魏少军,清华大学微电子学研究所教授,主要研究方向为可重构计算和超大规模集成电路设计方法学

可重构芯片的方法学原理

魏少军

清华大学微电子学研究所,北京 100084

摘要　可重构芯片技术(又称软件定义芯片技术)是集成电路领域非常有希望的差异化技术,具有广泛的适用性。概述了动态可重构芯片的属性分类、潜在特点及其演进过程和方法学原理,总结了动态可重构芯片的若干难点及对应的核心技术,并展望了未来发展趋势。

关键词　可重构计算芯片;方法学原理;控制密集型任务;映射技术

可重构芯片具备软件、硬件双编程的特性,硬件架构和功能随软件变化而实时动态变化,因而又被称为软件定义芯片。可重构芯片的出现打通了“应用定义软件、软件定义芯片”进而实现“应用定义芯片”这一人们长期追求的通道,而广泛的适应性也使其成为替代专用集成电路、可编程器件和经典处理器的有力竞争者。可重构芯片技术的发展经历曲折。尽管可重构的概念早在20世纪60年代就被提出,但半个多世纪后才获得突破。可重构芯片最早的技术源头可追溯到20世纪80年代末诞生的高层次综合理论和方法。进入21世纪后,中国学者经过10多年的不懈努力,突破了一系列核心关键技术,成为可重构芯片领域的全球领跑者。了解可重构芯片的方法学原理,可深入理解这一全新技术路线的重要意义。

1　动态可重构芯片的属性分类及潜在特点

可重构芯片很多时候被误解成现场可编程门阵列(FPGA)。不仅在国内,国际上也是如此。在国内,对于这方面的理解显得更少一些。

为了说明可重构芯片的独特性,以软件可编程性和硬件可编程性为坐标轴,构建一个四象限的坐标图,分析现有的芯片均处在哪个象限。有意思的是,为人们所熟知的处理器,例如中央处理器(CPU)、数字信号处理器(DSP)等,处在第二象限,因为它们的软件可编程,硬件基本不可动。SoC(system-on-a-chip)及专用集成电路(ASIC)处在第三象限,而FPGA、可擦除可编辑逻辑器件(EPLD)在第四象限(图1)。

第一象限中的芯片不仅软件可编程,还要硬件可

引用格式:魏少军.可重构芯片的方法学原理[J].科技导报,2019,37(3):69-76;doi:10.3981/j.issn.1000-7857.2019.03.011

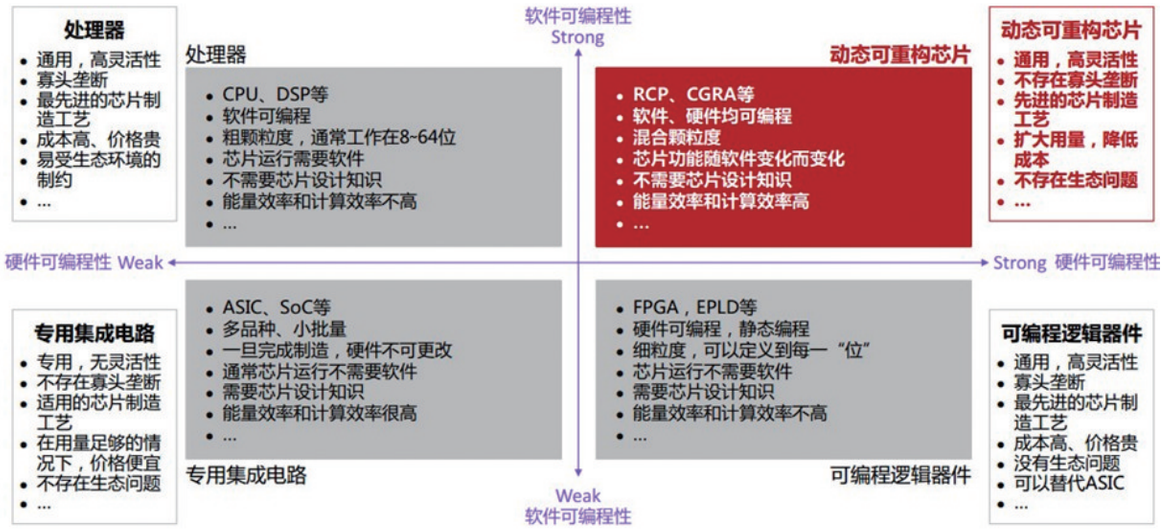


图1 动态可重构芯片的属性分类

编程,甚至还要硬件在软件编程下可编程。显然,现有的芯片均不属于此类。第一象限可以归纳为动态可重构的芯片,有时称为RCP或CGRA等,其特点是:软件硬件均可以编程,混合粒度,芯片的硬件功能随软件变化而变化,应用改变软件、软件再改变硬件。而且它与CPU等处理器有很多类似的地方,开发者不需要底层芯片设计知识。这类芯片与ASIC一样,具备很好的能量效率和计算效率等。显然,这种芯片的属性分类与以前不同,不可以将其与FPGA等混为一谈。

动态可重构芯片有别于传统芯片的预期特点和潜在能力可总结为:(1) 软硬件可编程;(2) 硬件架构的动态可变性及高效的架构变换能力;(3) 兼具高计算效率和高能量效率;(4) 本征安全性;(5) 应用简便性,不需要芯片设计的知识和能力;(6) 软件定义芯片;(7) 实现智能的能力。

其中软件定义芯片的功能值得重点关注。如果能够实现“应用定义软件,软件又能定义芯片”就等效于应用可以定义芯片。打通了这个链条,就成为一种“通用的专用芯片”,既具备了通用性又具备了专用性。

另外一个值得重点关注的是实现智能的能力。不难理解,设计规格定义的芯片差异化只存在于产品产出的初期;一旦芯片安装于设备,其差异化就不再增加;随着时间的推移,差异化只会越来越小。因此,需要研究如何以芯片可以理解的方式实现对芯片的“教育”,以及芯片在接受“教育”的过程中如何能够实现“学习”。如果芯片在使用过程中可以通过“教育”不断地自我“学习”并改进,则差异化可以不断增强。因此,芯片应该具备学习的能力、架构不断变化的能力和功能不断提升的能力。

2 动态可重构芯片的演进过程及方法学原理

20世纪80年代,采用硬件描述语言(HDL)进行芯片设计是集成电路设计方法学的一大进步。而如何最优地实现一款符合HDL设计描述定义的芯片,则是设计方法学要解决的核心问题。显然,在不考虑实现的代价和复杂性的前提下,对该HDL描述而言,一个拓扑结构和与HDL描述中相应的运算——对应的硬件方案,是最直接的实现,性能效率也最高。早先人们在讨论这一问题时,首先有图2(b)的电路图,再去想象应该如何描述它,即图2(a)的描述。但是,随着集成电路设计方法学的发展,从描述出发设计一款功能相同的电路成为主流,这成就了逻辑综合等从HDL进行芯片设计的基本想法。只是这个基本想法并非如一般人想象的那么简单,中间还有一整套方法需要考虑。

不难理解,图2(b)电路图可以实现图2(a)描述的功能,但这个电路并不是一个足够优化的结果,包含了3个乘法器、2个除法器、2个加法器和3个减法器等。每一个运算器价格都很昂贵,需要考虑如何对该结构进行优化。

20世纪80年代末、90年代初形成的高层次综合

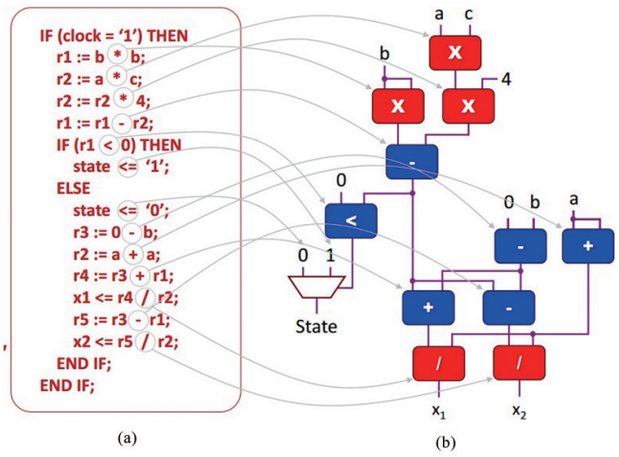


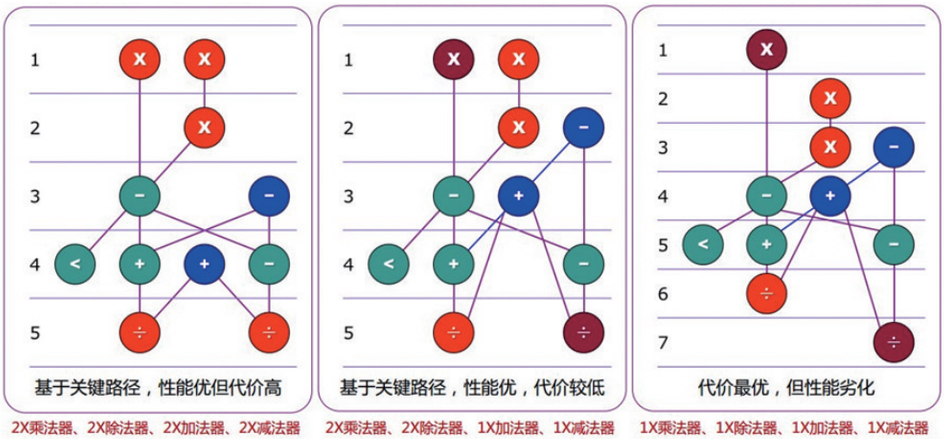
图2　硬件描述语言与硬件实现方案

(high-level synthesis)理论和方法就是一种从行为描述到电路的优化设计方法。首先要找到数据依赖关系，当有了数据依赖关系以后，可以通过运行时间上的分割，对运算进行调度实现计算资源的复用。

如图3所示，可以将一个计算的执行过程分成若干个时间间隔，凡是不在同一个时间间隔执行的运算，就是不同时发生的，因此上下之间就可以共享同一个资源，称为资源的复用。在图3(a)的方案1中各用了两个乘法器、除法器、加法器、减法器。

通过对数据依赖关系的分析，可以发现图3(a)中有两个运算器(1个加法器和1个减法器)与之前的任何运算均无依赖关系，可以把它们向前移动一下，结果就减少了1个加法器、1个减法器，最后就可以只需用2个乘法器、2个除法器、1个加法器、1个减法器，如图3(b)所示。当然，硬件资源的使用还可以进一步减少，但是会导致关键路径增长，性能劣化(图 3(c))。

通过上述过程不难发现，一个电路的实现方案存在多种可能性，可以用2个乘法器、2 个除法器、2 个加法器和2 个减法器来实现，也可以只用1个乘法器、1个除法器、1个加法器和1个减法器来实现。通过巧妙地安排运算的时间以实现资源的复用，这就是算子调度。



(a) 方案1　(b) 方案2　(c) 方案3

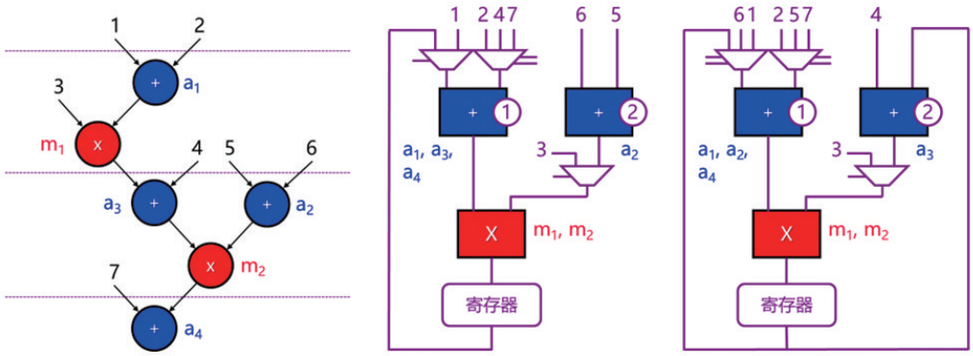
图3　面向资源复用的算子调度

当算子调度完成后，要进行资源的分配。当在一个时间间隔中有多个同类型资源同时出现时，它们与运算的捆绑方式不同，资源的分配方式也不同，造成的整个互连结果不同。如图4所示，由于加法操作捆绑到加法器的方式不同，图4(b)将a_1、a_3和a_4捆绑到加法器1，a_2捆绑到加法器2；图4(c)将a_1、a_2和a_4捆绑到加法器1，a_3捆绑到加法器2；对应地产生了不同的互连网络结构，因而多路器的复杂度也不同。

在互连结构产生后，可以生成微控制码。将微控制码按照所在控制步逐一施加于对应的控制端口，就可以得到所要实现的功能(图 5)。

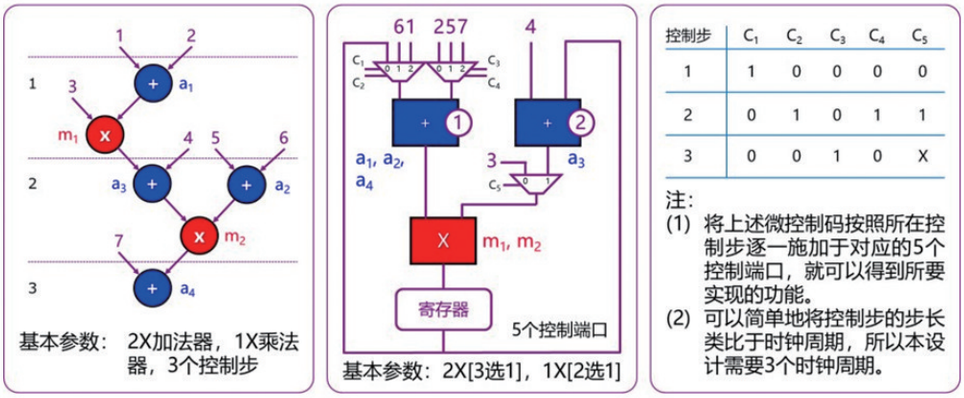
需要指出的是，资源分配与互连方案中给出的多路选择开关的控制方案并不唯一，可以有多种选择。因此，不同的多路选择开关的控制方案将导致不同的微控制码。

图6所示电路架构由专用数据通道和专用微控制器构成，显然这样的电路就是个专用集成电路。这就是20世纪八九十年代发展起来的高层次综合设计生成专用集成电路的基本理念。高层次综合系统实现过程为：系统输入用HDL写成的系统行为描述，例如VHDL



（a）调度方案 （b）资源分配方案1 （c）资源分配方案2

图4 面向互连网络优化的资源分配



（a）算子调度方案 （b）资源分配与互连方案 （c）微控制码方案

图5 面向控制单元优化的微控制码生成与优化

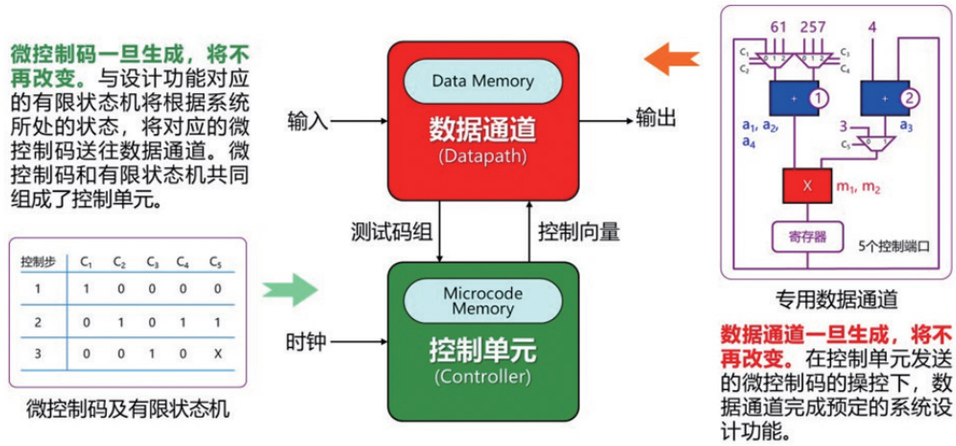


图6 高层次综合生成的专用集成电路架构

或Verilog;根据这些行为描述,通过高层次综合的编译器,生成包含数据和互连网络配置信息的微控制码以及与系统功能相关的有限状态机。这里虽然借用了“编译器”这个名词,但实际上它与传统的计算机编译器并没有关系,其核心是一整套高层次综合方法学的内容。高层次综合系统使设计过程变得非常有序,是20世纪八九十年代集成电路设计方法学中最好的选择。

但是,随着集成电路工艺技术进步到14 nm或者更小的特征尺寸,就会发现专用集成电路的研发成本太高了。在14 nm节点,研发一个芯片的综合成本高达1.5亿~2亿美元,通常要销售3000万颗以上,才能将研发成本合理地摊销到每颗芯片上。但是以多品种、小批量为特色的专用集成电路(ASIC)销量不高,难以有效摊销高昂的研发成本。ASIC面临巨大的挑战,难以为继。

如何解决这个问题,最好的办法是找到一种复用的方法,不是简单的复用一个资源,而是复用一个芯片。设想一下,如果只生产一种"通用"的芯片,其功能可以通过软件改变,当不同的软件写入就变成了"专用"的芯片,这将是非常理想的情况。如果这个想法能实现,可以认为软件定义芯片成为现实。

但面对的一个挑战是软件可以无限复杂,执行时间可以无限长,而硬件不管多大都有边界。因此,直接的想法就是将一个软件按照硬件规模分块,形成一系列的子任务,并按照子任务间的依赖关系将其一块块送到硬件上执行。这就要求芯片的架构和功能上必须是动态可变的,可以按照软件实时的变化。这就是可重构芯片(软件定义芯片)的最基本的思路。

图7给出了可重构芯片的基本架构。可以看出,仍然采用图6专用集成电路基本架构,在控制单元添加软件,同时将数据通道变成一个通用数据通道,控制单元变成一个通用的控制单元。通用数据通道是一个二维的处理单元(PE)阵列。每个PE的功能可以根据需要自行定义;PE可以是单功能的,也可以是多功能的;PE阵列可以是同构的,也可以是异构的;PE阵列的运行采用数据驱动方式。

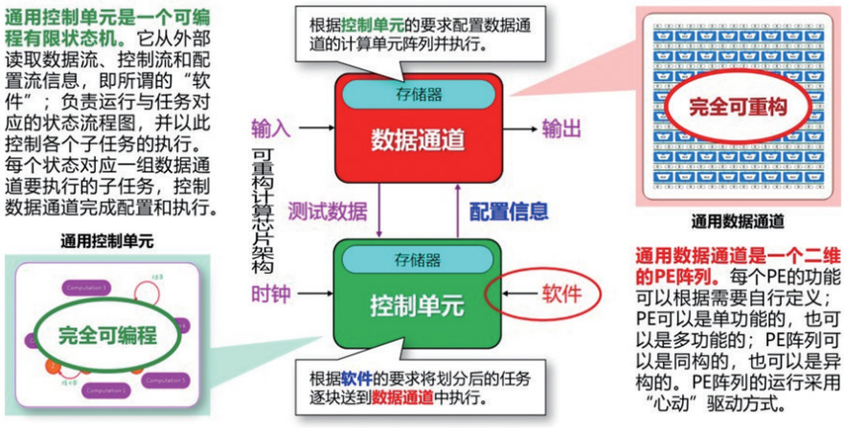


图7　可重构芯片的基本架构

通用控制单元是一个可编程有限状态机(FSM)。它从外部读取数据流、控制流和配置流信息,即所谓的"软件";负责运行与任务对应的状态流程图,并以此控制各个子任务的执行。每个状态对应一组数据通道要执行的子任务,控制数据通道完成配置和执行。这样无论是数据通道还是控制单元,要么是可重构、要么是可编程,这样的结构就可以满足动态可重构芯片的基本结构。

动态可重构芯片系统与高层次综合系统有很多相似之处,例如高层次综合系 统的算子调度与可重构系统的任务调度;二者均有资源的分配及互联的产生过程。

可重构芯片系统与高层次综合系统的区别在于:(1)高层次综合系统的输入采用的是硬件描述语言,动态可重构芯片系统用高级编程语言(如C语言);(2)高层次综合系统使用高层综合的编译器,动态可重构芯片系统采用可重构芯片的编译器;(3)动态可重构芯片系统软件定义芯片技术不再使用指令,而是通过数据流、控制流和配置流实现对芯片功能的再定义。

当然,从编译器的输出所产生的结果也不相同,出现了数据流、控制流和配置流来控制可重构芯片的运行。因此可以看到,20世纪发展的有关计算架构的所有理论在可重构芯片中均可以得到进一步验证和扩展,从理论的完整性和方法学原理上经过了时间的考验,但又有了新的发展。

冯·诺依曼计算架构既是经典的计算机体系结构,也是当代数字集成电路的基本架构。几乎所有的数字

电路的架构最终都可以归结为冯·诺依曼计算架构的变形,可重构计算也不例外。

经典的数字电路架构与可重构计算架构的区别如图8所示。

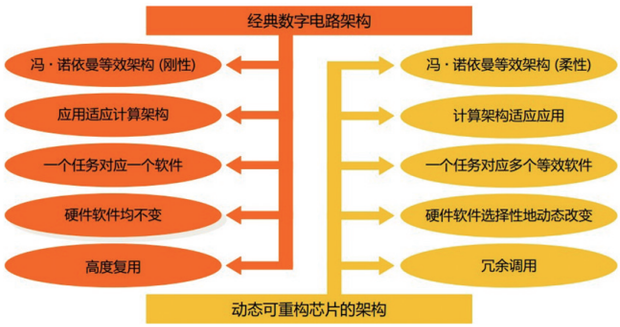


图8 动态可重构芯片架构与经典计算架构的比较

3 动态可重构芯片的若干难点及对应的核心技术

早在20世纪60年代初,可重构的概念就已经被提出,经过了将近60年时间,可重构的问题才真正意义上得到突破,说明该技术的难度非常高。下面将重点介绍动态可重构芯片的若干难点及对应的核心技术。

3.1 配置信息量大幅减少及“隐式配置–数据驱动”技术

要不断地改变一个硬件的架构,需要不断进行配置,那么就要考虑配置的信息量有多大。一般的FPGA的配置信息大概要十几兆、几十兆字节,耗费几百毫秒到几秒的配置时间。要在很短的时间内实现配置的变化,首先需要减少配置信息量。通过对计算流图的分析,基于子图同构相似度匹配的层次化配置信息生成技术,按照子图间相似度匹配和交叉索引的方式,提取数据流图共性,形成层次化的配置信息组织结构,可以使配置信息总量减少 70%以上。

3.2 配置信息高效加载及相关性感知缓存及轮摆式加载技术

配置信息量减少后还需要把它加载到数据通道上,配置仍要消耗时间。经过研究发现,很多时候并不需要时刻发送配置信息,有一些配置信息是可以常驻在存储器当中。因此就要一个相关性感知的配置流的缓存策略,加载时一部分加载、一部分运算,也就是轮摆式动态加载机制:采用基于计算任务对配置信息进行分组的片上高速缓存结构及预取方法,消除各层配置流冗余传输,按层向下汇聚配置集合,并采用流水均衡方法优化流水配置间隙,实现轮摆式动态快速加载。通过这样的方式可以使配置信息的读取和加载的速度平均提高12倍,配置量减少,配置速度提升,自然配置时间就变得非常短。

这些技术克服了动态可重构芯片配置信息优化生成、存储和加载难题,通过配置和执行过程的最大限度并行化,实现了纳秒级的功能重构,突破了制约能效提升的技术瓶颈,为动态可重构芯片能够同时实现高能效和高灵活奠定基础。

3.3 高效阵列架构及控制密集型任务并行化方法

可重构计算架构对计算密集型任务很有效,但是如何执行控制密集型任务是一个较难的问题,需要探索控制密集型任务在集中式控制计算阵列上的并行化方法。通过给出通用映射流程,采用执行体和条件计算合并、配置融合、配置分支优化等技术减小控制任务的配置和执行时间,这些优化可提升大约 40%的性能。

同时,对于分布式控制系统,采用控制密集型任务在分布式控制计算阵列上的并行化方法,支持触发式的符合配置的运算单元及其控制,有机结合触发式机制和复合配置结构,高效实现复杂控制流的指令级并行,降低控制流造成的等待和执行代价。通过这些方法使控制密集型任务的处理速度进一步提升,提升20%~140%。

3.4 时域空域协同映射技术

将一个用高级程序语言写成的应用映射到可重构芯片上运行是一个非常复杂的问题。映射过程可以合并使用多种技术,例如面向可重构计算系统不规则应用的激进流水并行技术,针对非规则应用中静态分析难以预测的控制流,利用空域计算资源在运行期激进地并发执行任务,从而高效开发应用中的细粒度并行, 通过一系列组合方法,可以将计算性能提升一个数量级。

还可以采用面向性能优化的多面体模型映射技术,综合考虑动态重构、阵列计算和缓存访问等参量,采用仿射变换和循环分块的联合优化方法建立性能模型、功耗模型,可以使任务执行时间进一步缩短 20%左右。

4 动态可重构芯片的部分成果

可重构芯片领域相关学者经过多年不懈努力,突破了一系列核心关键技术。下面结合清华大学动态可重构芯片课题组近年来取得的部分研究成果,阐述可

重构芯片的广泛应用。

FPGA一般的重构时间是几百毫秒,甚至需要几秒。课题组将可重构芯片技术的配置信息生成和管理技术应用于FPGA,实现的重构时间只有20~40 ns,速度比FPGA快100万倍。课题组与深圳国微技术有限公司进行了卓有成效的合作,将专利技术转移进入商用可编程逻辑器件,已应用于十数种装备。

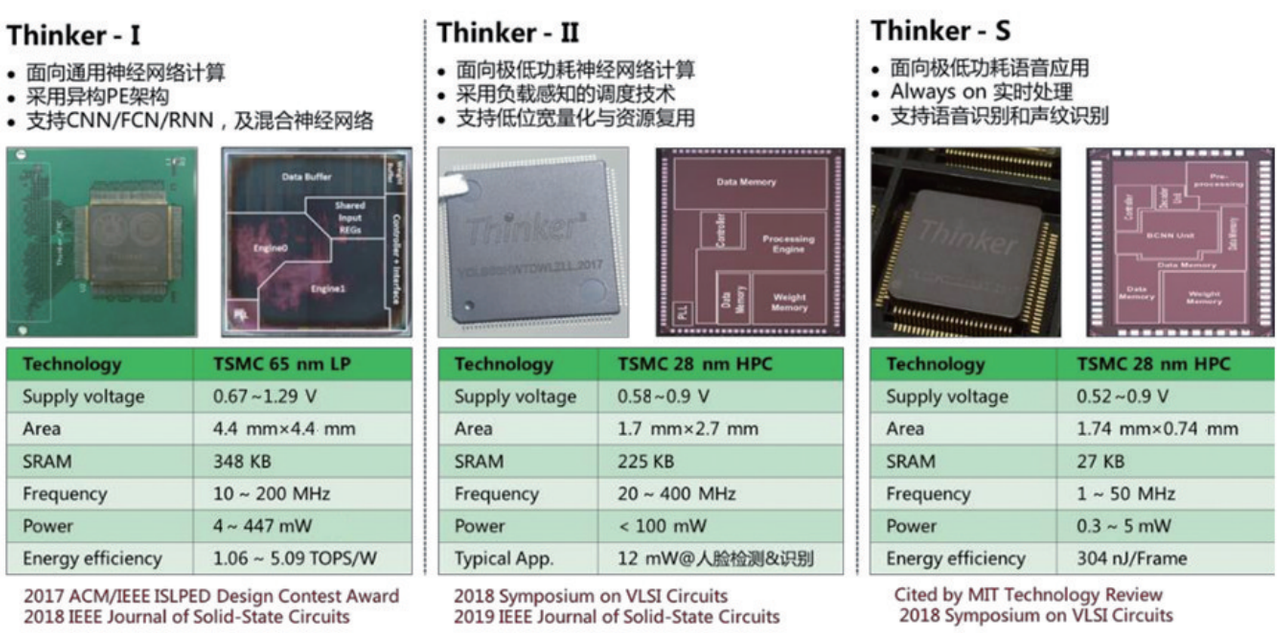
可重构芯片通过让硬件随软件变化而实时、动态地变化,实现电路架构和功能主动适应算法,因此作为信息安全芯片可以保障核心部门的信息保密安全和畅通。课题组研发的可重构加解密计算验证芯片作为国家信息安全领域的核心技术,被北京信息科学技术研究院采用,成为今后持续发展的重要技术基础。

与Intel公司合作,将可重构芯片的部分技术应用到了可穿戴计算机Edison上。该产品在2014年国际消费类电子产品展览会上获得了4项大奖,课题组也获得了教育部2014年度高等学校科学研究优秀成果奖技术发明奖一等奖。Intel公司给予其很好的评价:"动态重构和部分重构等技术大幅提升芯片能效的同时能够满足众多新兴应用对功能灵活性的需求。"

课题组基于动态可重构芯片架构研制出动态可重构芯片RCP,并与中国电子集团和Intel公司合作,将RCP作为硬件加速器集成进服务器CPU芯片。Intel公司向清华大学投入研发经费1.38亿美元,用于研发同时具备高速硬件加速功能和高能量效率的处理器。这是Intel公司第一次与高校合作研发服务器CPU,也是其向单一高校做出的最大单笔研发投资。这一新型处理器很快会走向市场,通过中国电子集团下属澜起科技进行销售,它是一个与X86完全兼容,同时又兼具中国可重构芯片特点的全新的数据中心的服务器芯片。

课题组同时将可重构芯片构架应用到了人工智能(AI)领域,研制成了Thinker系列AI芯片(图9)。以Thinker-II为例,它面向极低功耗的神经网络通用计算,将其配置为人脸识别功能,在LFW数据集上识别率高达99%,同时功耗又非常低,人脸识别功耗仅为12 mW。Thinker-S可用于极低功耗智能语音应用,支持语音识别和声纹的识别,可以广泛用于超便携设备的人机交互,最低功耗不到300 μW。

2018年1月24日,《MIT Technology Review》专题报道了可重构芯片的研究成果,认为该技术能动态调整计算和内存参数来满足实时AI软件的不同需求,是中国取得的一个"Crowning Achievement"。这是近5年来中国大陆半导体技术成果第二次被《MIT Technology Review》报道,成果得到了国际同行的高度认可。



Technology	TSMC 65 nm LP
Supply voltage	0.67~1.29 V
Area	4.4 mm×4.4 mm
SRAM	348 KB
Frequency	10 ~ 200 MHz
Power	4 ~ 447 mW
Energy efficiency	1.06 ~ 5.09 TOPS/W

2017 ACM/IEEE ISLPED Design Contest Award
2018 IEEE Journal of Solid-State Circuits

Technology	TSMC 28 nm HPC
Supply voltage	0.58~0.9 V
Area	1.7 mm×2.7 mm
SRAM	225 KB
Frequency	20 ~ 400 MHz
Power	< 100 mW
Typical App.	12 mW@人脸检测&识别

2018 Symposium on VLSI Circuits
2019 IEEE Journal of Solid-State Circuits

Technology	TSMC 28 nm HPC
Supply voltage	0.52~0.9 V
Area	1.74 mm×0.74 mm
SRAM	27 KB
Frequency	1 ~ 50 MHz
Power	0.3 ~ 5 mW
Energy efficiency	304 nJ/Frame

Cited by MIT Technology Review
2018 Symposium on VLSI Circuits

图9　应用于Thinker系列AI芯片

5 结论

《国际半导体技术发展路线图(ITRS)2015版》认为,粗颗粒度可重构架构(CGRA)是未来最有发展前途的新兴计算架构之一。在过去的5年当中,清华大学可重构芯片课题组发表SCi论文83篇、Ei论文62篇,申请发明专利126项(美国专利4项)。

软件定义芯片是可以替代专用集成电路的新型电路架构技术,它是一条重要的和全新的技术路线,未来有望使中国集成电路研发摆脱跟随模仿,实现赶超。2018年美国国防高级研究计划局(DARPA)正式启动旨在支撑美国2025—2030年电子技术能力的"电子复兴计划"(ERI),其关键技术"软件定义硬件"(SDH)的提出比国内相关研究团队晚了10年,其设定的关键性能指标(重构时间300~1000 ns)远低于国内相关团队已有成果(重构时间20~40 ns)。

"应用定义软件、软件定义芯片"从而实现"应用定义芯片",是集成电路设计技术一次根本性的改变。经过10年的努力,中国已经在该方面取得了很重要的突破。只要继续努力,中国很快可以在可重构芯片(软件定义芯片)研究领域取得更大、更好的成果,中国集成电路发展将大有可为。

(责任编辑 刘志远)