Press Release

2023 年 7 月 5 日
東京工業大学

CPU/GPU とメモリをハイブリッド 3 次元積層

－超並列 3 次元配線によりテラビット伝送を
世界最小エネルギーで実現－

【要点】

○バンプレス Chip-on-Wafer（COW）、Wafer-on-Wafer（WOW）プロセスにより CPU/GPU とメモリを 3 次元に積層

○バンプレス COW、WOW プロセスの高密度垂直配線を用い、CPU/GPU とメモリを 16,000 本もの信号線で超並列に接続する「BBCube 3D」を提案

○BBCube 3D のデータ伝送は PC サーバ用メモリの 13 倍、AI などに使われる高帯域メモリ（HBM2E）の 4 倍となる高帯域を実現しながら、電力は PC サーバ、高帯域メモリそれぞれに対し 1/20、1/5 に抑えることができることを世界で初めて明らかにした

【概要】

東京工業大学 科学技術創成研究院 異種機能集積研究ユニットの大場隆之特任教授は、**WOW アライアンス**（用語 1）との共同研究により、バンプレス **Chip-on-Wafer**（COW、用語 2）プロセスおよび **Wafer-on-Wafer**（**WOW**、用語 3）プロセスによって CPU/GPU とメモリを 3 次元実装するハイブリッド 3 次元実装技術「BBCube 3D」を創出した。

現在、2 次元的な配線でのデータ伝送能力の向上が、物理的、消費電力的に困難となる中、さらなる高帯域化を目指した 3 次元積層半導体の開発が進められている。しかし、従来の 3 次元積層技術はチップの垂直配線 **Through Si Via**（**TSV**、用語 4）同士の接続に**マイクロバンプ**（用語 5）を用いることから垂直配線の高密度化が難しく、また、データ伝送時の消費電力増大の一因となる**寄生容量**（用語 6）の低減が困難という問題があった。

BBCube 3D のバンプレス COW、WOW プロセスは、銅（Cu）を配線に用い、埋め込み・研磨によって垂直配線を行う Cu ダマシン TSV 配線を用いることでバンプレス化し、垂直配線の 16 倍の高密度化、1/20 の寄生容量低減を実現できる。さらに CPU/GPU とメモリを 16,000 本もの信号線で超並列に接続する。BBCube 3D のデータ伝送は PC サーバ用メモリの 13 倍、AI などに使われる高帯域メモリ（HBM2E）の 4 倍となる高帯域を実現しながら、電力は PC サーバ、高帯域メモリそれぞれに対し 1/20、1/5 に抑えることができることを世界で初めて明らかにした。

この成果は 2023 年 6 月 11 日～ 6 月 16 日に開催の半導体回路・実装技術に関する国際会議「VLSI Symposium 2023」（主催：IEEE）で発表された。

●背景

AI や HPC（High Performance Computing）では CPU/GPU とメモリとの間で大量のデータ伝送することが要求される。これに応えるためにデータ伝送のスピードを 1 秒間に 100 億ビットまで増加させ、**Si インターポーザ**（用語 7）で多くのデータを送受する場合、4 千～6 千本までの配線を必要としていた。しかし、従来の 2 次元的な配置では配線数をこれ以上増やすことは物理的に難しい。また、伝送距離とデータ伝送速度に比例した消費電力の増大が課題となっている。

こうした 2 次元的な配置の限界を突破し、半導体デバイスのさらなる高性能化、低消費電力化を実現するために、CPU/GPU とメモリとを 3 次元積層するハイブリッド 3 次元実装技術の開発が急ピッチで進められている。

しかし従来の 3 次元積層による半導体パッケージでは、チップの垂直配線（TSV）同士の接続にマイクロバンプを用いており、垂直配線の高密度化やデータ伝送消費電力増大の一因となる寄生容量の低減が困難という問題があった。

本研究では、銅（Cu）を配線に用い、埋め込み・研磨によって垂直配線を行うバンプレスプロセスを用いて CPU/GPU とメモリとを 3 次元積層する BBCube 3D を提案することで、AI HPC 向け半導体の CPU/GPU とメモリ間データ伝送の高帯域と消費電力低減の両立を目指した。

●研究成果

本研究におけるバンプレス WOW、COW プロセスの流れを説明する（図 1）。チップとチップの TSV 垂直配線は **Via-Last 法**（**Via-Last**、用語 8）を用いている。

まず、ワッフル状にしたウエハである**ワッフルウエハ**（用語 9）に CPU または GPU のチップを搭載し（図 1-1a）、モールディングして（図 1-2a）、ウエハを薄化（図 1-3）、TSV を形成する（図 1-4, 5）。同様にキャッシュチップをワッフルウエハに搭載し（図 1-1b）、モールディングして（図 1-2b）、CPU/GPU のウエハに積層し薄化する（図 2-6）。その後 TSV を形成して CPU/GPU とキャッシュとを接続する（図 1-7, 8）。DRAM は DRAM ウエハをキャリアウエハに貼付け、薄化する（図 1-2c）。その後、CPU/GPU、キャッシュを搭載したウエハに積層し（図 1-9）、TSV を形成して、キャッシュと DRAM を接続する（図 1-10, 11）。図 1-9 から図 1-11 を繰り返して DRAM を必要層数積層した後、ダイシングして個別のチップに分ける。

本研究で検証を進めている WOW、COW プロセスの断面写真を図 2 に示す。

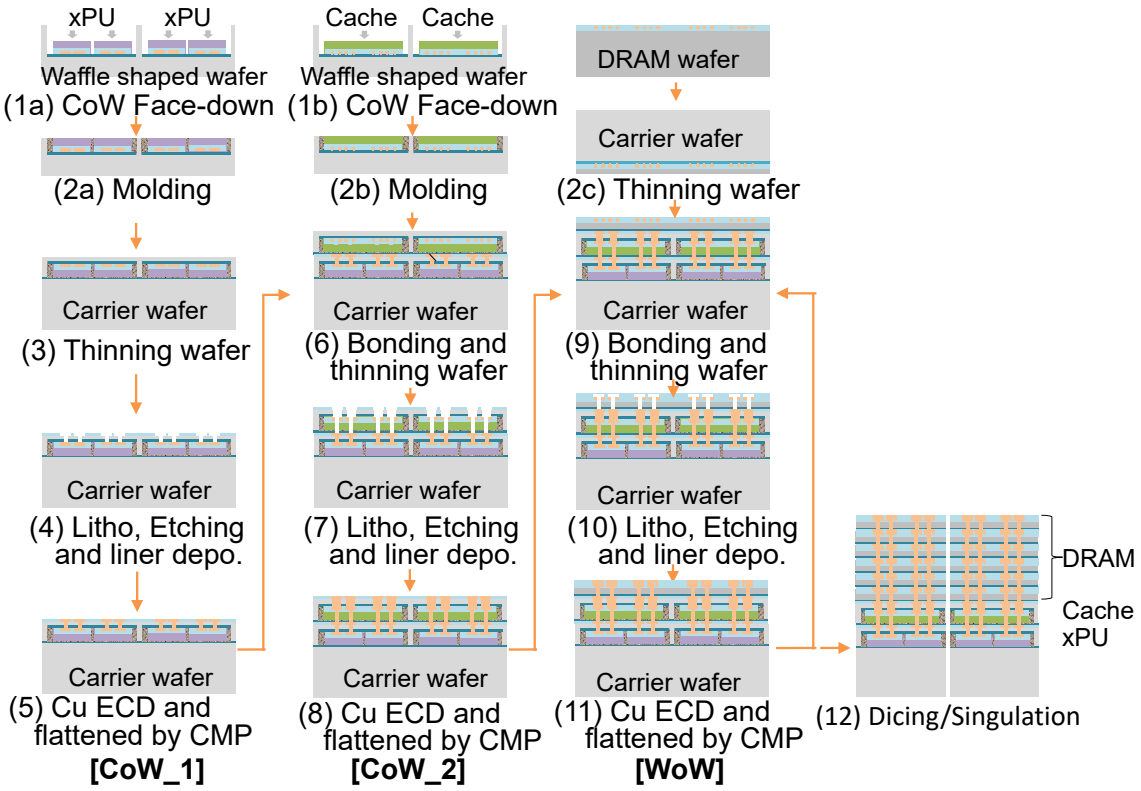

図 1 バンプレス WOW、COW プロセスフロー

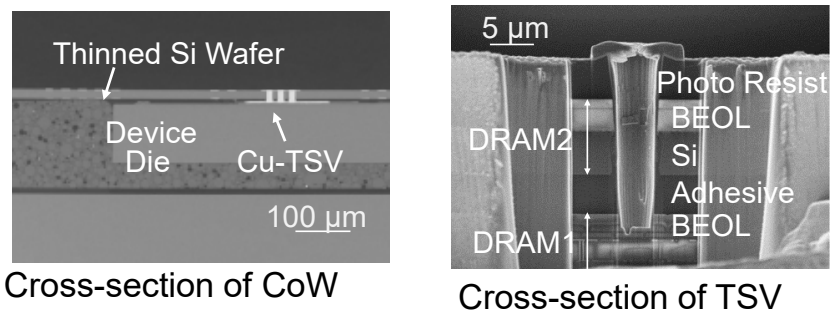

図 2　バンプレス WOW、COW の断面写真

BBCube の垂直配線（TSV）の寄生容量を 3 次元電磁界解析を用いて計算した。モデルと結果を図 3 に示す。BBCube の TSV は従来の 3 次元実装の TSV に対し 16 倍の密度を実現しながら 1/20 の寄生容量となる。BBCube の TSV の寄生容量は Si 上に 2 次元的に配置した配線の長さに換算すると、わずか 30 μ m 分の寄生容量に過ぎない。従来 CPU/GPU とメモリの間は、配線長さを短くできる Si インターポーザを用いたものでも 600 μ m ある。つまり CPU/GPU を垂直に配置して BBCube 3D の TSV を介して接続することで、配線の寄生容量を 1/20 に減らすことができ、データ伝承消費電力を低減することができる。さらに CPU/GPU を垂直に配置すると、CPU/GPU 間を接続する TSV は面で配置することができる。チップサイズ（側壁の長さ）で制限されていた従来の実装に対し、BBCube 3D は、CPU/GPU とメモリを 16 倍の 16,000 本もの信号線で超並列に接

続することができる。BBCube 3D の構成を図 4 に示す。

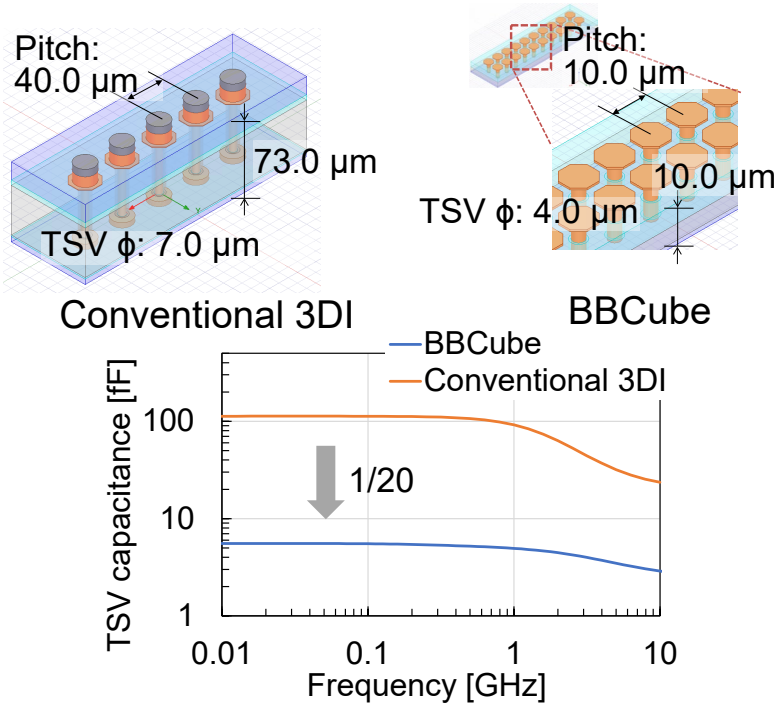

図 3　垂直配線（TSV）モデルと寄生容量計算結果

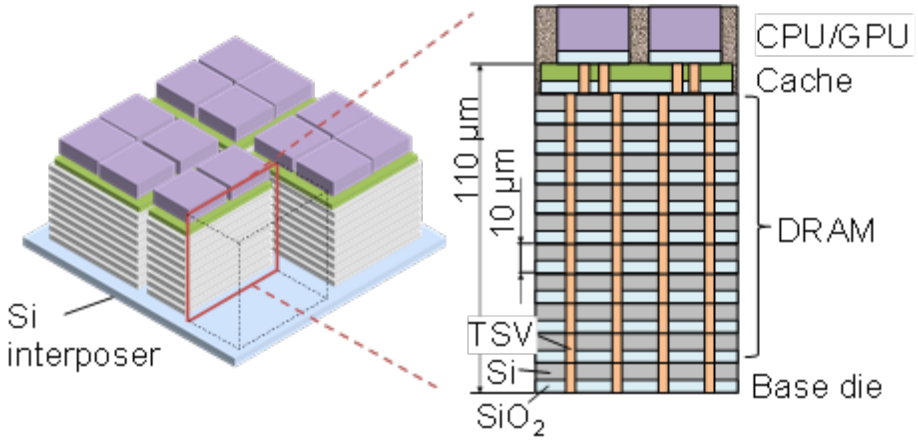

図 4　　BBCube 3D の構成

CPU/GPU とメモリ間のデータ伝送にかかるエネルギーを解析し、PC/サーバ用メモリ（DDR5）、AI などに使われる高帯域メモリ（HBM2E）との比較を行った。結果を図 5 に示す。BBCube3D は DDR5 の 13 倍、HBM2E の 4 倍の高帯域なデータ伝送を実現しながら、電力はそれぞれに対し 1/20、1/5 に抑えることが可能となる。

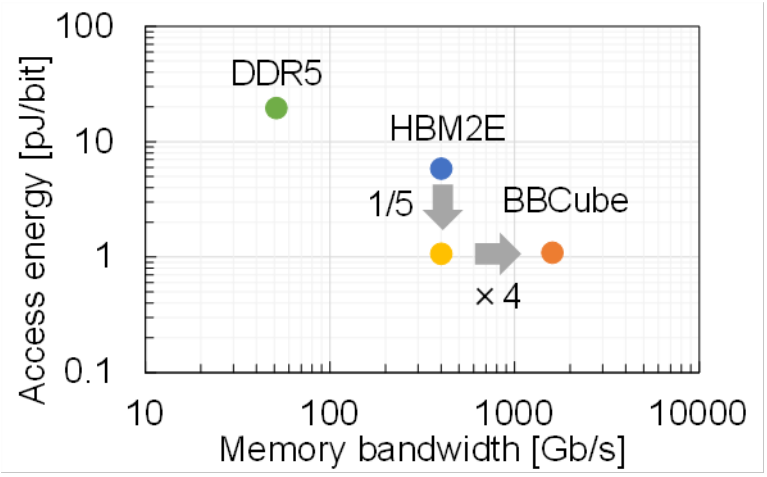

図 5　データ伝送容量とデータ伝送にかかるエネルギー

●社会的インパクト

微細化とともに 300 兆円まで巨大化した半導体産業は、各国の経済安全保障、社会インフラすべてわたって要になっている。微細化技術の発展は限界を迎えつつあり、DRAM や CPU デバイス単体の競争力は小さくなり、異なるデバイスをまとめたシステムの大量生産が次の競争力となっている。本技術は、半導体産業の悩みである微細化の終焉に応えるものであり、次世代半導体システムの道を示すものである。特に性能と生産性の両輪が求められる半導体産業において、300mm ウエハプロセスを軸に開発された三次元集積技術は、社会実装に直結し、高い競争力を得ることができる。

同技術は本学拠点の産学アライアンスで開発されたものであり、技術と組織、それぞれのシームレス化、柔軟な開発計画に基づく。半導体において国プロによらない産学協調開発は初めての試みである。

●今後の展開

今回開発した BBCube 3D により、AI、HPC 用半導体の電力を大幅に低減することが可能となる。今後は、CPU/GPU とメモリをバンプレス WOW、COW で積層した BBCube 3D を試作し、大容量データ伝送と低消費電力の両立の実証を目指す。

【用語説明】

（1）　**WOW アライアンス**：東京工業大学を中心とした産学研究プラットフォーム。半導体関連の設計・プロセス・装置・材料などを手がける企業、および研究機関によって構成される。高度かつ簡便なウエハの薄化技術・積層技術を持ち、バンプレス TSV 配線を用いた 3 次元化技術を世界で初めて開発に成功した。

（2）　**Chip-on-Wafer（COW）**：チップをウエハ上に接合する技術。一般的にチップの接合は、樹脂材料でできた配線基板に対して行われており、それと区別するために「ウエハ上に（Chip-on-Wafer）」という表現が用いられている。チップをウエハ上に接合することにより、以降の半導体製造工程において、各種装置を用いた高精度な加工が行えるようになる。

(3) **Wafer-on-Wafer（WOW）**：ウエハ上にウエハを接合する技術。COW と同様の技術であるが、多くのチップが形成されたウエハ同士を接合することにより、同時に多数のチップの積層ができ、COW よりも効率よい。ただしチップのサイズは同一である必要があるため、メモリ等同一チップを積層する場合に向く。

(4) **Through-Silicon-Via（TSV）**：シリコン（Silicon）ウエハを貫通（Through）させて開けた接続孔（Via：ビア）。上下に積層したチップを、埋め込み配線によって接続させる。最近では、シリコン材料以外にも配線するため、前工程における垂直配線（vertical interconnects）とした方が分かりやすい。

(5) **マイクロバンプ**：電極部に半田で形成した配線接続のための突起。半田を溶かして圧力をかけて接合させるため、隣のバンプと短絡しないよう間隔が必要。

(6) **寄生容量**：電圧のかかった導体の間に、設計の意図から外れて発生する電気的な負荷。

(7) **Si インターポーザ**：半導体のパッケージにおいて、端子ピッチが異なる半導体とパッケージ基板の間を中継する Si 製の電源基板。主に、ハイエンド向け半導体のパッケージにおいて用いられる。

(8) **Via-Last 法**：チップ間の電気配線接続方法の一つ。チップやウエハを積層した後にエッチング加工で接続孔を形成し、Cu などの金属をスパッタとメッキで充填して配線として利用する。従来は Via-First とよばれ、チップに予め Cu 金属を埋設しておき、チップやウエハ同士を積層する際に、同時に金属と金属を機械的に接触させ、熱処理や圧縮応力を利用して導通界面を形成する。Via-Last は前工程、Via-First は実装由来の方式であり、微細化と高信頼性配線には Via-Last が用いられる。

(9) **ワッフルウエハ**：その名の通り菓子の「Waffle」を由来としたウエハ表面の加工形状で、ウエハ表面に四角い溝を規則的にエッチング加工したものである。ドライエッチングの最適化で幾何学的な段差と平坦な底部を形成し、接着層を塗布後、底部にチップを Face-down でボンディングする。ボンディングが完了したらモールディングし、モールド材料とチップ裏面のシリコンを同時に研削、そして平坦化する。チップで専有された四角い溝の残りがモールド体積となる。

【論文情報】

学会名：IEEE 2023 Symposium on VLSI Technology and Circuits (2023 VLSI Symposium)
論文タイトル：Bumpless Build Cube (BBCube) 3D: Heterogeneous 3D Integration Using WoW and CoW which Enables TB/s Bandwidth with Low Bit Access Energy
発表者：Norio Chujo, Koji Sakui, Shinji Sugatani, Hiroyuki Ryoson, Tomoji Nakamura and Takayuki Ohba

【問い合わせ先】

東京工業大学 科学技術創成研究院 異種機能集積研究ユニット
秘書　川島

Email: kawashima.t.ae@m.titech.ac.jp
TEL: 045-924-5866

【取材申し込み先】

東京工業大学 総務部 広報課
Email: media@jim.titech.ac.jp
TEL: 03-5734-2975　FAX: 03-5734-3661