

Halbleitertechnologie von A bis Z

Advanced Packaging / 3D-Integration

Grenzen der klassischen Integration	3
<i>Warum „More Moore“ an Grenzen stößt</i>	3
<i>„More than Moore“ als Alternative</i>	3
<i>Einordnung zu klassischem Wire-Bonding und Flip-Chip</i>	3
Through-Silicon Via (TSV)	4
<i>Das Prinzip</i>	4
<i>Herstellungsprozess</i>	4
<i>Isolation und Cu-Füllung</i>	5
<i>Herausforderungen beim Aspektverhältnis</i>	5
2.5D-Integration: Der Interposer	6
<i>Das Grundprinzip</i>	6
<i>Silizium- vs. organischer Interposer</i>	6
<i>Redistribution Layer (RDL)</i>	6
<i>Beispiel: HBM auf Interposer</i>	6
3D-Stacking	7
<i>Die-to-Die vs. Wafer-to-Wafer Bonding</i>	7
<i>Hybrid Bonding: Cu-Cu direkt</i>	8
<i>Thermische und mechanische Herausforderungen</i>	8
Chiplets & heterogene Integration	9
<i>Disaggregation von SoCs</i>	9
<i>Vorteile: Yield und Kosten</i>	9
<i>Nachteile: Interconnect-Overhead</i>	10
<i>Standardisierung: UCIe</i>	10
Anwendungsfelder & Ausblick	10
<i>HBM als Paradebeispiel</i>	10
<i>KI-Beschleuniger</i>	11
<i>Foundry-Vergleich</i>	11
<i>Ausblick</i>	11

Grenzen der klassischen Integration

Warum „More Moore“ an Grenzen stößt

Über Jahrzehnte folgte die Halbleiterindustrie dem Mooreschen Gesetz: Alle 18-24 Monate verdoppelte sich die Anzahl der Transistoren pro Chip, primär durch Verkleinerung der Strukturbreiten. Mit Strukturgrößen im einstelligen Nanometerbereich stößt diese Skalierung jedoch an physikalische und ökonomische Grenzen. Leckströme steigen durch Kurzkanaleffekte, die Lithografiekosten pro Belichtungsschritt explodieren (EUV-Anlagen kosten teils über 200 Millionen Euro), und die Ausbeute (Yield) sinkt bei großflächigen monolithischen Chips überproportional, da die Wahrscheinlichkeit eines Defekts mit der Chipfläche wächst.

Gleichzeitig verlangt die Nachfrage nach Rechenleistung – getrieben etwa durch KI-Beschleuniger – nach immer mehr Transistoren pro System. Die Antwort der Industrie liegt zunehmend nicht mehr in kleineren Transistoren, sondern in einer intelligenteren Anordnung und Verbindung mehrerer Chips: Advanced Packaging.

„More than Moore“ als Alternative

Anstatt sämtliche Funktionen eines Systems auf einem einzigen monolithischen Die zu integrieren, werden einzelne Funktionsblöcke – Prozessorkerne, Speicher, I/O, analoge Schaltungen – auf separaten, oft in unterschiedlichen Technologien gefertigten Chips realisiert und anschließend eng gekoppelt zu einem Gesamtsystem zusammengeführt. Man spricht von heterogener Integration. Diese Strategie erlaubt es, jeden Baustein in der für ihn optimalen und kostengünstigsten Technologie zu fertigen (z. B. Logik in 3 nm, Speicher in einem älteren, günstigeren Knoten), anstatt Kompromisse für einen einzigen Fertigungsprozess einzugehen.

Die Herausforderung verlagert sich damit vom Transistor auf die Verbindungstechnik zwischen den Chips – von der Frage „Wie klein kann ein Transistor werden?“ zur Frage „Wie dicht und wie schnell lassen sich zwei Dies miteinander verbinden?“

Einordnung zu klassischem Wire-Bonding und Flip-Chip

Die in den Kapiteln zur Montage behandelten Verfahren – Wire-Bonding und Flip-Chip – verbinden jeweils einen einzelnen Die mit einem Gehäuse oder Substrat. Die elektrischen Verbindungen liegen dabei am Rand oder auf der

Chipoberfläche und sind in ihrer Dichte begrenzt: Wire-Bonds erreichen typischerweise einen Pitch (Abstand zwischen benachbarten Verbindungen) von 40–50 µm, Flip-Chip-Bumps liegen im Bereich von 100–150 µm.

Advanced-Packaging-Verfahren wie TSV, Interposer und Hybrid Bonding (Gegenstand der folgenden Kapitel) erreichen Verbindungsdichten, die um Größenordnungen darüber liegen – bis in den Sub-Mikrometer-Bereich. Erst dadurch wird es möglich, mehrere Dies so eng zu koppeln, dass sie sich elektrisch nahezu wie ein einziger, monolithischer Chip verhalten.

Through-Silicon Via (TSV)

Das Prinzip

Ein Through-Silicon Via ist eine vertikale, elektrisch leitfähige Verbindung, die komplett durch das Silicium-Substrat eines Dies hindurchgeführt wird. Anders als klassische Verdrahtungsebenen, die ausschließlich auf der Chipoberfläche liegen, erschließt das TSV die dritte Dimension: Signale und Versorgungsspannungen können von der Vorderseite eines Dies direkt zur Rückseite – und damit zum nächsten gestapelten Die – geführt werden, ohne den Umweg über Bond-Drähte am Chiprand.

Herstellungsprozess

Der Prozess gliedert sich grob in drei Schritte, die sich durch den Zeitpunkt der Via-Erzeugung relativ zur restlichen Fertigung unterscheiden:

- Via-first: Das TSV wird vor der Transistorfertigung in den nackten Wafer geätzt. Dieser Ansatz erlaubt hochtemperaturstabile Füllmaterialien wie polykristallines Silicium, ist aber mit modernen CMOS-Prozessen schwer vereinbar.
- Via-middle: Das TSV entsteht nach der Transistorfertigung (FEOL), aber vor der Metallisierung (BEOL). Dies ist heute der industriell dominante Ansatz, da er die etablierte Transistorfertigung unangetastet lässt.
- Via-last: Das TSV wird erst nach vollständiger Chipfertigung, teils sogar von der Waferrückseite aus, geätzt. Vorteil ist die vollständige Entkopplung von der Vorderseiten-Prozessierung.

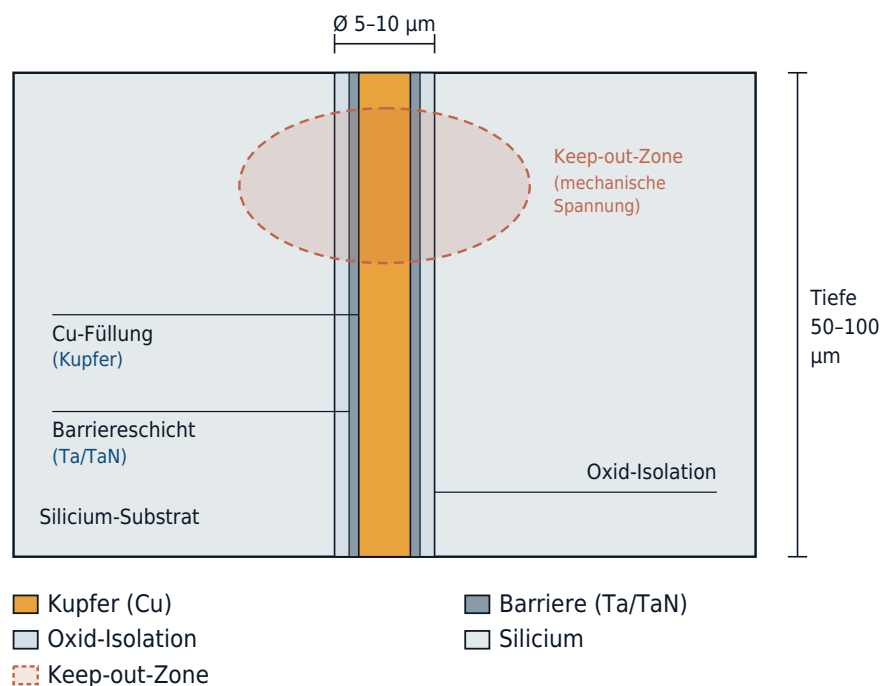
Das eigentliche Via wird mittels Deep Reactive Ion Etching (Bosch-Prozess, siehe Trockenätzen-Kapitel) geätzt und erreicht dabei Aspektverhältnisse von 10:1 oder höher – typische Durchmesser liegen bei 5–10 µm, bei Tiefen von 50–100 µm.

Isolation und Cu-Füllung

Da das umgebende Silicium selbst halbleitend ist, muss die Via-Wand zunächst mit einer dünnen Oxidschicht (meist per PECVD, siehe Abscheidung-Kapitel) elektrisch isoliert werden. Anschließend folgt eine Barrierschicht (typischerweise Ta/TaN), die die Diffusion von Kupfer in das Silicium verhindert – dieselbe Problematik wie bei der Kupfermetallisierung (siehe Kupfertechnologie-Kapitel). Erst danach wird das Via elektrochemisch mit Kupfer aufgefüllt.

Aufbau einer TSV im Querschnitt

Der Aufbau eines Through-Silicon Via (TSV)
Querschnitt durch ein gefülltes Via im Silicium-Substrat



Herausforderungen beim Aspektverhältnis

Je höher das Aspektverhältnis, desto schwieriger die vollständige, void-freie Kupferfüllung – ähnlich der Spaltfüllungsproblematik bei CVD-Verfahren, nur in deutlich größerem Maßstab. Zusätzlich erzeugt das eingefüllte Kupfer aufgrund seines gegenüber Silicium stark abweichenden thermischen Ausdehnungskoeffizienten mechanische Spannungen im umgebenden Material, die die Transistor-Performance in unmittelbarer TSV-Nähe beeinträchtigen können („Keep-out Zone“).

2.5D-Integration: Der Interposer

Das Grundprinzip

Bei der 2.5D-Integration werden mehrere fertig prozessierte Dies nicht direkt übereinandergestapelt, sondern nebeneinander auf einem gemeinsamen Trägersubstrat – dem Interposer – platziert. Der Interposer selbst enthält keine aktiven Transistoren, sondern dient ausschließlich als hochdichte Verdrahtungsebene, die die einzelnen Dies untereinander sowie mit dem darunterliegenden Package-Substrat verbindet. Der Name „2.5D“ verweist darauf, dass hier zwar eine zusätzliche vertikale Ebene hinzukommt, die eigentliche Chip-Integration aber weiterhin lateral (nebeneinander) erfolgt.

Silizium- vs. organischer Interposer

- Silizium-Interposer: Gefertigt aus einem Silicium-Wafer, durchsetzt mit TSVs zur Durchkontaktierung und mehreren Metallisierungslagen (RDL, siehe unten) auf der Oberseite. Silizium erlaubt sehr feine Verdrahtungsstrukturen (Sub-Mikrometer-Pitch), ist aber vergleichsweise teuer in der Herstellung und in der Fläche begrenzt (typischerweise auf die Größe eines Reticles bei der Belichtung).
- Organischer Interposer: Basiert auf Leiterplattenmaterialien ähnlich klassischer PCB-Technik, deutlich kostengünstiger und großflächiger herstellbar, erreicht jedoch nur gröbere Verdrahtungsdichten.

Redistribution Layer (RDL)

Die Redistribution Layer ist eine zusätzliche Metallisierungsebene auf dem Interposer (oder direkt auf einem Die), die die vergleichsweise grob gerasterten Kontaktflächen eines Chips auf die viel feineren Anschlussstrukturen des Interposers „umverteilt“ – daher der Name. RDL-Strukturen werden mit ähnlichen Lithografie- und Kupfertechnologien gefertigt wie die BEOL-Verdrahtung eines regulären Chips, jedoch mit gröberen Strukturbreiten und dickeren Leiterbahnen, da hier eher Stromtragfähigkeit als maximale Dichte im Vordergrund steht.

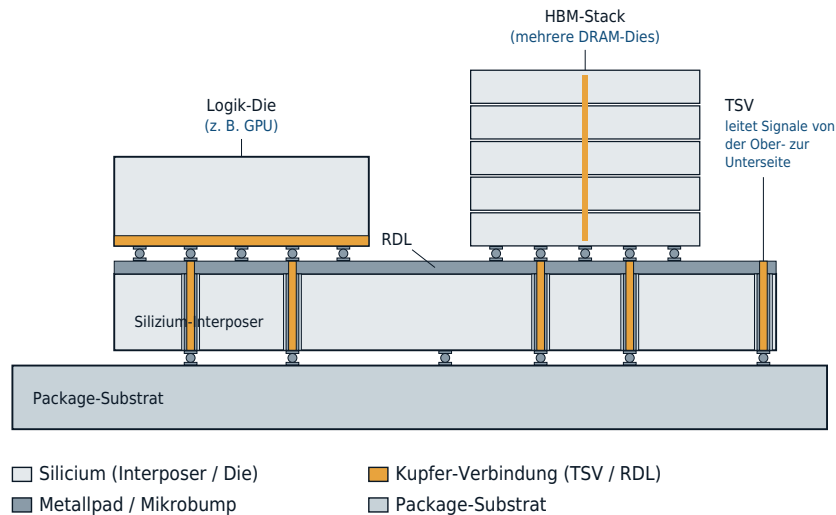
Beispiel: HBM auf Interposer

Das prominenteste Anwendungsbeispiel ist die Kopplung eines Logik-Dies (z. B. GPU) mit mehreren High-Bandwidth-Memory-Stapeln auf einem gemeinsamen Silizium-Interposer. Die extrem kurzen, dichten Verbindungswege über den

Interposer ermöglichen Speicherbandbreiten, die über klassische Leiterplattenverdrahtung nicht erreichbar wären (vertiefte Behandlung in Kapitel 6).

Interposer mit Logik-Die und HBM-Stack

2.5D-Integration: Interposer mit Logik-Die und HBM-Stack
Seitenschnitt eines gemeinsamen Silizium-Interposers



3D-Stacking

Die-to-Die vs. Wafer-to-Wafer Bonding

Während die 2.5D-Integration Dies nebeneinander anordnet, werden beim echten 3D-Stacking mehrere Dies direkt übereinandergestapelt und über TSVs vertikal verbunden. Zwei grundlegende Fertigungsansätze stehen zur Wahl:

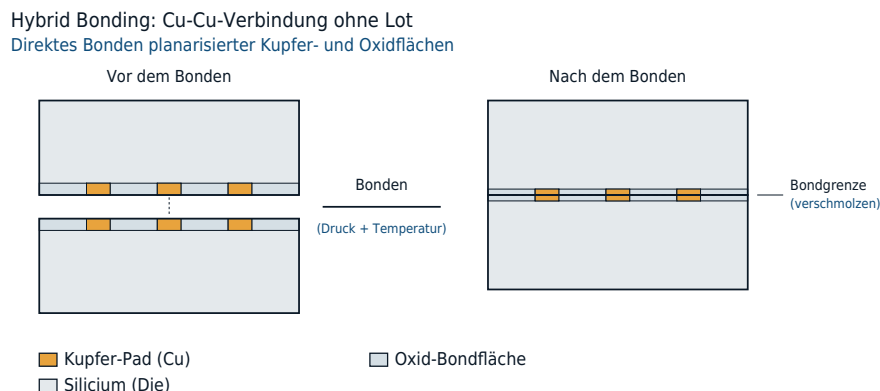
- **Die-to-Die (D2D):** Einzelne, bereits vereinzelte und getestete Dies werden übereinandergestapelt. Vorteil: Vor dem Stacking kann jeder Die einzeln auf Funktionsfähigkeit geprüft werden („Known Good Die“), was die Gesamtausbeute des Stapels deutlich verbessert.
- **Wafer-to-Wafer (W2W):** Zwei komplette, unvereinzelte Wafer werden als Ganzes miteinander verbunden und erst danach gemeinsam vereinzelt. Vorteil: Deutlich höherer Durchsatz und Positioniergenauigkeit, Nachteil: Ein defekter Die auf einem der beiden Wafer führt zum Ausfall des gesamten gestapelten Bauelements an dieser Position, da keine Vorselektion möglich ist.

Ein hybrider Ansatz, Die-to-Wafer (D2W), kombiniert die Vorteile beider Verfahren: geprüfte Einzeldies werden auf einen kompletten Wafer gestapelt.

Hybrid Bonding: Cu-Cu direkt

Die derzeit dichteste Form der 3D-Verbindung ist das Hybrid Bonding. Dabei werden auf beiden zu verbindenden Oberflächen gleichzeitig Kupferflächen (für die elektrische Verbindung) und Oxidflächen (für die mechanische Verbindung und Isolation) hochpräzise planarisiert – mittels Chemisch-Mechanischem Polieren, siehe entsprechendes Kapitel – und anschließend direkt, ohne zusätzliches Lot oder Bump-Material, aneinandergesetzt. Unter Druck und moderater Temperatur verschmelzen die Oxidflächen zunächst chemisch, das Kupfer dehnt sich anschließend thermisch aus und stellt eine direkte metallische Verbindung her. Auf diese Weise lassen sich Verbindungs-Pitches von unter einem Mikrometer erreichen – eine Größenordnung, die mit klassischen Löt- oder Bump-Verfahren unerreichbar ist.

Hybrid Bonding vor und nach dem Verbindungsprozess



Thermische und mechanische Herausforderungen

Gestapelte Dies erzeugen im Betrieb Wärme, die durch mehrere aktive Schichten hindurch abgeführt werden muss – ein Die in der Mitte eines Stapels hat einen deutlich schlechteren thermischen Pfad zur Wärmesenke als ein einzelner, plan aufliegender Chip. Zudem verstärken sich mechanische Spannungen durch unterschiedliche thermische Ausdehnungskoeffizienten der gestapelten Materialien mit jeder zusätzlichen Lage. Beides begrenzt in der Praxis, wie viele aktive, leistungsstarke Logik-Dies sinnvoll

übereinandergestapelt werden können, während passive oder stromsparende Speicherlagen (wie bei HBM) deutlich unkritischer sind.

Chiplets & heterogene Integration

Disaggregation von SoCs

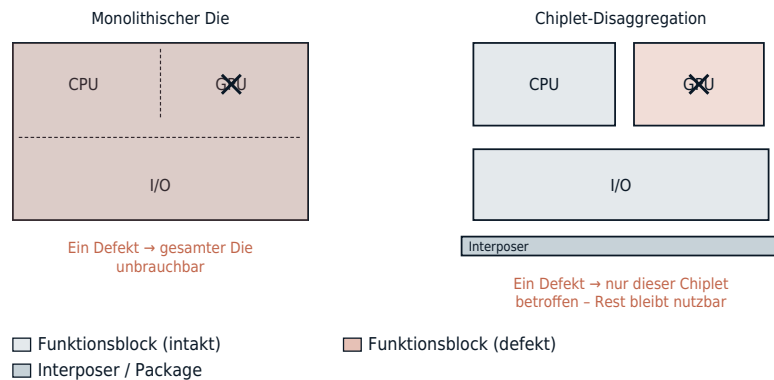
Ein klassisches System-on-Chip (SoC) integriert sämtliche Funktionsblöcke – CPU-Kerne, GPU, I/O-Controller, Speicherinterfaces – auf einem einzigen monolithischen Die. Der Chiplet-Ansatz zerlegt dieses SoC gezielt in mehrere kleinere, funktional abgegrenzte Dies („Disaggregation“), die anschließend über die in den vorherigen Kapiteln beschriebenen Advanced-Packaging-Verfahren (Interposer, Hybrid Bonding) wieder zu einem funktional zusammenhängenden System zusammengeführt werden.

Vorteile: Yield und Kosten

Der zentrale Vorteil liegt in der Fertigungsausbeute: Da die Defektwahrscheinlichkeit eines Chips überproportional mit seiner Fläche wächst, sinkt die Ausbeute eines großen monolithischen Dies drastisch schneller als die mehrerer kleiner Dies gleicher Gesamtfläche. Ein defekter kleiner Chiplet-Baustein wird verworfen, während der Rest des Systems weiterverwendet werden kann – bei einem monolithischen Chip führt derselbe Defekt zum Totalausfall. Zusätzlich lässt sich, wie in Kapitel 1 angesprochen, jeder Chiplet-Baustein im für ihn wirtschaftlich und technisch optimalen Fertigungsknoten produzieren, statt den gesamten Chip auf den teuersten benötigten Prozess festzulegen.

Auswirkung eines Defekts: monolithischer Die vs. Chiplet-Disaggregation

Monolithischer Die vs. Chiplet-Disaggregation Auswirkung eines Fertigungsdefekts auf die Ausbeute



Nachteile: Interconnect-Overhead

Diesen Vorteilen steht ein Nachteil gegenüber: Die Verbindung zwischen Chiplets – selbst über dichteste Advanced-Packaging-Verfahren – weist höhere Latenz und höheren Energieverbrauch pro übertragenem Bit auf als eine on-Chip-Verdrahtung innerhalb eines monolithischen Dies. Systemarchitekten müssen daher sorgfältig abwägen, welche Funktionsblöcke eng genug gekoppelt sind, um von monolithischer Integration zu profitieren, und welche sich für eine Trennung in separate Chiplets eignen.

Standardisierung: UCle

Damit Chiplets unterschiedlicher Hersteller in einem gemeinsamen System kombiniert werden können, braucht es standardisierte physikalische und protokollseitige Schnittstellen. Der Industriestandard Universal Chiplet Interconnect Express (UCle) definiert hierfür sowohl die elektrische Schicht (Signalpegel, Pitch, Bump-Anordnung) als auch das Übertragungsprotokoll, mit dem Ziel, einen herstellerunabhängigen „Chiplet-Marktplatz“ analog zur heutigen Standard-IC-Beschaffung zu ermöglichen.

Anwendungsfelder & Ausblick

HBM als Paradebeispiel

High Bandwidth Memory ist die konsequenteste kommerzielle Umsetzung der in

diesem Abschnitt behandelten Technologien: Mehrere DRAM-Dies (siehe DRAM-Zelle-Kapitel) werden mittels TSV und Wafer-to-Wafer- bzw. Die-to-Wafer-Bonding zu einem vertikalen Stapel von acht, zwölf oder mehr Lagen verbunden und anschließend über einen Silizium-Interposer mit einem Logik-Die gekoppelt. Die daraus resultierende Speicherbandbreite übersteigt die klassischer, planar auf der Leiterplatte verdrahteter DRAM-Module um ein Vielfaches, bei gleichzeitig deutlich geringerem Energieverbrauch pro übertragenem Bit.

KI-Beschleuniger

Moderne KI-Trainingsbeschleuniger sind der Haupttreiber der beschriebenen Packaging-Technologien: Sie kombinieren typischerweise mehrere Logik-Chiplets mit umgebenden HBM-Stapeln auf einem gemeinsamen Interposer, um sowohl die für KI-Workloads kritische Speicherbandbreite als auch die Rechendichte zu maximieren, ohne auf einen einzigen, unwirtschaftlich großen monolithischen Die angewiesen zu sein.

Foundry-Vergleich

Verschiedene Foundries haben eigene Markennamen für ihre Advanced-Packaging-Plattformen etabliert, die die in diesem Abschnitt beschriebenen Grundprinzipien – Interposer, TSV, Hybrid Bonding – jeweils unterschiedlich kombinieren und vermarkten. Diese Plattformen unterscheiden sich unter anderem in der Wahl von Silizium- versus organischem Interposer, dem unterstützten Aspektverhältnis der TSVs und der maximalen Anzahl kombinierbarer Dies.

Ausblick

Der Trend geht klar zu immer engerer, immer feiner aufgelöster 3D-Integration – von TSV-basierten Stapeln hin zu Hybrid Bonding mit stetig sinkendem Pitch. Gleichzeitig bleiben thermische Ableitung und Kosten die limitierenden Faktoren, weshalb Advanced Packaging heute weniger ein reiner Ersatz für, sondern eine notwendige Ergänzung zu klassischer Transistorskalierung darstellt.