Semiconductor Technology Roadmap

2nm, HBM4, Advanced Packaging Milestones

Process Technology Evolution

TSMC Roadmap:
- 3nm (N3): Mass production Q4 2022, Apple A17 Pro first customer
- 3nm Enhanced (N3E): Q4 2023, improved yield and cost
- 2nm (N2): Volume production H2 2025, GAA transistors
- 2nm Plus (N2P): 2026, 10-15% performance improvement
- 1.4nm (A14): 2027 target, next-generation GAA

Samsung Foundry Roadmap:
- 3nm GAA (3GAP): Q4 2022 launch, yield challenges
- 2nm (2GAP): 2025 target, competing with TSMC N2
- 1.4nm: 2027 target

Note: Samsung has faced persistent yield issues with GAA transition, losing market share to TSMC.

Intel Process Roadmap:
- Intel 4 (7nm): Meteor Lake launched Q4 2023
- Intel 3 (enhanced 7nm): 2024, 18% performance gain
- Intel 20A (2nm equivalent): H1 2024, first Intel GAA
- Intel 18A (1.8nm): 2025, targeting external customers
- Intel 14A: 2026-2027

Critical Technology Transitions:
GAA (Gate-All-Around) transistors replace FinFET at 2nm node. This enables continued transistor density scaling but requires entirely new manufacturing processes. Development cost: $30B+ per company.

HBM (High Bandwidth Memory) Evolution

HBM3 (Current Generation):
- Bandwidth: 819GB/s per stack
- Capacity: up to 24GB per stack
- Power efficiency: 3.2 pJ/bit
- Production: SK Hynix (50% share), Samsung (30%), Micron (20%)

HBM3E (Enhanced, 2024):
- Bandwidth: 1.15TB/s per stack (40% increase)
- Capacity: up to 36GB per stack

- Power efficiency: 2.8 pJ/bit
- Mass production: SK Hynix Q3 2024, Samsung Q4 2024
- Primary customer: Nvidia H200, AMD MI325X

HBM4 (2026 Target):
- Bandwidth target: 2TB/s per stack
- Capacity: up to 48GB per stack
- New features: Error correction, higher stack height (16-hi)
- Technology: Through-Silicon Via (TSV) improvements
- Development status: SK Hynix and Nvidia co-developing

Supply Chain Bottleneck:
HBM production requires specialized equipment and has 12-month qualification cycles. Current shortage limits AI accelerator production. SK Hynix HBM capacity fully allocated through 2025. New fabs coming online in 2026 will ease constraints.

Advanced Packaging Technologies

CoWoS (Chip-on-Wafer-on-Substrate) - TSMC:
- Current: CoWoS-S (interposer-based), used in Nvidia H100
- 2024: CoWoS-L (LSI interposer), 3x size increase
- 2025: CoWoS-R (RDL-based), cost reduction
- Capacity: 15K wafers/month (2024) → 30K wafers/month (2025)

Bottleneck: CoWoS capacity limits H100/H200 production

InFO (Integrated Fan-Out) - TSMC:
- Used for mobile processors (Apple A-series)
- Lower cost than CoWoS, less suited for HBM integration
- 2025: InFO_oS (on Substrate) for improved power delivery

Foveros - Intel:
- 3D stacking of active logic dies
- Used in Meteor Lake (compute + graphics stacking)
- 2025: Foveros Direct (10μm bump pitch)
- Target: compete with TSMC for external customers

X-Cube - Samsung:
- Hybrid bonding technology for 3D integration
- Currently in development, lagging TSMC/Intel
- Target production: 2025-2026

EUV Lithography Evolution

Current EUV (0.33 NA):
- Minimum pitch: 13nm

- Used for: 7nm through 2nm processes
- ASML production: 60 systems/year capacity

High-NA EUV (0.55 NA):
- Minimum pitch: 8nm (enables sub-2nm processes)
- First system delivered to Intel in December 2023
- Cost: $380M per system (vs $200M for standard EUV)
- ASML production: 10-20 systems/year initially
- Customer roadmap: Intel (2024), TSMC (2025), Samsung (2026)

Technology Challenges:
High-NA systems require complete process redesign. Wafer size changes from 300mm to effectively 200mm throughput, increasing costs. Multi-year learning curve expected.

AI Accelerator Roadmap

Nvidia GPU Evolution:
- 2023: H100 (Hopper architecture, 4nm)
- 2024: H200 (HBM3E upgrade, 141GB vs 80GB)
- 2024: B100/B200 (Blackwell architecture, 4nm)
- 2025: B200 Ultra (3nm process shrink)
- 2026: Next-gen (codename Rubin, 3nm+ or 2nm)

AMD GPU Competition:
- 2023: MI300A/MI300X (chiplet architecture, 5nm+6nm)
- 2024: MI325X (HBM3E, 256GB)
- 2025: MI350 (3nm target, competing with B100)

Software ecosystem gap remains AMD's challenge vs CUDA.

Custom Silicon Trend:
- AWS Trainium2 (2024): competing with Nvidia for training workloads
- Google TPU v5 (2024): 4x performance vs v4
- Microsoft Maia (2024): Azure OpenAI infrastructure
- Meta MTIA v2 (2025): inference optimization

Custom chips now represent 20% of AI accelerator market, growing to 35% by 2027.