中国科学 E 辑：信息科学 2008 年 第 38 卷 第 6 期：933 ~ 940
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

# MPP 系统芯片体系结构技术的发展

沈绪榜

西安微电子技术研究所，西安 710054
E-mail: shenxubang@163.net

收稿日期: 2007-12-26; 接受日期: 2008-02-19

**摘要** 首先讨论芯片体系结构的演变，然后，分析 3 种计算模式的 MPP 系统芯片体系结构，在此基础上，提出统一改变的阵列处理器体系结构，同时实现了数据并行算法与非数据并行算法编程的简单性、高效性与通用性.

**关键词**
MPP
系统芯片
阵列处理器
体系结构

芯片体系结构的演变，是与计算机的体系结构的发展史密不可分的. 直到如今，计算机都是按照 von Neumann 体系结构的计算过程数字化，依照一定的程序进行运算，并按照电子学工作原理来设计的. 1958 年 Kilby 研制成功世界上的第 1 块芯片，从此，计算机体系结构就是在芯片体系结构不断创新的过程中发展的.

## 1 芯片体系结构的演变

芯片体系结构(chip architecture)是随应用需求与芯片集成度的提高而演变的[1]，如图 1 所示，是提高器件密度(device density)与计算模式(computing paradigm)的桥梁. 1971 年的微处理器芯片的问世，使原来体现在计算机设计上的指令集合体系结构(ISA, instruction set architecture)设计，就完全转到了微处理器芯片的指令集合体系结构设计上，是芯片体系结构的转折点. 按照Moore定律处理器芯片的性能每两年提高 1 倍，但还不能跟上应用要求的发展速度，1990 年代中期, Intel公司研制了首台由 1000 个处理器组成的万亿次计算机Option Red系统; 2000 年, IBM公司的 10 万亿次的Option White MPP计算机，有 8000 个处理器; 而 2007 年全球高性能计算机 500 强排名第 1 的IBM蓝色基因/L(Blue Gene/L)计算机，处理器数目超过了 13 万(131072). 相应地多核处理器(multi-core processor)悄然兴起. 例如，双核处理器芯片有IBM公司的Power6(2006)[2]，Intel公司的Conroe(2006)，AMD公司的Opteron(2004)[3]，Sun公司的Ultra SPARC IV(2004)等. 2007 年四核处理器芯片也已经问世. 多核处理器的芯片体系结构从串行计算走向了并行计算. 其实，早在 1987 年人们就提出了系统芯片(SoC, system on chip)的概念，研究如何进一步将计算机系统的设计转移到芯片设计上来，用系统芯片计算机取代由处理器芯片、存储器

芯片、接口电路芯片与传感器芯片等 4 种单功能芯片组成的现代计算机.

系统芯片概念提出 20 年后，已逐渐形成了正在走向市场化的两种系统芯片的发展途径. 一种是以现有单功能芯片组成的 PCB 系统为基础而发展起来的，由异构的多处理器与存储器、接口电路和传感器等总线互连的 MP 系统芯片(MP SoC, multi processor SoC)；另一种是根据并行计算机技术与深亚微米技术的发展需要而发展起来的，仅由许多同构的处理元 PE 组成的，体现高性能应用的并行性与高集成度发展的规则性，以网络作为互连“总线”的大规模并行处理系统芯片(MPP SoC, massively parallel processing SoC)，简称 MPP 系统芯片. 因为采用了 PE 阵列的实现方法，又叫做阵列处理器(AP, array processor)，使处理器的芯片体系结构从多核处理器，进一步走向 MPP 系统芯片的阵列处理器.

人们估计到2010年后，基于光刻技术采用SiGe的CMOS工艺的制造能力将达到它的30nm极限[4]，如图 1 中所示，将出现所谓的红墙(red wall)问题. 一是线的延迟比门的延迟越来越重要，在过去的技术中，导线是用来连接逻辑门的，在今后的技术中，情形则相反，导线是用逻辑门来连接的，芯片上的互连线不仅有传输延迟问题，而且还有能耗问题，为此，处理元与处理元之间，处理元与存储器之间，在设计上要采用局部通信技术；二是特征尺寸已小得使芯片缺陷不可避免，设计上要从缺陷容忍，故障容忍与差错容忍等 3 个方面研究避错的自主重构技术；三是漏电流与功耗变得非常重要，当特征尺寸小于 65 nm之后，静态功耗将超过 50%,设计上要采用功耗的自主管理技术. 如何解决这些技术上的“红墙”问题推动了MPP系统芯片的体系结构研究和发展. MPP系统芯片体系结构是芯片集成度发展到上亿晶体管之后，芯片体系结构也要从功能设计上采用更规则的体系结构的必然趋势，并且形成了基于指令流计算模

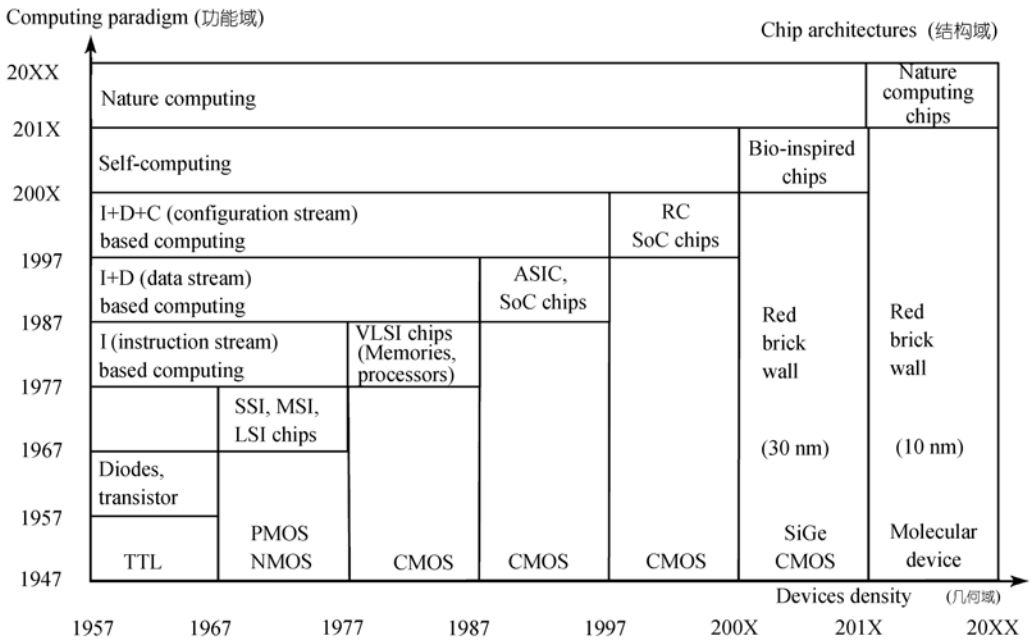


**图 1　芯片体系结构随集成度提高与计算模式发展的变化情形**

式、数据流计算模式与构令流计算模式的三类计算模式的体系结构, 进一步的发展将是自主计算模式与自然计算模式的体系结构[5,6], 如图 1 中所示.

## 2 指令流计算模式的 MPP 系统芯片体系结构

按照 1966 年Flynn的分类[7], 主要有两种指令流计算模式的并行计算的体系结构, 一是数据级并行的SIMD(single instruction multiple data)体系结构, 其控制部件(control unit)发送同样的指令流到每个PE(processing element), 对不同的数据完成相同的运算; 二是指令级并行的MIMD(multiple instruction multiple data)体系结构, 每个处理元PE都有它自己的控制部件而成为处理器, 每个处理器彼此独立地对不同的数据执行不同的指令, 如图 2(a)中所示. 数据并行算法的程序设计是适合于在SIMD体系结构上完成的, 但非数据并行算法的程序设计还没有合适的体系结构, 是在MIMD体系结构上完成的, 带来了程序设计的复杂性.

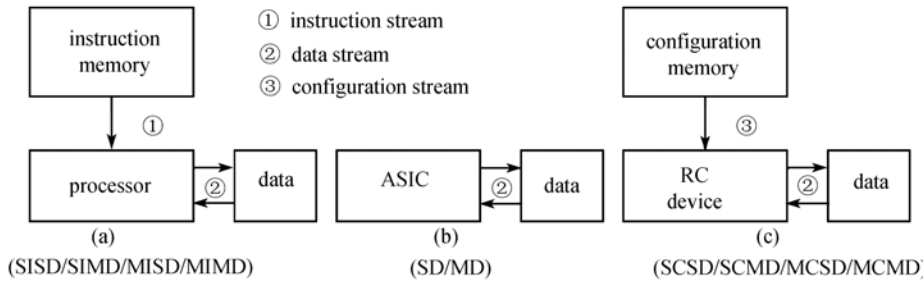


**图 2 3 种计算模式及其 10 种体系结构[8]**

(a) 指令流计算模式; (b) 数据流计算模式; (c) 构令流计算模式

早在 1980 年, 国外针对航空航天遥感图像处理的数据并行计算的特点, 研制出了许多图像处理的 SIMD PE 阵列体系结构. 1976 年英国 ICL 公司研制了 64×64 PE 阵列(processing element array)的 DAP(distributed array processor)计算机. 1980 年英国伦敦大学学院推出了一种 96×96 PE 阵列的 CLIP4(cellular logic image processor)计算机. 1983 年美国 Goodyear Aerospace 公司研制了一种 128×128 PE 阵列的 MPP(massively parallel processor)计算机, 用来分析处理航天飞机发回的地貌图像. 1987 年美国 Thinking Machine 公司推出了 CM-2, PE 阵列中可有 65536 个处理元 PE, 主要用于人工智能和图像处理. 这些图像处理的 SIMD PE 阵列是一些应用针对性很强的 MPP 系统芯片, 充分体现了芯片体系结构随芯片集成度提高与计算模式发展的变化特点.

用于加速图形处理运算的nVIDIA公司的Geforce 8800图形处理器(GPU, graphic processing unit)[9], 采用了SIMD PE阵列的体系结构. GPU的前身是图形加速芯片, 是一种加速台式电脑进行三维图形显示的专用处理器. 随着图形加速芯片从CPU接管的工作越来越多, 图形加速芯片从一种专用芯片演变成了今天的MPP系统芯片. 2006 年DirectX 10 规范的到来更是打开了GPU通用计算的大门. GPU的计算任务可以分为 5 类, 顶点渲染(vertex shading)、像素渲染(pixel shading)、几何渲染(geometry shading)、物理计算(physical calculation)、通用计算(general purpose calculation). GPU是由能完成所有 5 类任务的通用处理元PE所组成的PE阵列, 而不像

前几代产品为每一种计算设计独立的计算单元和流水线. Geforce 8800 系列GPU是第一个实现Directx 10 规范的GPU产品.

ATI的HD2000 系列GPU同nVIDIA的Geforce 8800 系列GPU一样[10], 也是支持Directx 10 规范的统一架构的GPU, 也是一种两维空间的MPP系统芯片, 有 320 个流处理元. 每 80 个PE组成一个SIMD PE阵列, 共有 4 组. 每个SIMD PE阵列包含 16 个SIMD处理元, 所以一个SIMD处理元包括 5 个PE.

圆片规模集成(WSI, wafer scale integration)技术是通过扩大芯片面积来提高芯片集成度的一种新途径. 例如, 欧洲四国的ELSA计划的主要目标, 就是采用 2-D WSI技术, 研制一种集成大量处理元PE阵列的处理器[11]. 将一个大圆片分成 20 个方格区域, 每个方格又分成 4 个子区域, 子区域之间用标准开关和缓冲开关连接, 每个子区域上有 7×12 个一位的处理元PE. 其中, 6×12 个PE为实用的PE, 另外的 12 个为备用的PE, 所以, 每个大圆片可集成 6720 个一位的PE(其中 960 个备用).

对于非数据并行算法, 在当前的高性能计算机中, 采用的是单核或多核的高性能处理器, 处理器之间采用共享路由器的复杂的线接技术. 在指令流计算模式的 MISD 体系结构中, 指令流水线的实现方法是行之有效的, 但流水线的深度, 也就是空间并行度是有限的; 而在 MIMD 体系结构中, 虽然也有空间上的一维二维或三维的并行实现技术, 但受算法本身内在并行度低的限制, 只有 VLIW 与 SMT 等指令级空间并行度低的实现技术. 在这些指令流计算模式的体系结构中, 需要通过软件将一个大的处理任务划分为若干个子任务, 由分开的处理器执行, 编程相对复杂.

实践表明, 自动并行识别技术是不成功的. 在现代的并行计算机系统中, 非数据并行算法的 MPP 计算的任务划分主要是靠操作系统的任务调度(task scheduling)程序管理的. 美国国家科学基金会认为, 高性能计算机正处于重要的转折期, 从 2001 年开始启动一系列项目, 鼓励“革命性体系结构概念”的研究. 从当前支持高性能计算的 MPP 系统芯片体系结构的变化来看, MPP 系统芯片上的处理元 PE 要比高性能处理器核简单很多, PE 之间采用分布路由器的简单的邻接技术. 不仅嵌入式计算机是要面向应用领域优化设计的, 而且今后的高性能计算机也是要面向应用领域优化设计的.

1999 年 12 月 6 日IBM宣布将耗资 1 亿美元研制代号为“蓝色基因”(blue gene)的用于模拟计算的高性能计算机, 通过对各种蛋白质分子聚合到一起的多种力量加以测量, 来研究人类蛋白质分子的折叠方式. 根据 2007 年全球高性能计算机 500 强评选的排名, IBM的蓝色基因/L(blue gene/L)排名第一, Linpack基准测试的峰值速度为 280.6Teraflops(1 Teraflops=$10^{12}$ flops, 万亿次); IBM估计, 2007 年底将建成蓝色基因/P(blue gene/P), 拥有 100 万个处理元, 峰值速度为 1 petaflops (1 Petaflops=$10^{15}$flops,千万亿次). 有 64 个 6 英尺高的机柜, 每个机柜有 8 块主板, 每块主板上有 64 个MPP系统芯片, 每个芯片上含有 32 个简单的处理元, 每个处理元的运算速度是 10 亿次/秒[12~14].

Intel提出了一个名叫Tera-Scale的计划, 2007 年初采用 65 nm工艺的MPP系统芯片, 总共集成了 1 亿(100 Million)晶体管, 一共采用了 80 个比现代处理器简单的处理元PE, 每个PE有单独

的路由器, 每个路由器有 5 个通信路径, 带宽达到 80 GB/s, 延迟为 1.25 ns, 其中 4 个通信路径用来形成一个二维的格状网络, 实现PE之间的数据通信; 另一个用于连接重叠的SRAM,实现PE与存储器之间的数据通信, 如图 3(a)所示, 是一种网格互连的MPP系统芯片. PE与SRAM互连的 3-D 二次工艺集成技术是一种缩短连线的途径, 很有发展前途[15~17]. 这些芯片最终实现的性能如表 1 所示, 每种MPP系统芯片的性能都在 1 Teraflops以上.

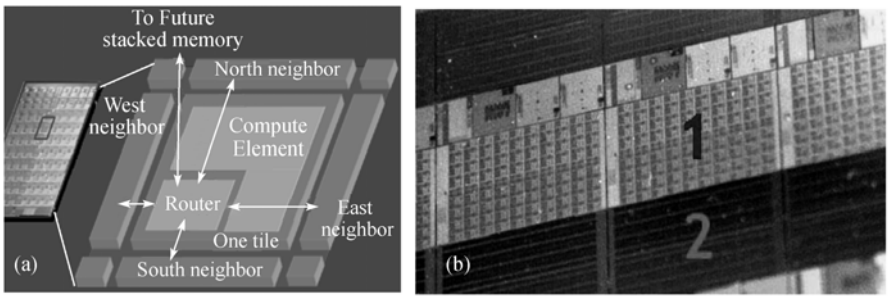


**图 3 Intel 的 MPP 系统芯片**

(a) PE 互连结构图; (b) PE 阵列

**表 1 Intel MPP 系统芯片的性能**

| Frequency/GHz | Voltage/V | Power/W | Aggregate bandwidth | Performance |
|---|---|---|---|---|
| 3.16 | 0.95 | 62 | 1.62 Terabits/s | 1.01 Teraflops |
| 5.1 | 1.2 | 175 | 2.61 Terabits/s | 1.63 Teraflops |
| 5.7 | 1.35 | 265 | 2.92 Terabits/s | 1.81 Teraflops |

## 3 其他计算模式的阵列芯片体系结构

由于指令流计算模式中的MISD/MIMD体系结构, 对实现非数据级并行算法的空间并行度的不足, 随着应用要求与芯片集成度的进一步提高, 在现在的控制流体系结构中, 提高计算机性能的办法, 已从图 2(a)的指令流计算模式的体系结构发展到了数据流计算模式的体系结构, 如图 2(b)中所示, 出现了按算法设计的ASIC电路, 例如, Systolic Array[18], 是一种两维的功能模块的阵列芯片体系结构, 提高了计算的空间并行度. 通过数据流计算模式的ASIC电路, 每个执行周期能完成一次算法的计算, 实质上这些电路就成了基于数据流计算模式的非数据并行算法的算法处理器, 具有并行计算的高效性, 但存在专用性的缺点. 为了解决ASIC电路的多次可应用性, 出现了静态可重构的FPGA电路, 是一种两维的门阵列芯片体系结构, 但还是没有程序设计的灵活性. 为了克服数据流计算模式的阵列芯片体系结构的这个缺点, 人们又提出了构令流计算模式的体系结构, 如图 2(c)中所示, 出现了通过构令(configuration)动态可重构的RC device(re-configuration device)电路, 是一种功能模块的阵列芯片体系结构.

## 4 统一改变的阵列处理器体系结构[19]

综上所述, 有 3 种计算模式, 共 10 种体系结构, 如图 2 中所示. 但是, 后两种计算模式的体系结构虽然都具有 MPP 的特点, 但都要求使用者从硬件的逻辑设计上, 而不是从软件的程序设计上来实现应用算法到体系结构的映射的, 降低了应用设计的抽象层次. 从解决实际工

程问题出发, 这 10 种芯片体系结构, 可以多至 1023 种可能的体系结构组合. 为了使 SIMD PE 阵列同时对数据级并行算法与非数据级并行算法都能有效映射, 并行编程简单, 最终能取代 ASIC, FPGA 与 RC Device 等大规模集成电路芯片, 在找出这 10 种体系结构的共同点的基础上, 我们提出了统一改变的阵列处理器体系结构, 以减少芯片体系结构设计的可能组合, 实现并行编程的统一性、简单性、高效性与通用性.

指令流计算模式的体系结构是通过程序设计(改变)来实现不同计算的, 是应用设计(改变)上最灵活的一种体系结构, 现代的微处理器都采用了 ISA 模型. 数据流计算模式的体系结构是采用全定制的 ASIC 或半定制的 FPGA 芯片, 通过芯片设计(改变)来实现不同计算的. 构令流计算模式的体系结构, 是通过流件(stream-ware)设计(改变)可重构电路的结构, 来实现不同计算的. 这些不同计算模式的体系结构是通过不同的应用设计(改变)方式来实现计算的. 从而造成了应用设计(改变)与电路设计(实现)的多样性.

从映射方式上来看, 指令流计算模式的体系结构是一种时间映射的体系结构, 而数据流计算模式与构令流计算模式的体系结构是一种空间映射的体系结构. 为了提高 MISD 与 MIMD 体系结构的并行计算能力, 可以将它们从按时间映射的指令流计算模式, 改变成在 SIMD PE 阵列上按空间映射/时空映射的数据流计算模式, 如表 2 中所示, 这样一来, 就可以在 SIMD PE 阵列上, 建立时空映射的阵列处理器芯片的体系结构, 使 SIMD PE 阵列也能成为一个非数据并行算法的算法处理器. 将数据流计算模式与构令流计算模式上的应用设计(改变)的抽象层次, 从芯片设计层次或逻辑设计层次都提高到程序设计的层次, 降低应用设计的门槛, 使所有计算模式的应用设计都是通过统一的程序设计(改变)完成的, 从而能消除应用设计(改变)方式的多样性与实现的多样性, 实现指令级并行计算的编程简单性、高效性与通用性.

**表 2　体系结构的映射方式分类**

| mapping paradigm | data stream | |
|---|---|---|
| | sigle data | multiple data |
| temporal mapping | SISD | SIMD |
| | MISD | MIMD |
| spatial mapping | MISD | MIMD |

## 5　结束语

统一改变的阵列处理器体系结构, 也是在 von Neumann 的指令流计算模式的体系结构基础上发展起来的. 如图 4 中所示, 指令集合(体系结构)的 IC(实现)化就是处理器; 算法(应用)的指令集合(体系结构)化就是目标程序; 算法(应用)的直接 IC(实现)化就是 ASIC 电路, 或者是可重构的 RC device 电路; 而 PE 阵列就是这 3 种实现的并行化发展. 当 MISD/MIMD 体系结构在 PE 阵列上的空间映射实现 ASIC 电路或 RC device 电路时, 便得到统一改变阵列处理器, 如图 4 中间所示.

由于统一改变的阵列处理器可以用来将 ASIC 电路与 FPGA 电路的应用设计统一为“指令(语句)阵列”的程序设计, 将 RC device 电路看作由多套 ASIC 组成的, 它的应用设计也统一为多套“指令(语句)阵列”的程序设计, 这样一来, 就可以使 ASIC, FPGA 与 RC Device 电路的芯片设计都转化为粗粒度或细粒度的 PE 阵列设计.

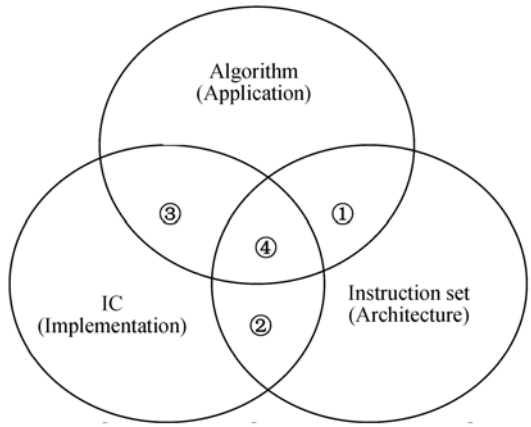


**图 4 Evolution of array processor**

① Program; ② Processor; ③ ASIC/RC Device; ④ Array Processor

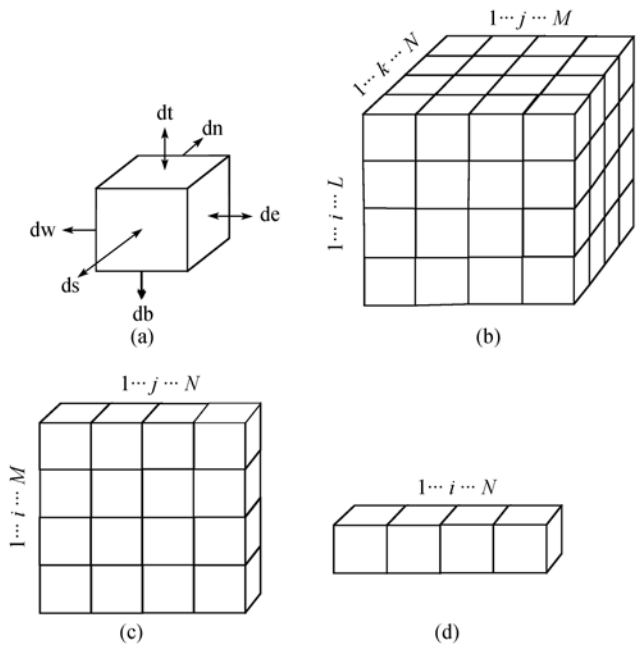


**图 5 AP 的 PE 阵列系列化的示意图**

(a) 1×1×1 的 PE 单元与数据接口; (b) $L\times M\times N$ 三维 PE 阵列; (c) $1\times M\times N$ 的二维 PE 阵列; (d) $1\times 1\times N$ 的一维 PE 阵列

科学和艺术都是探索四维的时空关系的, 统一改变的阵列处理器是用来探索四维的时空并行计算关系的. 物理世界是三维的, 所以, 对应的阵列处理器 AP 自然是三维的好, 如图 5(b)中所示; PE 阵列中的处理元 PE 的数据接口应当是三维的, 处理元的路由器有东南西北的通信路径(dn, data north; de, data east; dw, data west 与 ds, data south)和上下的通信路径(dt, data top 与 db, data, bottom)等 6 个, 如图 5(a)中所示. 当前计算机的显示器输出与传感器输入是空间上的两维阵列, PE 阵列也是两维的, 如图 5(c)中所示, 在一些简单的应用中甚至是一维的, 如图 5(d)中

所示. 艺术家是在两维的画面上加上透视、阴影, 而产生视觉上的三维立体感的, 类似地, 计算机专家是通过立体视觉算法在两维阵列的显示器上形成三维的立体感的.

由于工艺技术的限制, 单个芯片上的 I/O 引脚数目不能随芯片集成度的提高而成比例地增长, 为了支持阵列处理器体系结构的四维时空探索, 从工艺实现技术上就要有三维的实现技术, 支持阵列处理器、阵列存储器与阵列传感器等各种阵列芯片的三维集成. 因为三维的一次集成技术是比较困难的, 到目前为止还只有松下电器工业有限公司研制过一次集成技术的 3-D 图像处理芯片; 而三维的二次集成技术是比较现实的. 前述 Intel 公司的 Tera-Scale 研究计划的 MPP 系统芯片, 就采用了芯片级的 3-D 二次集成技术. 图 3(b)中的"1"是 80 个处理元 PE 部分, 而"2"则是 2M 字节的二级 SRAM 缓存, 这两种芯片采用了芯片规模的 3-D 二次集成技术, 将处理元 PE 与 SRAM 芯片采用立体叠加结构的方式连接. 同采用 PIM(processor in memory)方法相比, 3-D 二次集成技术, 也解决了所谓"存储墙(memory wall)"问题.

## 参考文献


1 Manners D, Makimoto T. Living with the Chip. London: Chapman & Hall, 1995
2 Le H Q, Starke W J, Fields J S, et al. IBM POWER6 microarchitecture. J Res Dev, 2007, 51(6): 639—662
3 AMD Corp. AMD Opteron™ Product Data Sheet. 2004
4 Ratner M A, Ratner D. Nanotechnology: a gentle introduction to the next big idea. New Jersey: Person Education, Inc, Prentice Hall, 2003
5 Macias N J, Durbeck L J K. Adaptive method for growing electronic circuits on an imperfect synthetic matrix. Biosystem, 2004, 73(3): 173—204 [DOI]
6 Adleman L M. Computing with DNA. Sci Am, 1998, 279(2): 54—61
7 Flynn M J. Very high speed computing systems. Proc IEEE, 1966, 54: 1901—1909
8 沈绪榜, 张发存, 冯国臣, 等. 计算机体系结构的分类模型. 计算机学报, 2005, 28(11): 1759—1766
9 NVIDIA Corp. NVIDIA GeForce 8800 Architecture Technical Brief. 2006
10 Persson Emil. ATI Radeon™ HD 2000 programming guide. AMD Graphics Products Report. 2007
11 Boubekeur A, Patry J L, Saucier G, et al. Lessons learnt from designing a wafer scale 2D array. In: Proceedings of IEEE Defect and Fault Tolerance in VLSI Systems 1992. New York: IEEE, 1992. 137—146
12 Allen F, Almasi G, Andreoni W, et al. Blue Gene: a vision for protein science using a petaflop supercomputer. IBM Syst J, 2001, 40(2): 310—327
13 Gara A, Blumrich M A, Chen D, et al. Overview of the Blue Gene/L system architecture. IBM J Res Dev, 2005, 49(2/3): 195—212
14 IBM Blue Gene team. Overview of the IBM Blue Gene/P Project. J Res Dev, 2008, 52(1/2): 199—220
15 Held J, Bautista J, Koehl S. From a few cores to many: a Tera-scale computing research overview. Intel Tera-Scale Computing Research White Paper. 2006
16 Vangali S, Howard J, Ruhi G, et al. An 80-Tile 1 .28TFLOPS Network-on-Chip in 65nm CMOS. In: Proc IEEE ISSCC'07. New York: IEEE, 2007. 98—99
17 Wechsler O. Inside Intel core microarchitecture: setting new standards for energy-efficient performance. Intel Corporation White Paper. 2006
18 Kung H T. Why systolic architecture? Computer, 1987, 15: 37—46
19 沈绪榜, 刘泽响, 王茹, 等. 计算机体系结构的统一模型. 计算机学报, 2007, 30(5): 729—736