半導体入門講座・第 26 回

# 先端パッケージと新しい材料：縮めるだけではない

プロセッサの計算はどんどん速くなるのに、メモリからデータを送る速さが追いつかず、この「メモリの壁」のせいで AI チップはよく止まって待つことになります。メモリを積み重ねてプロセッサのすぐとなりに置く、チップを 1 つのパッケージにまとめる、遠い道のりは光で送る。どれもこの壁と関係があります。材料を変える道も加えて、この道筋がわかれば、ニュースに次々出てくる新しい言葉がつながります。

半導体入門講座：チップの一生をたどる
授業や非営利の目的なら無料で使えます。fab24hr と明記してください。販売することや、自分の作品として使うことはできません。

# わかること

1　メモリの壁とは何か、つまりプロセッサがなぜよく止まってデータを待つのか、データを運ぶほうがなぜ計算より電気を使うのか、そしてなぜ AI がいちばん強くぶつかるのかを説明できる。

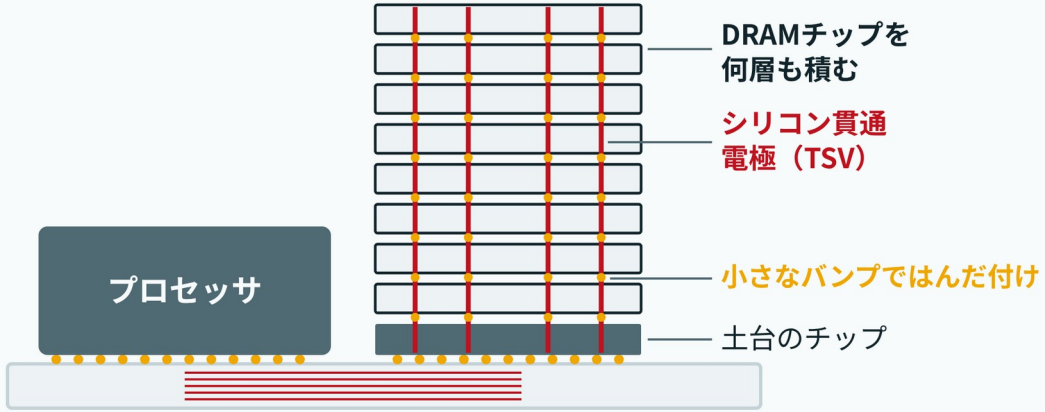
2　キャッシュ、 HBM 、チップレット、 2.5D 、 3D がそれぞれ何をしているかを説明でき、これらの道はいっしょに使えること、どれにも代償があることを知る。

3　ワイドバンドギャップの材料とシリコンフォトニクスがそれぞれどんな問題を解くのかを区別でき、どちらもシリコンや銅線を全部置き換える話ではないことを知る。

**ふつうのメモリ：マザーボードに差し込み、プロセッサから数センチ**

**メモリモジュール1枚で、一度に0か1を64個**

**HBM：DRAMを積み重ね、インターポーザの上でプロセッサのとなりに**




インターポーザ：薄いシリコンの板。配線は回路基板よりずっと細く密

**積み重ね1つにデータ線1,024本。メモリモジュールの16倍**
**最新世代の規格では2倍の2,048本**

HBM の中身も DRAM で、違うのは並べ方とつなぎ方です。積み重ね、シリコンを貫いてつなぎ、プロセッサのすぐとなりに置きます。線が 16 倍（最新世代はさらに 2 倍）あるので一度にたくさん運べ、道が短いぶん 0 か 1 を 1 個送るエネルギーも少なくてすみます。模式図で、層や穴、線の数は実際どおりではありません。

# メモリの壁：計算は速いのに、データが届かない

プロセッサは計算を受け持ち、メモリは計算に使うデータを置いておきます。この 2 つは伸びる速さが違います。 1980 年代半ばから 2000 年代初めまで、プロセッサは毎年およそ 5 割以上速くなりましたが、メモリからデータを 1 つ読むときの待ち時間は、 1 年で 1 割も縮みませんでした。メモリチップがいちばん重んじるのは、たくさん入って安いことで、速さは二の次です。データが届かなければ、プロセッサはどれだけ速く計算できても止まって待つしかありません。この差をメモリの壁と呼びます。いちばん強くぶつかっているのは AI です。モデルの数字は多すぎてプロセッサには入りきらず、外のメモリに置かれます。チャットボットが文章を少し出すたびにその多くを読み出し、 1 人のためだけに計算するときは、運んできた数字 1 つを掛け算 1 回と足し算 1 回くらいに使うともう次の数字なので、計算する回路は多くの時間を待って過ごします。ある研究が 20 年分のサーバーを調べたところ、計算の速さはおよそ 6 万倍になったのに、メモリがデータを送る速さはおよそ 100 倍にしかなっていませんでした。データを運ぶには電気もかかります。 0 と 1 が切り替わるたびに配線全体に電気をためたり抜いたりし、チップの外のメモリまでは数センチの道のりです。ある研究の見積もりでは、 1 回読むのに、簡単な足し算 1 回の数百倍から千倍以上の電気を使うことがあります。

# データを近くに置く：キャッシュと HBM

業界のやり方は台所にたとえられます。料理人の手はとても速いのに、食材が遠くの倉庫にあるなら、よく使う食材を手元に置き、倉庫を近くに移し、台車を大きくすればいい。手元にあたるのがキャッシュで、プロセッサと同じチップの上にあり、よく使う数字をそこに置いておけば速く、電気も節約できます。ただキャッシュは SRAM でできていて、1 ビット覚えるのにトランジスタを 6 個使い、場所をとるので、大きな AI のモデルは入りきりません。昔の壁はおもに「1 回待つのにどれだけかかるか」でしたが、いまは「1 秒にどれだけ送れるか」のほうがよく話題になります。そこで、メモリ（DRAM）のチップを何層も積み重ねて、プロセッサのとなりに置きます。これが HBM（広帯域メモリ）です。積み重ね 1 つにデータ線が 1,024 本あり（最新の世代はさらに倍）、ふつうのパソコンのメモリモジュール 1 枚が一度に送るのは、0 か 1 を 64 個並べた分だけです。線が多いので一度にたくさん送れ、道のりが短いので 0 か 1 を 1 つ送る電気も少なくてすみます。そのかわり、どの層も良品でなければならず、とても薄く削る必要があり、熱も逃げにくく、高くてつくるのが難しいのです。

# チップを 1 つにまとめる：チップレット、 2.5D 、 3D

HBM をプロセッサのすぐとなりに置けるのは、先端パッケージのおかげです。計算するチップとメモリ、または計算するチップ何個かを、 1 つのパッケージに入れます。最新の世代の AI チップのとなりに HBM を並べるように、これらの道はよくいっしょに使われます。組み合わせ方はおもに 2 つ。とても細い配線を描いた薄い板（インターポーザ。シリコンでつくることが多い）の上に横に並べるのが 2.5D 、そのまま積み重ねて、シリコンを貫いて金属を詰めた細い穴でつなぐのが 3D です。回路基板の上の数センチの配線を通るのに比べて、道は短く、線も多いので、データは速く届き、電力も少なくてすみます。大きなチップを、チップレットと呼ばれる小片に分けることもよくあります。チップが大きいほど、ウェーハの上の欠陥に当たりやすく、使えるものの割合（歩留まり）が下がり、露光装置は 1 回に 1 区画ずつしか焼けず、いちばん大きなチップはもうほぼ 1 区画の大きさだからです。それぞれつくってテストしてから組み合わせ、計算を受け持つ部分は最新の世代で、新しい世代でつくってもあまり小さくならない部分は、何世代か前の安くて安定したつくり方でつくります。そのかわりパッケージは難しく高くなり、中の 1 個が不良なら 1 組まるごとむだになります。積み重ねる場合はチップをとても薄く削らなければならず、あいだにはさまれた層は熱も逃げにくくなります。

# ほかの 2 つの道：遠い道のりは光で、材料を変える

データを速く送るほど、銅線で届く距離は短くなり、信号を立て直す電気も増えます。そこで、サーバーのラックとラックのあいだのような遠い道のりでは、電気を光に変えて光ファイバーで送ります。シリコンフォトニクスは、チップをつくる方法で、シリコンの上に光を通す細い道や、電気を光に、光を電気に変える小さな部品をつくる技術です。 CPO （コパッケージド・オプティクス）は、さらに「電気を光に変える」部分を、データを送り分けるチップと同じパッケージの中に移します。電気は数ミリから数センチ走るだけで光に変わります。材料を変える道が解くのは別の問題です。炭化ケイ素と窒化ガリウムのバンドギャップ（電子が電気を通すために跳び上がらなければならない「段差」）は、シリコンのおよそ 3 倍の高さで、ワイドバンドギャップ材料と呼ばれます。高い温度や電圧に耐えられ、電力を扱うスイッチにするとむだが少なく、電気自動車や急速充電器に使われています。ただ、結晶を育てるのが難しく、プロセッサやメモリはこの強みを生かせないので、チップの大半は今もシリコンでつくられています。多くがまだ研究や立ち上げの段階にある道もあります。簡単な計算をメモリの中でやってしまうこと（いちばん安上がりな運び方は、運ばないこと）や、超ワイドバンドギャップ（中国語圏では「第 4 世代半導体」）と呼ばれる酸化ガリウムやダイヤモンドを使うことです。

# よくある誤解

**よく聞く：ワイドバンドギャップ半導体（中国語圏でいう「第 3 世代半導体」）は、「2 ナノ」より新しい世代のプロセスだ。**

実は：この「世代」は、チップをつくる材料の話です。たいてい炭化ケイ素と窒化ガリウムのことで、新しい世代のシリコンのプロセスではありません。「2 ナノ」は同じ面積にスイッチをいくつ詰められるかの話、炭化ケイ素や窒化ガリウムは高い電圧や大きな電流、熱に耐えられるかの話です。電力を扱うスイッチは大きな電流を通すので、もともと大きくつくり、いちばん細い線は要りません。両者は別の問題を解いていて、どちらが進んでいるということはありません。

**よく聞く：ニュースが「銅線から光へ」と言うのだから、これからはチップの中も光で計算するようになる。**

実は：ニュースに出てくるシリコンフォトニクスは、たいてい「送る」ためのもので、光で「計算する」話ではありません。チップの中の計算は、いまもトランジスタがしています。「銅線から光へ」は、たいてい、どんどん速く、長くなっていく区間が光に変わるという話で、チップの中や、すぐ近くのチップどうしは、いまもほとんど電気です。光にもまだ難しさがあります。シリコンはほとんど光らず、光ファイバーはとても正確に合わせる必要があり、レーザーは熱に弱いのです。

# 遊ぶ

## 第 11 章「ボールが浮いた」から：シューフイと組立ラインを歩き、チップが抵抗やコンデンサといっしょに小さなモジュールに収まるのを見る

スマホで読み取って、ここから遊ぶ

fab24hr.com/ja/?story=11&utm_source=slides

# 考える

1　大きなチップをチップレットに分ければ、歩留まりがよくなるかもしれず、お金もいちばん必要なところに使えます。でも、つなぐための場所が余分に要り、パッケージングも難しくなり、中の 1 個が不良なら 1 組まるごとむだになります。あなたなら、どんなチップなら分ける価値があると考えますか？ どんなチップなら 1 個のままにしますか？

2　メモリの壁に対しては、キャッシュを増やす、 HBM をもっと並べる、といった道があり、メモリの中で計算する研究も進んでいます。どの道にも代償があります。 AI で質問に答えるだけの機械をつくるとしたら、まずどの道にお金をかけますか？ 決める前に、何を知りたいですか？

# 必読

### メモリの壁とは？ チップの計算は速いのに、データが追いつかない理由

fab24hr.com/ja/blog/memory-wall/

### ニュースの「先端パッケージ」とは？ チップを横に並べたり積み重ねたりする理由

fab24hr.com/ja/blog/advanced-packaging/

### ニュースの「 HBM 」とは？ AI が特別なメモリを必要とする理由

fab24hr.com/ja/blog/what-is-hbm/

## 関連

ニュースの「ワイドバンドギャップ半導体」とは？ SiC と GaN をやさしく

ニュースの「シリコンフォトニクス」とは？ 電気の代わりに光でデータを送る理由

# 半導体入門講座：チップの一生をたどる

## https://fab24hr.com/ja/learn/beyond/

公開されている教科書、業界の標準、大学の講義資料、論文をもとにまとめました。ゲームに出てくる会社、人物、装置の番号、数字は架空のものです。

授業や非営利の目的なら無料で使えます。 fab24hr と明記してください。販売することや、自分の作品として使うことはできません。

公開資料をもとに AI の助けで作った一般的な模式図。誤りがあるかもしれません。特定の会社の装置ではありません