

》學技術

## 創造殺手級3D IC產品 CPU/記憶體堆疊勢在必行

3D IC系列

唐經洲

中央處理器(CPU)的記憶體使用量極高，且占據CPU近三分之二的面積，經常成為效能、良率與測試的技術瓶頸；因此，若能利用矽穿孔(TSV)技術，將CPU與記憶體進行堆疊，將可打造晶片間傳輸速度更快、雜訊更小且效能更佳的一代三維晶片(3D IC)。

根據全球半導體聯盟(GSA)調查，記憶體在部分異質堆疊中是必要的元素。而中央處理器(CPU)使用記憶體量最多，主因是CPU的快取(Cache)與暫存器(Register)皆為繞線複雜度相當高的功能元件，從減輕繞線複雜度的角度觀之，此兩個區塊特別適用3D IC。

另外，因為快取記憶體介於CPU與主記憶體之間，其速度亦需相當快，記憶體面積也經常占據CPU近三分之二的面積(圖1)。在此種情形下，無論效能、良率或測試均為CPU的技術瓶頸，可稱之為記憶體的設計障礙(Memory Wall)。這個現象等同於一個都市內的車子越來越快，住家與辦公大樓也越來越豪華，惟都市的街道卻無法滿足這兩者之間的高速運輸量。

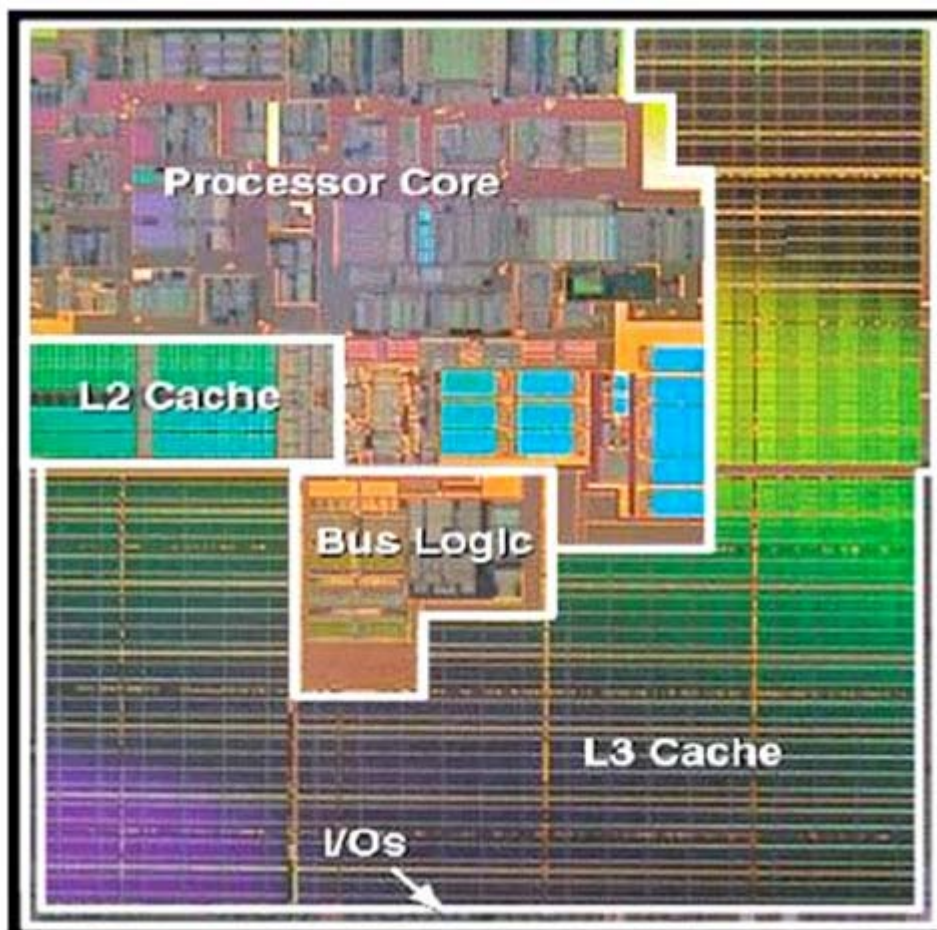


圖1 Itanium 2 MPU內部各功能區塊的分布情形 資料來源：Intel

### 快取扮演CPU與記憶體緩衝器角色

將Cache音譯成快取，自然是考慮到它將記憶體階層中下一層的存取速度提升到它的存取速度。快取記憶體是使用者透明(User-transparent)的快速記憶體，用以當作CPU和主記憶體的緩衝器，主要是利用程式的區域性特質記錄程式常用的資料及指令。

圖2是快取記憶體管理系統的示意圖。快取記憶體放在CPU與主記憶體之間，當CPU要讀取主記憶體資料或寫入資料到主記憶體，會先檢查資料是否早已放在快取記憶體，若有，則從快取記憶體存取資料；若無，則再想辦法從主記憶體拿取資料。由於快取記憶體係由靜態隨機存取記憶體(SRAM)組成，所以其存取資料的速度可較由動態隨機存取記憶體(DRAM)組成的主記憶體快。

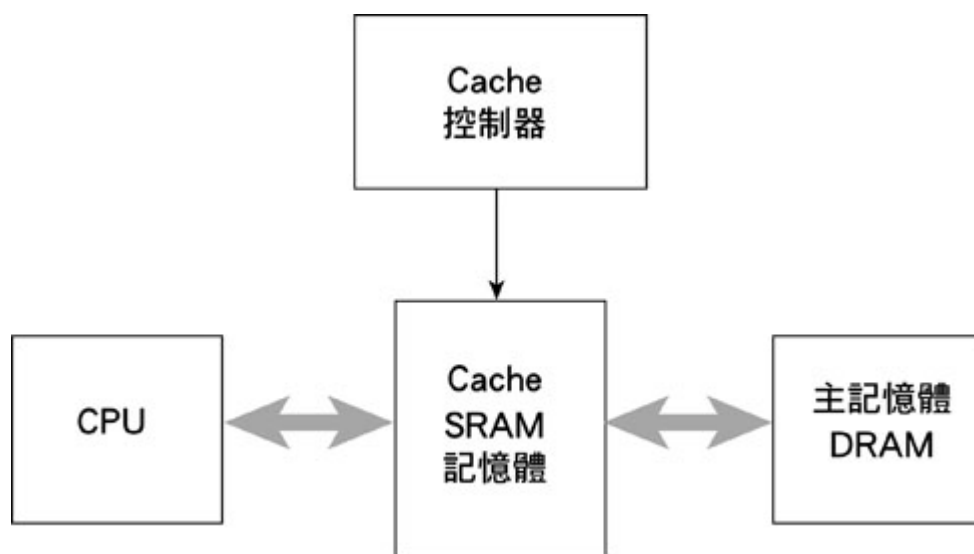


圖2 快取記憶體管理系統

簡言之，快取原理乃是利用程式區域性與動態重置性，讓使用者感覺擁有如下一層DRAM同樣的大容量，且像快取記憶體般快的記憶體系統。雖然80x86及其他大多數的微處理器系統中，快取係指介於主記憶體與暫存器之間的層次，但快取原理卻可應用在整個電腦記憶體系統中的任兩個層次之間。這裡所謂的快取記憶體，在L1當然大部分是暫存器，L2則為SRAM，現僅特別針對記憶體與CPU的堆疊整合提出說明。

### SoC面臨異質製程整合挑戰

過去的SoC目標係將各獨立的IC整合，透過SoC可減少面積與封裝需求，並提高其可靠度。然不幸的是，以一顆消費性或汽車電子的SoC而言，其可能要將邏輯電路、類比電路、快閃記憶體等整合，但其製程與設計的整合太多，代價亦太高，電路效能不一定好。英特爾(Intel)也早就指出若使用三個製程於一個大晶片，製程增加的費用係三倍。此論點，美國賓夕法尼亞州立大學(PSU)的教授謝源亦有提到。此因當晶片所需要的製程技術增加，如由單純的邏輯製程(運算邏輯單元(ALU)計算用)增加類比製程(倍頻用)，則費用就會增加為原有的兩倍。若是，再加上記憶體製程，費用為三倍。因此3D IC被視為是化解半導體製程發展危機的良策。

另外，從製程的觀點而言，一般邏輯製程與記憶體製程並無法匹配，這是因為邏輯製程會考量到可提供較多的金屬繞線層(Metal Layer)，而記憶體製程則希望提供較高密度

的儲存元件。此製程上的限制，對於要設計一個CPU是很巨大的挑戰，畢竟記憶體占絕大的面積。其實台積電早在2001年即有將內嵌式DRAM由兩個獨立的晶片(邏輯加上DRAM)組成或以多晶片模組(MCM)來取代的想法。其實，要將邏輯搭配記憶體並非CPU設計公司的專利，即使連爾必達(Elpida)這類型記憶體廠商也會想要堆疊。因此，以3D IC的方式來實現CPU將是未來趨勢，諸多公司咸認邏輯與記憶體的堆疊是3D IC的殺手級產品。

### 邏輯與記憶體位居上層各有優劣

由以上的討論得知，3D IC可減小外觀尺寸、增加頻寬與速度、降低功耗、減低生產費用、改善可靠度與測試品質、提供異質整合、減少靜電放電(ESD)需求、提高散熱效果等。但以邏輯與記憶體的堆疊問題是：哪一個元件會安排在上層？安排在上層與下層的考量因素又是什麼？以下針對幾家公司或者研究單位，討論邏輯在上層或記憶體在上層的不同可行性與其優缺點。

### PSU選擇將記憶體置於上方

2010年3月，謝源堆疊了五層3D IC(圖3)，其中兩層是邏輯層，而邏輯層是用特許(Chartered)的130奈米製程，各為一個H.264解碼器(Decoder)與一個32位元UniCore CPU；另外三層是DRAM，係由Tezzaron的製程完成。其中記憶體的面積較大：12.3毫米×21.8毫米，而邏輯層的面積則是2.5毫米×5毫米。由圖3可知，此設計是將記憶體放置於上方，而邏輯單元放置於下方。

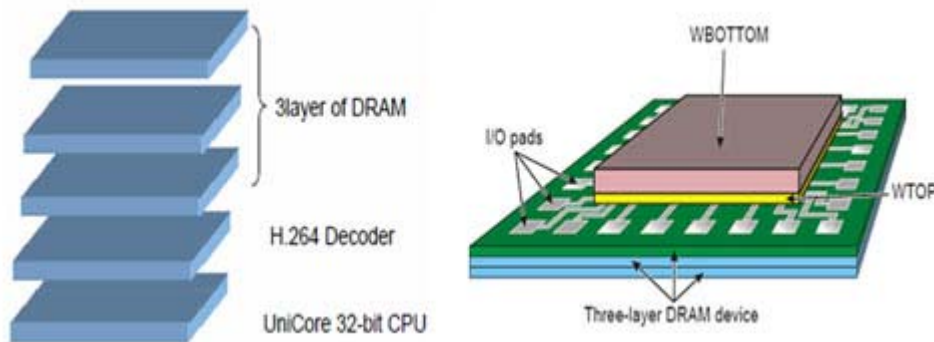


圖3 左為謝源教授的記憶體堆疊邏輯的3D IC設計規劃，右為面積概觀，WTOP即是H.264，WBOTTOM即是UniCore 32位元CPU。

### 益華電腦設計記憶體在上層

益華電腦(Cadence)為電子設計自動化(EDA)公司，早在DAC 2010說明到，其已與高通(Qualcomm)成功開發出堆疊Memory over Logic(28奈米)、Logic over Analog與Logic over Logic的三層堆疊。

該公司於2011年提出業界第一個三維記憶晶片介面標準規範Wide I/O記憶體控制器矽智財(IP)，此IP符合Wide I/O的要求，亦即512位元的資料寬度與12.8Gbit/s的資料傳遞速度。提供有運輸感測(Traffic Sensing)的機制，也就是系統會根據匯流排上面的資料傳遞情形來調整系統的功率消耗。另外更進一步提供動態調整電源電壓與操作頻率(DVFS)功能，以提供使用者可改變操作電壓與頻率來降低功耗。此設計使用到TSV，同時記憶體在邏輯上方。

### 喬治亞理工學院採記憶體階層設計

喬治亞理工學院的3D IC研究團隊著重於利用TSV於邏輯加上記憶體體的架構設計。該校的教授Sung Kyu Lim所設計的3D MAPS(3D MAssively Parallel processor with Stacked memory)是一個六十四核心(上層)與256KSRAM 記憶體(下層)堆疊的3D IC，特別用來處理一些須平行計算的應用程式。這是學術界第一顆成功使用3D IC技術於多核心與記憶體堆疊的設計，它整合KAIST、Tezzaron、Amkor與Board Lab幾家公司的技術(圖4)。如圖4，上層的厚度為12微米，下層為765微米，整體厚度就高達0.8毫米。這顆IC只有四十二個指令，操作頻率僅有277MHz，但透過TSV垂直訊號的平行傳遞，因此理論上的頻寬可高達70.9GB/s。該研究團隊分析，相對於以一個英特爾i7 CPU加上外部1.6GHz的第三代雙倍資料率(DDR3)其頻寬也只有 $1,600\text{MHz} \times 2 \text{通道} \times 8 \text{位元組} = 25.6\text{GB/s}$ ；另外以英特爾Xeon E7加上外部1.066GHz的DDR3，其頻寬= $1.066\text{MHz} \times 4 \text{通道} \times 8 \text{位元組} = 34.1\text{GB/s}$ 。若是用45奈米SoC製程完成，假設操作頻率因為使用前瞻45奈米製程使得頻率可提升五倍，所以其最多可以 $277\text{MHz} \times 5 \text{ (透過45奈米增加速度)} \times 64 \text{通道} \times 9 \text{ (更多區域)} \times 4 \text{ (更小核心)} \times 4 \text{位元組} = 12,764\text{GB/s}$ 。由此看來，3D IC真的可以大量提升速度。

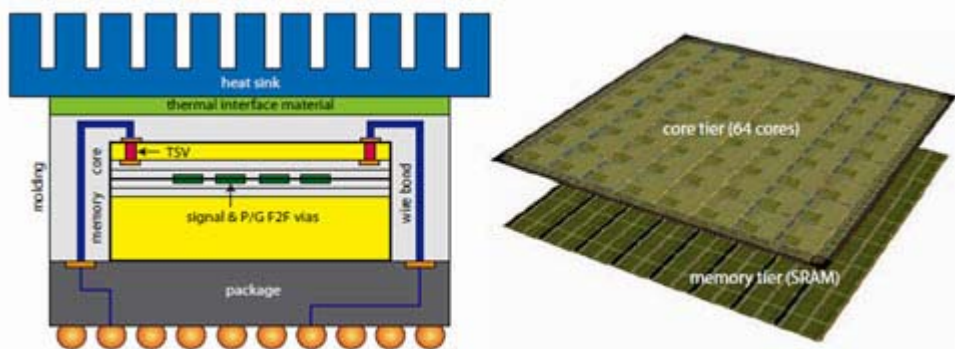


圖4 喬治亞理工學院使用Tezzaron的TSV F2F技術與Amkor之封裝技術所設計之3D IC。

從面積的角度來看，這個3D IC僅有5毫米×5毫米，若是使用等同i7的製程，面積會高達15毫米×15毫米。這顆IC的TSV是用在電源部分，晶片與晶片的訊號傳遞透過所謂的F2F(Face to Face)Pad。因為這顆3D IC內部的TSV是用在上下兩層的電源相接，最後上層的輸入輸出(I/O)Cell又以打線(Wire Bond)的方式外接到最下層的基底封裝部分，因此為提高良率，每個I/O Cell都有二百零四個冗餘(Redundant)的TSV(圖5)。

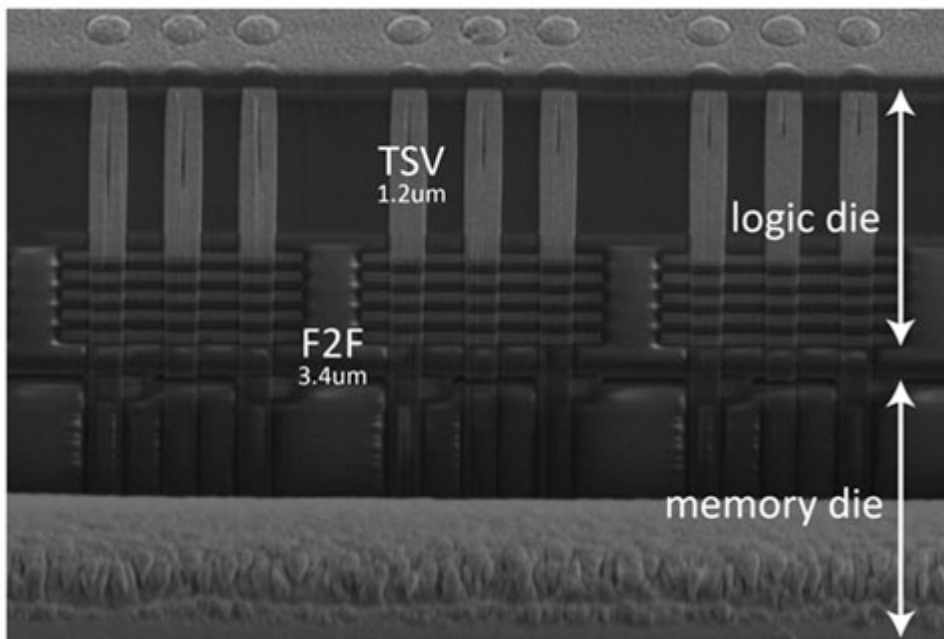
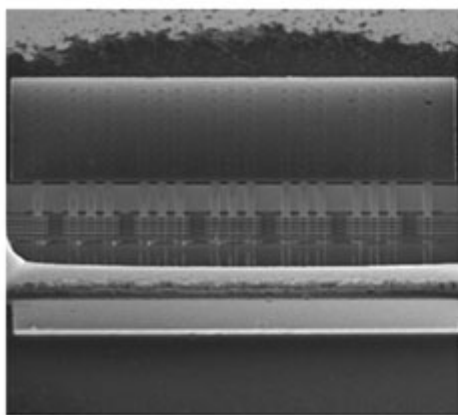
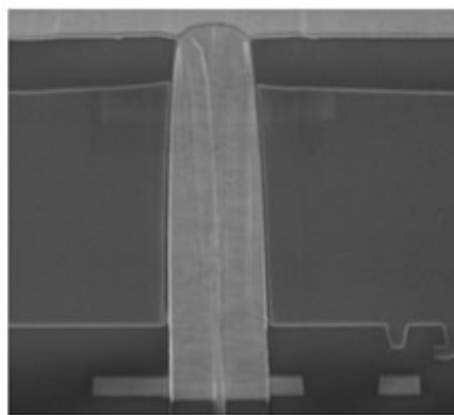


圖5 每個I/O Cell都有204個冗餘的TSV

圖6左為每一個I/O Cell與其下方對應的諸多TSV，圖6右則為單獨一個TSV與落地接觸點(Landing Pad)的剖面圖。圖7左為此顆3D IC的實體照片，此為上層IC的背端。在圖7右中間白色部分是所謂虛擬(Dummy)TSV，此因為TSV是由Tezzaron所完成，該公司對於TSV有類似互補式金屬氧化物半導體(CMOS)製程的布局密度(Layout Density)需求。可以看到TSV有開路錯誤，但因有做冗餘，所以不致對於整體有影響。



single IO cell, TSVs, BEOL of core die



single TSV and its landing pads

圖6 (左)每一個I/O與其對應的TSV，(右)單獨一個TSV與落地接觸點的剖面圖。資料來源：喬治亞理工學院

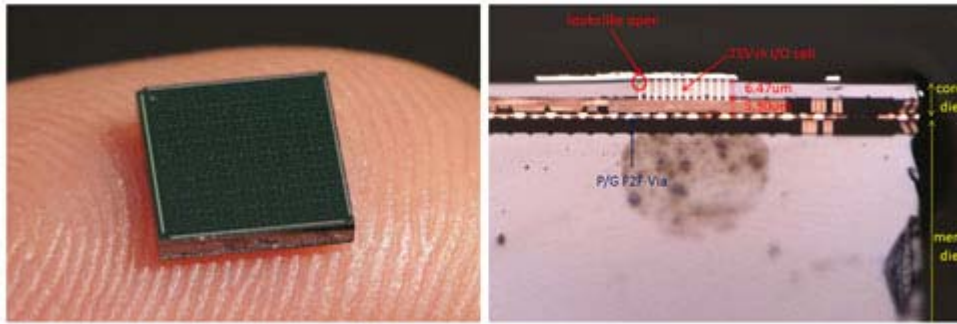


圖7 (左)3D MAPS實體照片，(右)看來有一個TSV有開路錯誤，但是因為有做冗餘，所以對於整體功能沒有影響。資料來源：喬治亞理工學院

喬治亞理工學院的另一個研究團隊教授為李憲信，同樣對於3D IC相當有研究。圖8是該團隊針對一個3D DRAM堆疊於邏輯上的可能架構所做的分析。基本上，此記憶體系統都是由一些記憶體區塊(Tile)所組成，整體系統共有一百二十八個 DRAM區塊。假設每個區塊有 $256M=28 \times 220$ 位元DRAM，每個DRAM區塊的架構上為二百一十四個列(Row)，而每個列有二百一十四個位元，所以整體3D記憶體可高達 $27 \times 28 \times 220 = 25 \times 30 = 32\text{Gbits} = 4\text{GB}$ 。根據這樣的記憶體容量與3D架構的需求，研究出可能的架構設計為以下三項：

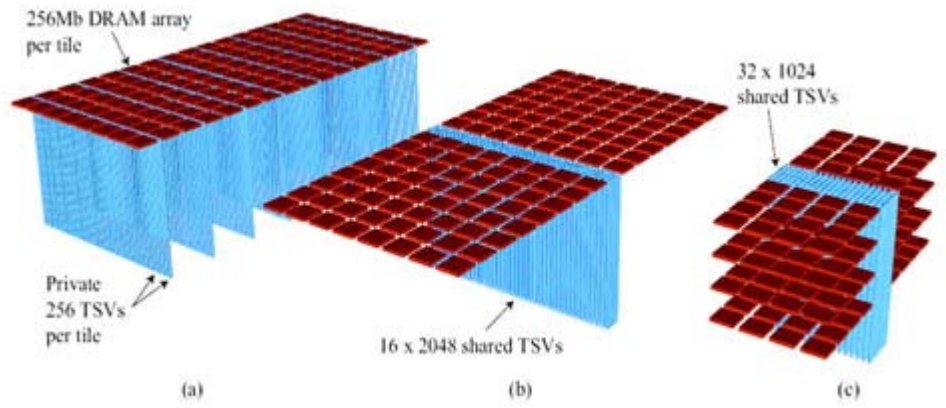


圖8 記憶體堆疊於CPU上的3D記憶體的不同形式

- 個別DRAM有個別的TSV

根據圖8a模擬，實際大小為29.36毫米×12.56毫米。

- 每層分享一叢TSV

根據圖8b模擬結果，面積需要約29.36毫米×12.77毫米，約比上個方式大一些。

- 所有的DRAM分為四層，再分享一叢TSV

根據圖8c的模擬，每一層的面積為14.68毫米×6.68毫米，所以四層的整體面積就會高達 $14.68\text{毫米} \times 6.68\text{毫米} \times 4\text{層} = 392.24\text{平方毫米}$ ，也就是雖然單一層面積減小，但是整體面積卻是變大。

圖9則是該團隊利用TSV的一種記憶體階層(Hierarchy)設計。整體系統由最下層到最上層分別為：處理器+L2(含TSV)→記憶體控制器(含TSV)→DRAM。其中，最下層的L2快取雖然是平面製程，但是會以64B一個子區塊(Subbank)的方式，共有六十四個TSV匯流排連接到記憶體控制器。希望透過TSV可以在L2與DRAM間同時提供六十四個 $64\text{B} = 64 \times 64 = 26 \times 26 = 4\text{KB}(1\text{Page})$ 的資料存取頻寬。亦即希望可一次傳輸就可完成

從DRAM到L2共4KB資料的傳輸寫回(Write Back)與快速填充(Cache Fill)，用以減少遺失。其實DRAM內部的這些資料會先同時到緩衝器/記憶體控制器，再以一次64B的方式填入L2的線(Line)。

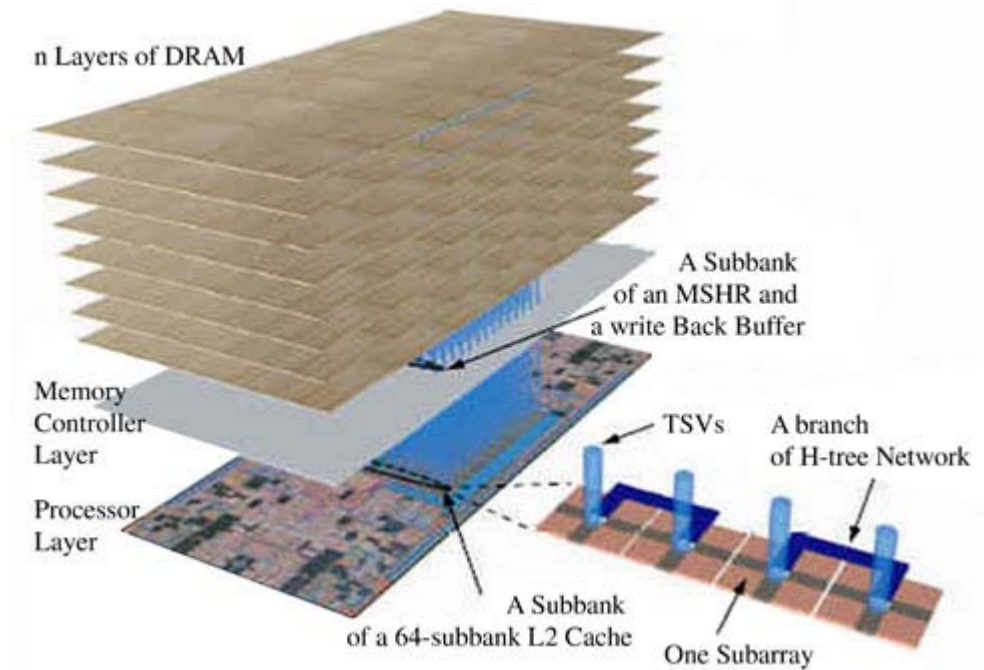


圖9 Smart-3D快取記憶體設計

### Tezzaron開發首款3D微處理器

早在2004年，Tezzaron已成功發表幾個利用該公司FaStack技術所設計的3D IC，採用0.18微米CMOS製程。第一個3D IC為一個CPU內常用到的暫存器，單純用來展示一個記憶體可以用3D堆疊完成。第二顆IC為一個超級8051(Super 8051)單晶片，產品編號為TSCR8051Lx；TSCR8051L2表示有128KB的SRAM；TSCR8051L3則表示有兩層128KB的SRAM，也就是共有256KB的SRAM；TSCR8051L5，表示有四層的SRAM，共有512KB的SRAM。這也是全球第一顆3D微處理器(MPU)，此單晶片是一個高速的8位元8051微處理器加上128KB的SRAM。這個8051可比其他8051速度快五至一百倍，可提供200MHz/200MIPS/100MFlops的運算速度。

### HRI 實現操作速度快/省電3D IC

日本的本田研究機構(HRI)在2008年也用200毫米、0.18微米CMOS製程，成功設計三層3D IC(圖10)，第一層為邏輯，第二層是類比數位轉換器(ADC)，第三層是64MB的DRAM。晶片的面積為8.44毫米×4.69毫米，在邏輯與ADC間有高達一千零五十六個TSV，整個晶片則有六十五萬五千五百八十四個TSV；ADC與DRAM間則有二百八十七個TSV，整個晶片則有十八萬零一百六十個TSV。整體厚度在60~100微米間。實驗結果顯示，此系統可比原有以MCM設計的系統提供二倍的操作速度，三分之一的功率消耗。

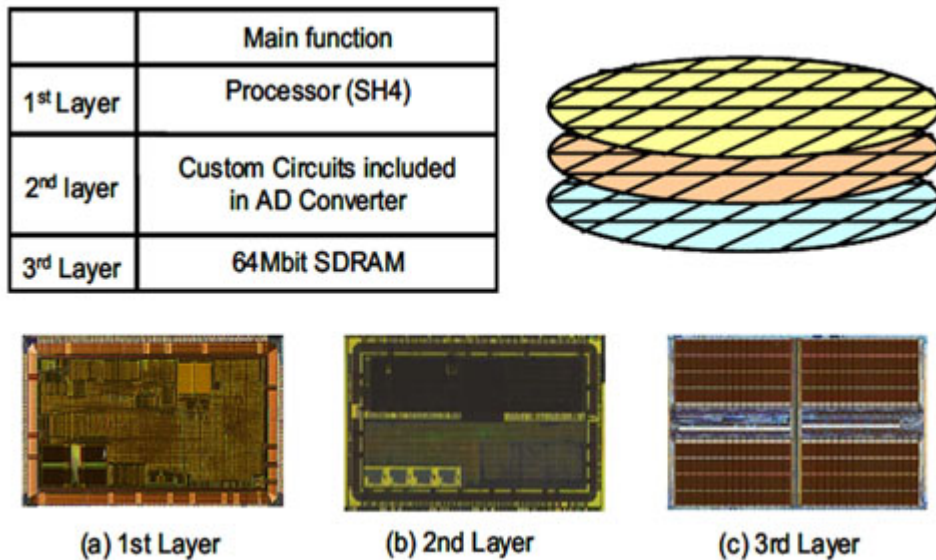


圖10 HRI的3D IC

### 英特爾展示80核心CPU堆疊

圖11說明英特爾八十核心的CPU堆疊構造，上層為處理器(含SRAM)，稱之為Polaris。每一個核心有256KB的SRAM，因為有八十個核心，所以共有256K×80=20MB記憶體。CPU是透過TSV與下層DRAM相接，此層稱之為Freya。TSV的數目高達八千四百九十個，TSV的間距為190微米。下方的記憶體則透過Cu C4 (Controlled Collapse Chip Connection)碰撞與外界相連。記憶體的晶片厚度約為20~100微米，CPU的厚度則是300微米。

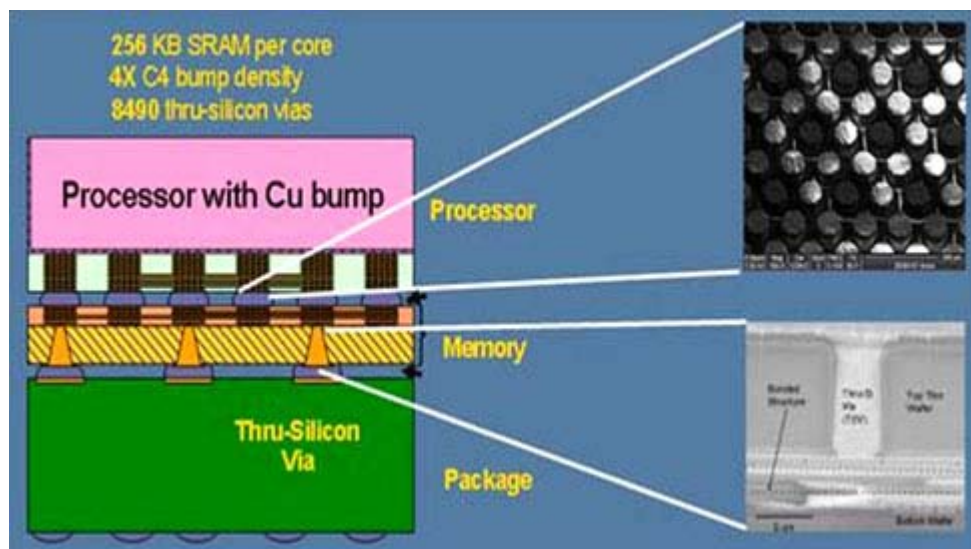


圖11 英特爾的80核心CPU/3D IC

### Alchimer由PoP轉向擁抱3D IC

Alchimer在2010年針對一個過去以封裝級封裝(PoP)封裝的手機IC改以TSV來實現。該原有的PoP封裝內包含ARM 11的MPU(65奈米, S3C 6410)、2GB儲存型快閃記憶體(NAND Flash)及1MB的DRAM；全部IC總共有五百個輸入/輸出，一半用於電源與接地，一半用於訊號傳遞。另外需要八十個內部連線連接三個IC，所以整體內部的訊號

線約為三百條。

新的設計使用TSV與堆疊技術，特性如處理器於上面，兩個記憶體在下面；TSV密度為每平方毫米十六個；處理器的面積為8毫米×8毫米，所以最多可以放 $8 \times 8 \times 16$ 個TSV；訊號TSV有三百三十個，電源與接地TSV的數目有六百六十個(根據Alchimer的經驗，電源與接地的TSV為訊號TSV數目的兩倍)，所以約使用一千個TSV。

該公司實驗三種不同的深寬比(AR)，亦即5：1、10：1與20：1。在相同的厚度情形下，AR越高，其所占的比例越小(12.3%、3.1%、0.8%)，基本上這與AR的加大比例成平方比，換言之，10：1的TSV會是5：1的TSV占晶片面積的四分之一。該公司認為20：1的TSV比5：1的TSV在300毫米的晶圓片上，可協助晶片設計者省下高達700美元的製造費用，此因20：1的TSV製造費用雖然高，但其減少的面積卻可製造更多的晶片。

(本文作者為南台科技大學電子系教授)

### 參考資料

1. The 3D-MAPS Processors, Available At: <http://www.gtcad.gatech.edu/3d-maps/>, 2011
2. Tezzaron, Tezzaron Announces Commercial 3D ICs, Available At: [http://www.tezzaron.com/about/Press/0404\\_3dic.html](http://www.tezzaron.com/about/Press/0404_3dic.html), 2008
3. C. Truzzi and S. Lerner (Alchmier), Electrografting: Unlocking High-Aspect-Ratio TSVs, Future Fab International, Vol. 31, No. 2, pp. 93-98, Oct. 2010

<http://www.mem.com.tw>