing 100× integration; "3D WSI" ~10× at board, rack, and server-farm level, ~1,000× overall; Sec. 4, memory 1 µm above logic.
35. Cerebras, "Ultrafast Frontier Inference: Cerebras Deep Dive at Hot Chips 2026," cerebras.ai blog, Aug 2026 (CS-6: DRAM stacked at wafer scale on the wafer-scale engine); Tom's Hardware, "Hot Chips 2026: Cerebras lays out the future of wafer-scale AI… CS-6 wafer to incorporate stacked DRAM," Aug 27, 2026; ServeTheHome, "Cerebras Talks Going Rack-Scale with Their WSEs at Hot Chips 2026," Aug 2026.
36. US patents issued June 2011 through May 2026: 402 issued; many more pending. Full list: monolithic3d.com.
37. Hot Chips 2026, Micron (Raghu Sreeramaneni), as reported: BigGo Finance, "Micron: AI Memory Bottleneck Worsening - Compute Triples Every Two Years While HBM Bandwidth Grows Less Than 2x," Aug 24, 2026 (memory silicon ~90% of a four-stack 16-high GPU SiP; 17% of Llama 3 training interruptions HBM-related; hybrid and fusion bonding named).
38. Xiaomi XRING O100 announcement, Aug 24, 2026, as reported: Mezha, "Xiaomi has unveiled the XRING O3 processor and showcased the AI Cube computer based on it"; VideoCardz infographic translation; EE Times China (first 6 nm wafer-on-wafer 3D implementation, 1.4 µm hybrid-bonding pitch).
39. Fujitsu Monaka, Hot Chips 2026 (Aug 24, 2026), as reported: Tom's Hardware, "Fujitsu's Monaka CPU stacks its entire cache on a separate 5nm die and narrows to 256-bit SVE2"; Chips and Cheese, "Hot Chips 2026: Fujitsu's Monaka CPU"; ServeTheHome, "Fujitsu's Arm-based Monaka Data Center CPU at Hot Chips 2026" (2 nm core die hybrid-bonded on 5 nm SRAM die with last-level cache and LDOs; 2 nm under 30% of area; volume 2027).
40. Applied Materials media briefing, Seoul, Aug 31, 2026, as reported: The Elec, "Applied Materials Touts 3D Scaling Technologies, Develops New Materials for HBM and CPO"; Digital Today, "AMAT says AI chip performance hinges on data movement, solution is 3D scaling"; Korea Times, Aug 31, 2026; ChosunBiz, "Applied tackles AI memory wall with 3D scaling, HBM bonding, CPO in Korea." Besi Q2 2026 results as reported: Seeking Alpha, "BE Semiconductor: Hybrid Bonding Is Becoming AI's Next Bottleneck," Aug 31, 2026 (orders roughly doubled year on year).
41. SemiVision Research, "Hot Chips 2026: The Memory Wall Is Making Hybrid Bonding a Necessity," Aug 27, 2026. https://tspasemiconductor.substack.com/p/hot-chips-2026-the-memory-wall-is
42. Samsung CUBE strategy, SEMICON Taiwan Memory Executive Summit, Sept 1, 2026, as reported: The Korea Times, "Samsung showcases chip 3D architecture strategy at SEMICON Taiwan," Sept 1, 2026; The Elec, "Samsung Targets Memory Limits With 3D-Stacked HBM," Sept 2, 2026; Seoul Economic Daily, "Samsung Unveils 'CUBE' Strategy to Break Memory Limits," Sept 1, 2026 (Omdia capacity figures); Korea JoongAng Daily, Sept 2, 2026; Global Sources Sourcing Digest, "Samsung's CUBE strategy aims to redefine AI memory with 3D integration and giant production capacity."
43. Huawei "Tau Scaling Law" / LogicFolding: IEEE ISCAS, Shanghai, May 25, 2026 (as reported: Tom's Hardware, May 25, 2026); second paper with engineering detail and measured data, July 3, 2026 (as reported: Pandaily; Geopolitechs, July 4-6, 2026); Tingbo He, "Huawei's τ Chip Was Supposed to Melt?", arXiv 2609.04287, submitted Sept 3, 2026 (ChinaXiv DOI 10.12074/202609.00031); South China Morning Post, Sept 6-7, 2026; TrendForce, Sept 7, 2026.
44. SK hynix Future Forum 2026, Sept 8, 2026: SK hynix Newsroom recap (3D direction as device integration and heterogeneous integration; data bus width, ultra-short-reach transmission, DRAM periphery as logic foundry; boundary between foundry and memory IDM).
45. d-Matrix, "d-Matrix Adopts NVIDIA NVLink Fusion Rackscale Infrastructure," Sept 10, 2026 (Raptor tape-out expected before end of 2026; MGX availability Q4 2027); NVIDIA Blog, "d-Matrix Adopts NVIDIA NVLink Fusion for Rack-Scale XPU Deployment"; d-Matrix, "Going Vertical: Why we created a 3D DRAM solution," Mar 16, 2026 (Pavehawk 3D DRAM test chip); The Next Platform, "Startup d-Matrix will pair its Raptor memory-based XPU to Nvidia rackscale iron," Sept 10, 2026; Blocks and Files, Sept 11, 2026.

46. Kepler Computing launch, Sept 9, 2026: keplercompute.com and "Act I: AI's New Memory - Beyond HBM & SRAM, Without EUV"; U.S. Department of Commerce letter of intent, July 2026 (up to $245M); press coverage Sept 10-11, 2026.

47. TechInsights, Apple M5 Pro (APL1X15) package analysis, June 30, 2026: CPU and GPU chiplets on a silicon interposer joined by TSMC SoIC-X face-to-face chip-on-wafer hybrid bonding; described as a 2.5D CoWoS-S-class architecture with bumps replaced by hybrid bonding.

48. "The CPO Debate Is Over: Silicon Photonics Enters the Manufacturing Race," TSPA Semiconductor (Substack), Sept 11, 2026 (SEMICON Taiwan 2026 Silicon Photonics Global Summit, fourteen presentations; roadmap from pluggable optics to optical interposer; TSMC COUPE roadmap; ASE on the optical engine as a chiplet). TSMC COUPE in volume production 2026: TechTimes, Aug 31, 2026; TrendForce, Apr 1, 2026 (COUPE stacks the electrical die on the photonic die by SoIC-X).

49. US patent application filed Oct 2011, "Semiconductor and Optoelectronic Devices" (filed Oct 14, 2011; issued Nov 24, 2015), Figs. 31A-35C: LED-driven and laser-driven chip-to-chip optical interconnect (31, 32) and on-chip optical interconnect (33, 35) constructed by monolithic stacking using layer transfer. PCT application with priority Sept 24, 2017, Figs. 21A-27B, paras. 133-189: heterogeneous integration of an optical interconnect level made on a different process and transferred over the logic (para. 135); optical structure prepared with a cut layer and hybrid-bonded with vias of about 100 to 400 nm (Figs. 24A-24B); generic, pre-tested optical fabric crossing dice lines and reticles up to wafer scale, cut to system size (Figs. 22A-23E, 25A-25C).

50. Nilesh Jasani and Vibhu Sharma, "The chip industry's future is memory by design," Asia Times, Sept 11, 2026.

51. Shane Downing, "Piecemakers bets edge AI devices will diverge from reliance on HBM," Tom's Hardware, Sept 16, 2026. https://www.tomshardware.com/tech-industry/semiconductors/piecemakers-bets-edge-ai-devices-will-diverge-from-reliance-on-hbm-custom-designed-memory-fuses-dram-stack-directly-to-the-processor-using-hybrid-bonding

52. BE Semiconductor Industries N.V. (Besi), Investor Presentation, August 2026 (hybrid-bonding customer adoption, HBM4E insertion point, 2030 installed-base and segment estimates, bonder roadmap). https://www.besi.com/investor-relations/

53. TechTimes, "Huawei Kirin 9050 Pro Launches: LogicFolding Moves From Roadmap to Silicon," Sept 7, 2026; Tech Wire Asia, "Kirin 9050 Pro proves Huawei's workaround," Sept 2026; Convequity, "Huawei After Sanctions: CIDM, Kirin 9050, and Logic Folding," Sept 10, 2026. https://www.techtimes.com/articles/326836/20260907/huawei-kirin-9050-pro-launches-logicfolding-moves-roadmap-silicon.htm

54. DigiTimes, "Micron says tightly coupled DRAM could deliver more than 10x HBM bandwidth," Sept 2, 2026. https://www.digitimes.com/news/a20260902VL223/micron-dram-hbm-bandwidth-nand.html

55. Tom's Hardware, "The current state of hybrid bonding in 2026: TSMC sits at 6 microns and the HBM delay that nobody expected," Sept 2, 2026. https://www.tomshardware.com/tech-industry/semiconductors/hybrid-bonding-roadmap-examined

56. Business Korea, "SK hynix Unveils Processing-In-Memory (PIM) for AI Bottlenecks," Sept 17, 2026 (Lim Eui-cheol, AI Infra Summit, Santa Clara, Sept 16). https://www.businesskorea.co.kr/news/articleView.html?idxno=277234

57. NeoTeo, "Huawei says the Kirin 9050 Pro uses no US components, but its supply chain remains undisclosed," Sept 17, 2026 (density and matched-performance power figures; independent benchmark coverage). https://www.neoteo.com/en/lp/huawei-says-the-kirin-9050-pro-uses-no-us-components-but-its-supply-chain-remains-undisclosed

58. Tom's Hardware, "Hot Chips 2026: Samsung reveals a three-phase HBM roadmap that puts logic and compute inside memory; zHBM ultimately stacks DRAM directly on top of the processor," Aug 2026; ServeTheHome, "Samsung Evolving HBM Base Die at Hot Chips 2026." https://www.tomshardware.com/tech-industry/semiconductors/hot-chips-2026-samsung-reveals-

a-three-phase-hbm-roadmap-that-puts-logic-and-compute-inside-memory-zhbm-ultimately-stacks-dram-directly-on-top-of-the-processor

59. Tom's Hardware, "Hot Chips 2026: Intel Xeon 7 'Diamond Rapids' comes with up to 256 P-cores, 1.28 GB of last-level cache," Aug 2026; ServeTheHome, "Intel Diamond Rapids, the 2027 Intel Xeon, at Hot Chips 2026." https://www.tomshardware.com/pc-components/cpus/intel-xeon-7-diamond-rapids-comes-with-up-to-256-p-cores-1-28-gb-of-last-level-cache-next-gen-18a-p-cpu-also-brings-avx-10-2-and-uses-ucie-s-instead-of-emib
60. PCT application filed Sept 19, 2017 with priority to Oct 10, 2016, Figs. 21A-25J: control-line arrangement for memory integration, stratum select, partition of an array of memory units, generic 3D memory structure, active thermal cooling; issued US continuations.
61. ZDNet Korea, report on Samsung's D0a DRAM roadmap revision, Sept 23, 2026; DigiTimes, "Samsung reportedly shifts sub-10nm DRAM roadmap toward hybrid bonding," Sept 24, 2026. https://www.digitimes.com/news/a20260924VL206/samsung-dram-roadmap-technology-development.html
62. SemiAnalysis, "The Memory Wall: Past, Present, and Future of DRAM," Mar 24, 2026. https://newsletter.semianalysis.com/p/the-memory-wall
63. SemiAnalysis, "Intel Panther Lake Teardown," Sept 26, 2026. https://newsletter.semianalysis.com/p/intel-panther-lake-teardown
64. Aveek Sarkar (TSMC), "Inside TSMC's Evolving Design Ecosystem: Shaping the Future of AI, with AI," EE Times, Sept 24, 2026 (figure: "TSMC's advanced nodes are enabling workload-specific architectures": 3D-stacked SRAM on SoIC, HBM on CoWoS, DRAM-on-Logic on SoIC).
65. TechInsights, "How Qorvo's 3D RF-SOI architecture is redefining smartphone RF front-end design" (Qorvo QM18566, UMC 55 nm 3DIC RF-SOI, Apple iPhone 17 Pro), 2026. https://www.techinsights.com/report/qorvo-3d-rf-soi-rf-front-end-analysis
66. SemiEngineering, "TSMC OIP: Chip Industry Growth Blows Past Forecast," Sept 2026 (TSMC 2029 forecasts: 3D DRAM about 1 pJ/bit vs HBM about 2.5 pJ/bit; five-year bandwidth growth 3D DRAM 115x vs HBM 34x). https://semiengineering.com/tsmc-oip-chip-industry-growth-blows-past-forecast/
67. Counterpoint Research, "$Trillion Bottleneck: TSMC OIP 2026 - AI Design Kit, Packaging Innovations, 3D DRAM Timeline," Sept 2026 (DRAM-on-logic on TSMC's 2029 roadmap at about 3x HBM5E bandwidth, under 1 pJ/bit).
68. Tech Wire Asia, "Huawei expands LogicFolding chips as China faces advanced manufacturing limits," Oct 2, 2026.