

先進封裝與新材料：不只靠縮小

處理器算得越來越快，記憶體送資料卻跟不上，這道「記憶體牆」讓 AI 晶片常常停下來等。把記憶體疊起來貼在處理器旁邊、把晶片包在一起、更遠的路改用光，都跟它有關；再加上換材料，懂了這幾條路，新聞上一個個新名詞就接得起來。

半導體入門：跟著晶片的一生

課堂與非營利可以免費使用，請註明 fab24hr ；不能販售，也不能改成自己的作品。

學到什麼

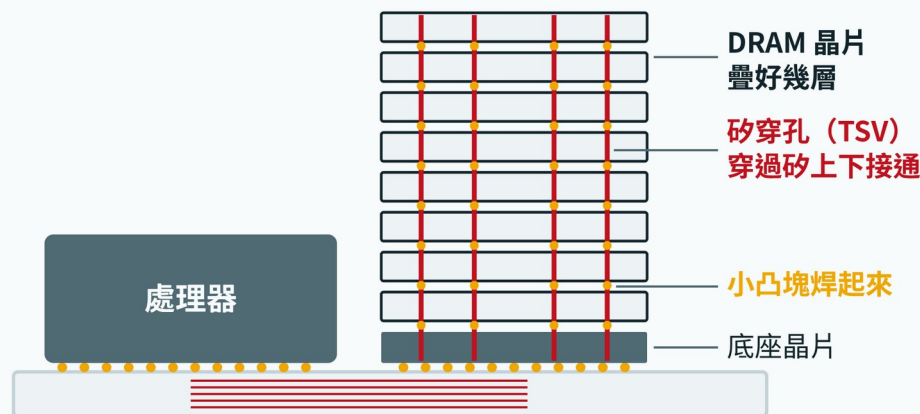
- 1 說明記憶體牆是什麼：處理器為什麼常停下來等資料、搬資料為什麼比算還耗電，以及為什麼 AI 撞得最重。
- 2 說明快取、HBM、小晶片、2.5D 和 3D 各在做什麼，也知道這幾條路可以一起用、各有代價。
- 3 分得清寬能隙材料和矽光子各解決什麼問題，也知道它們不是要把矽和銅線全部換掉。

一般的記憶體：插在主機板上，離處理器好幾公分



一條記憶體一次並排送 64 個 0 或 1

HBM：DRAM 疊成一疊，和處理器並排在中介層上



中介層：一塊薄矽片，上面的線比電路板細很多、密很多

一疊有 1,024 條資料線，是一般記憶體那一條的 16 倍
最新一代的標準再加倍，到 2,048 條

HBM 裡裝的還是 DRAM，差在擺法和接法：疊起來、穿過矽接通、貼在處理器旁邊。線多了 16 倍（最新一代再加倍），資料一次搬得多；路變短，送每個 0 或 1 用的電也比較少。示意圖，層數、孔和線的數目沒有照實際畫。

記憶體牆：處理器算得快，資料卻送不過來

處理器負責算，記憶體放著要算的資料。兩邊進步得不一樣快：從 1980 年代中期到 2000 年代初，處理器每年大約快五成以上，記憶體讀一筆資料要等的時間，每年只縮短不到一成；記憶體晶片最在意的是裝得多、便宜，速度排在後面。資料送不過來，處理器算得再快也只能停下來等，這道落差叫記憶體牆。AI 撞得最重：模型裡的數字多到放不進處理器，只能放在外面的記憶體；像聊天機器人每吐出一小段字，大量的數字都要讀一次，只替一個人算的時候，每個數字搬進來只做一次乘法、一次加法就換下一個，算的電路大半時間在等。有一篇研究整理了二十年的伺服器，能算的速度快了大約六萬倍，記憶體送資料只快了大約一百倍。搬資料也耗電：0 和 1 每換一次，就要把整條線充電或放電，到晶片外面的記憶體要走好幾公分；有研究估算，讀一次用的電可以是做一次簡單加法的幾百到上千倍。

把資料放近一點：快取和 HBM

業界的辦法像一間廚房：廚師手很快，食材卻放在遠方的倉庫，那就把常用的放在手邊、倉庫搬近一點、推車換大一點。手邊就是快取：跟處理器在同一顆晶片上，常用的數字留在這裡，快又省電；可是它用的 SRAM 一格要六個電晶體，很佔地方，大的 AI 模型還是裝不下。早年說的牆，多半是「等一次要多久」；現在更常說的，是「一秒送得出多少」。所以把好幾層記憶體（DRAM）晶片疊成一疊，放在處理器旁邊，叫 HBM（高頻寬記憶體）：一疊有 1,024 條資料線（最新一代再加倍），一般電腦的一條記憶體一次只送 64 個 0 或 1。線多，一次送得多；路短，送每個 0 或 1 用的電也少。代價是每一層都得是好的、要磨得很薄，熱也不容易散，又貴又難做。

把晶片包在一起：小晶片、2.5D、3D

HBM 能貼在處理器旁邊，靠的是先進封裝：把運算的晶片和記憶體，或好幾顆運算晶片，包進同一個封裝。所以這幾條路常一起用：最新一代做的 AI 晶片，旁邊並排幾疊 HBM。拼法主要有兩種：並排在畫了很細線路的薄片（中介層，常用矽做）上，叫 2.5D；直接疊起來，用穿過矽、填了金屬的細孔接通，叫 3D。比起隔著電路板上幾公分的線，路短、線多，資料跑得快也省電。一顆大晶片也常拆成幾小塊，叫小晶片（英文 chiplet）：晶片越大，越容易被晶圓上的缺陷砸中，能用的比例（良率）越低；曝光機一次只印一格，最大的晶片也已經差不多一格大。拆開來各自做、各自測過再拼：運算的那塊用最新一代，其他用新一代做也小不了多少的那幾塊，就用舊幾代、便宜又穩定的作法。代價是封裝更難、更貴，裡面一顆壞了整組都浪費；疊起來，晶片要磨得很薄，夾在中間的那一層熱也難散。

另外兩條路：更遠的路改用光、換材料

資料送得越快，銅線能送的距離越短，還要多花電把訊號補回來，所以比較遠的路，例如一櫃一櫃伺服器之間，改把電轉成光、走光纖。矽光子用做晶片的方法，在矽上做出讓光走的細路和電轉光、光轉電的小零件；共同封裝光學（CPO）再把電轉光搬進轉送資料那顆晶片的同一個封裝裡，電只走幾公釐到幾公分就變成光。換材料解的是另一件事：碳化矽、氮化鎵的能隙（電子要跳上去才能導電的「台階」）大約是矽的三倍高，所以叫寬能隙材料，撐得住更高的溫度和電壓，拿來做管電的開關，浪費的電少，用在電動車、快充充電器；可是晶體難長，處理器、記憶體也用不上它的好處，所以大部分晶片還是矽做的。還有些路大多還在研究或剛起步：把簡單的計算做進記憶體裡（最省的搬法就是不搬），或改用新聞有時叫「第四代半導體」的氧化鎵、鑽石。

常見誤解

常聽到：「第三代半導體」是比「2 奈米」更新的一代製程。

其實：「第三代」說的是做晶片的材料，多半指碳化矽和氮化鎵，不是新一代的矽製程。「2 奈米」比的是同一塊地方塞得下多少開關；碳化矽、氮化鎵比的是扛不扛得住高電壓、大電流和高溫。管電的開關要讓大電流通過，本來就得做大，用不到最細的線。兩邊解的是不同的問題，沒有誰比較先進。

常聽到：新聞說「用光取代銅線」，以後晶片裡都要改用光來算。

其實：新聞說的矽光子，多半負責「送」，不是用光來「算」；晶片裡的運算，還是電晶體在做。「光進銅退」多半是越來越快、越來越長的那幾段路改走光；晶片裡面、靠得很近的晶片之間，大多還是用電。光也還有難處：矽本身不太會發光，光纖要對得很準，雷射也怕熱。

玩

**從第 11 章「球掀起來了」開始：跟著淑惠姐
走一遍封裝線，看晶片和電阻、電容一起裝進
一個小模組**

用手機掃，從這裡開始玩

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



想一想

- 1 把一顆大晶片拆成小晶片，良率可能比較好，錢也花在最需要的地方；可是要多留接線的地方、封裝更難，裡面一塊壞了，整組都浪費。如果是你，什麼樣的晶片值得拆？什麼樣的你會留成一整顆？
- 2 對付記憶體牆，可以多放快取、在旁邊多放幾疊 HBM，也有人在研究把計算做進記憶體裡；每一條都有代價。如果你要做一台專門用 AI 回答問題的機器，你會先把錢花在哪一條？決定之前，你會想先知道哪些事？

必讀

記憶體牆是什麼？為什麼晶片算得快，資料卻送不夠快

fab24hr.com/blog/memory-wall/

新聞說的「先進封裝」是什麼？晶片為什麼要並排、疊起來

fab24hr.com/blog/advanced-packaging/

新聞說的「HBM」是什麼？AI 為什麼要特別的記憶體

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

延伸

新聞說的「第三代半導體」是什麼？碳化矽、氮化鎵白話解釋

新聞說的「矽光子」是什麼？為什麼要用光代替電來傳資料

半導體入門：跟著晶片的一生

<https://fab24hr.com/learn/beyond/>

內容整理自公開的教科書、產業標準、大學課程講義與論文；遊戲裡的公司、人物、機台編號和數字是虛構的。

課堂與非營利可以免費使用，請註明 fab24hr ；不能販售，也不能改成自己的作品。

通用示意：依公開資料、由 AI 協助製作，可能有錯；不代表任何一家公司的機台